The Pennsylvania State University

The Graduate School

College of Education

EVALUATING LANGUAGE GROUP DIFFERENCES IN THE SUBSKILLS OF READING USING A COGNITIVE DIAGNOSTIC MODELING AND DIFFERENTIAL SKILL FUNCTIONING APPROACH

A Dissertation in

Educational Psychology

by

Hongli Li

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2011

The dissertation of Hongli Li was reviewed and approved* by the following:

Hoi K. Suen
Distinguished Professor of Educational Psychology
Dissertation Advisor
Chair of Committee

Pui-Wa Lei
Associate Professor of Education

Bonnie Meyer
Professor of Education

Aleksandra Slavkovic
Associate Professor of Statistics

Yong-Won Lee
Assistant Professor of Language Assessment
Seoul National University
Special Member

Rayne Sperling
Associate Professor of Education
Professor-in-Charge of Educational Psychology

*Signatures are on file in the Graduate School

ABSTRACT

Using a sequential mixed-methods design, this study examined the differences between two native language groups—those with an East Asian language background and those with a Romance language background —in regard to reading subskills as represented in the Michigan English Language Assessment Battery (MELAB) reading test, so as to provide diagnostic information for second-language reading instruction. With a grounded theory approach that draws on think-aloud reports from a sample of ESL students, it was hypothesized that given the same overall English reading ability, Romance ESL learners would have more mastery of linguistic skills such as vocabulary and syntax, whereas East Asian ESL learners would have more mastery of comprehension skills such as extracting explicit information, and connecting and synthesizing information.

The hypotheses were tested using item response data from 669 examinees drawn from the MELAB dataset with L1 linguistic backgrounds in Chinese, Korean, or Japanese, or a Romance language. First, the subskill profile of each examinee was identified via an application of the item-skill Q-matrix to a Fusion Model of cognitive diagnostic modeling. Second, the specific hypotheses were then tested by comparing the subskill profiles of the East Asian examinees against the subskill profiles of those with a Romance L1 background via Differential Skill Functioning (DSF) analyses through logistic regression techniques.

This study confirmed the hypothesis that given the same overall English reading ability, it is more likely for Romance ESL learners to have mastery of the skill of vocabulary than East Asian ESL learners. Further, given the same overall English reading ability and gender, it is more likely for East Asian ESL learners to have mastery of the skill of connecting and synthesizing information than Romance ESL learners. In addition, given the same overall English reading

ability, female ESL learners are more likely to have mastery of the skill of syntax and the skill of connecting and synthesizing information than males. Instructional strategies are suggested to address the specific weaknesses in ESL learners' reading skills. Implications for the cognitive diagnostic assessment of reading are also discussed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This dissertation would not have been possible without many people. First and foremost, I would like to thank my committee members for their guidance and support.

Dr. Pui-Wa Lei and Dr. Bonnie Meyer are the two instructors who taught my first-year graduate seminar. You have both always been so encouraging and supportive since my very first day in the Educational Psychology program. Thank you for the training, mentoring, and support you've provided to me. And, thanks so much for reading my dissertation draft line by line and giving me so many helpful suggestions.

I also appreciate Dr. Aleksandra Slavkovic for being my minor advisor and a member of my committee. Thanks so much for your constructive feedback on my dissertation. You have always been such a great inspiration to me as I worked to make my dissertation reflect the best of my abilities.

I am indebted to Dr. Yong-Won Lee for many reasons. I still remember how interested I was in the special issue on the diagnostic assessment of reading that you edited for *Language Assessment Quarterly*. Suddenly, I told myself, "OK, this is going to be my dissertation." Thank you sincerely for bringing me into this field and also for serving on my committee.

The most fortunate part of my doctoral study was having Dr. Hoi K. Suen as my academic advisor and dissertation director. I am so glad that I finally found the right focus for my research interests and my abilities—and it is in large part because of you. Not only have you been my mentor for the past four years, but you have become my life-long role model in every respect. I hope that I will be as effective and kind in guiding my own future students as you have been with me. Thank you again for spending so much time cultivating me and this dissertation. I will never

forget all the steps that we took together in order to bring this dissertation to completion. I appreciate very much, too, that you have always had faith in me.

Special thanks go to Dr. Dorothy Evensen who walked me through every step of the grounded theory part of this dissertation. Your qualitative methods class helped me to put my ideas into a mixed-methods framework. Thank you for spending so much time coaching me in regard to think-aloud protocols and coding. Please accept my most sincere appreciation. I would not have been able to complete this dissertation without you.

I also wish to thank Dr. Peggy Van Meter for allowing me to sit in on your Psychology of Reading class. Many of the ideas in this dissertation were generated from your class. Thank you for being so supportive and helpful as always.

I also thank my colleagues at the Office of Learning Initiatives in the College of Information Sciences and Technology at Penn State. In particular, I sincerely appreciate Dr. Lisa Lenze for being such a wonderful supervisor, a devoted mentor, and a trusted friend. I have spent two wonderful and fruitful years in the IST building because of you.

I wish to take this opportunity to thank my professors, mentors, and friends in the Division of English as an International Language at the University of Illinois at Urbana-Champaign. I will never forget your support and guidance when I was working on my TESL degree. Special thanks must go, too, to my dearest friends Julieta Fernandez and Aziz Yuldashev. I am so lucky to have had you as schoolmates at both Illinois and Penn State. Thank you for assisting me in every aspect of my life. Writing in Pattee Library with the two of you sitting side by side is one of my best memories. Thanks to my dear friend Soo Hyon Kim for standing by for my dissertation defense via long distance. I also appreciate my fellow students in the Educational Psychology program at Penn State, Yu-Chu Lin, Yu Zhao, Melissa Ray, and Robin Tate for their friendship and support.

Special thanks go to my husband Zhouqiang Fan for helping me to fulfill my dream. I would not have been who I am without you. Another angel is my little Kevin, who was born when I was just about to start my dissertation. Thanks for bringing so much joy to my life. I actually enjoyed writing the dissertation with you on my lap. I also sincerely appreciate my parents for supporting me as I pursued this work. I cannot express how sorry I am to have been so far away from you over the past years.

Finally, I must thank the English Language Institute at the University of Michigan for providing the Spaan Fellowship in Second or Foreign Language Assessment to support this study. Thank you for generously providing me with both funding and the dataset that formed the basis of this dissertation. I also wish to thank the TOEFL program at the Educational Testing Service for supporting this dissertation with the Small Grants for Doctoral Research in Second or Foreign Language Assessment.

Last but not the least, I truly appreciate the students who participated in the think-aloud activity. Thank you for your interest in my work and your trust in me. I am thankful to the instructors at the Intensive English Communication Program (IECP) at Penn State and the Mid-State Council in State College, PA, for kindly helping me to recruit participants. I am also indebted to the four reading experts for providing important input into this study.

And of course, there are so many others I am not able to mention by name here. My sincerest gratitude goes to each and every one of you.

To My Husband Zhouqiang Fan

My Son Kevin Li Fan

My Mother Xiurong Zhang

My Father Zhoucun Li

CHAPTER 1

INTRODUCTION

1.1 Background

The concept of language transfer originates from the Contrastive Analysis Hypothesis (CAH). Common among its different versions, the CAH holds that "where two languages were similar, positive transfer would occur; where they were different, negative transfer, or interference, would result" (Larsen-Freeman & Long, 1991, p. 53). It is believed that a more effective pedagogy results when the differences between the first language (L1) and the second language (L2) are taken into consideration (Fries, 1945).

L2 reading theory is built based on L1 reading theory; however, L2 readers and L1 readers are distinct in many ways. The most important difference is that L2 readers draw on their prior literacy experience in L1 to facilitate their L2 reading, and thus L2 reading requires dual-language involvement. Some researchers (e.g., Clarke, 1980; Cziko, 1980) argue that L2 reading largely depends on L2 linguistic ability, whereas others (e.g., Cummins, 1984; Esling & Downing, 1986; Goodman, 1971) take the position that L2 reading performance most likely depends on L1 reading ability. In general, poor reading in L2 may be due to poor reading ability in L1, or poor linguistic ability in L2, or both (Alderson, 1984). For instance, L2 reader A is very competent in his L1 reading but has limited linguistic competence in L2. Though this reader may be very skillful with reading strategies, he may still have poor comprehension of the L2 text due to a lack of knowledge of L2 vocabulary and syntax. In contrast, L2 reader B, who is not a good reader in L1 due to ineffective reading strategies, may bring these ineffective reading strategies to his L2 reading. Thus, he may show poor L2 reading competence despite his relatively good L2

linguistic ability. Therefore, A and B may show similar overall reading performance in a reading test, but for different reasons.

The language-specific perspective of reading skills transfer, which emerged from cross-linguistic research, contends that the cognitive mechanism used in linguistic processing differs across languages and thus is language-specific (Koda, 2005). A large number of studies have investigated how L1 processing skills may be incorporated in L2 processing (e.g., Green & Meara, 1987; Hancin-Bhatt & Nagy, 1994; Juffs, 1998; Koda, 1990, 2000a, 200b; Wang, Koda, & Perfetti, 2003). For instance, Harrington (1987) found that Japanese learners of English use processing strategies similar to those used by Japanese L1 rather than those used by the English L1 group. Also, many studies have found that Spanish speakers show advantages in English vocabulary recognition because of the cognates (e.g., Chen & Hennning, 1985; Ryan & Bachman, 1992). It is accepted, therefore, that English as a Second language (ESL) learners with different native language backgrounds behave in different ways when learning the same foreign language (Ringbom, 1987).

1.2 Statement of the Problem

ESL learners from East Asian countries, especially China, Japan, and Korea, constitute the group that faces the greatest challenge in learning English. One commonality that the main languages of these three countries share is that they use scripts that are radically different from the Roman alphabet (Taylor, 1998). The grammar systems are different among the three languages in that Chinese belongs to the Sino-Tibetan family (Thurgood, & LaPolla, 2003), whereas Japanese is regarded as a Japonic language (Shibatani, 1990), and Korean is considered to be an Altaic language (Ramstedt, 1928). Additionally, the grammar system of each is very

different from that of English, which is an Indo-European language. The Defense Language Institute of the United States classifies Chinese, Japanese, and Korean (along with Arabic) as Category IV languages, meaning that 63 weeks of instruction (as compared to just 25 weeks for French, Spanish, Portuguese, and Italian) are required to bring an English-speaking student to a limited working level of proficiency (Raugh, 2008). Conversely, it takes enormous effort for East Asian ESL learners to learn English due to the huge differences between their native languages and English. The greater the difference between the mother tongue and the target language, the less useful the mother tongue is to learners attempting to acquire the latter (Corder, 1983).

Due to the differences between East Asian languages and Indo-European languages, it is not uncommon to find that East Asian ESL learners and Indo-European ESL learners differ in their English reading processes and skills. For instance, in Roman alphabetic systems, such as English and Spanish, each letter represents a phoneme, whereas in logographic systems, such as Chinese characters and Japanese Kanji, each symbol maps into a morpheme (Perfetti & Dunlap, 2008). Readers most familiar with the Roman alphabetic system and those most familiar with the logographic system appear to use different cognitive processes. This means that it is more difficult for East Asian ESL learners to read English than it is for, say, Spanish speakers. Furthermore, word order is a critical device for constructing syntactic relations in English sentences; however, Korean and Japanese learners of English may refer more to case-marking particles as a signaling device due to the syntactic features of Korean and Japanese (Koda, 1993). To summarize, East Asian ESL learners' English reading processes and skills may be different from those used by individuals whose native languages are Indo-European.

Another distinct feature of these three East Asian countries is their English instruction and testing practice. The civil service exam, which started over 2,000 years ago in China, not

only dominated the Chinese historical education system but also influenced neighboring countries such as Korea, Japan, and Vietnam (Suen & Yu, 2006). Currently, English language tests are used as gate-keeping devices for access to general employment and higher education in those countries (Ross, 2008). It is, therefore, not surprising that the teaching and learning of English is intensively test-oriented. Traditionally, the teaching of English in East Asian countries has been dominated by a teacher-centered, book-centered, grammar-translation method (Rao, 2001), which emphasizes rote memorization rather than communication and higher-level thinking skills. Those distinctive teaching and learning styles may influence East Asian ESL learners' reading skills and strategies. For instance, according to Abbott (2006), compared to Arabic ESL learners, ESL learners from China had an advantage in terms of extracting explicitly stated information in reading due to their intensive training with bottom-up reading skills (e.g., the skills focusing on word meaning, syntax, or text details), even though they were likely to find some higher-order reading skills (e.g., the skills focusing on the gist of a text, background knowledge, or discourse organization) to be challenging. Therefore, the teaching and learning styles of East Asian countries may shape their ESL learners' reading in different ways than they do for Indo-European ESL learners.

East Asian ESL learners constitute a large population in the ESL community, and this group also faces much greater challenges than do learners whose first language is Indo-European. It is of particular importance, therefore, to investigate how best to instruct members of this group in English reading. And, in order to collect detailed diagnostic information, it is necessary to conduct a cognitive diagnostic analysis of their reading skills as compared to those of Indo-European ESL learners.

1.3 Proposed Study

An overview of the study is outlined in this section. The purpose of this study was to examine native language group differences in the subskills of reading as represented in the Michigan English Language Assessment Battery (MELAB) reading test. Similar to the Test of English as a Foreign Language (TOEFL), the MELAB evaluates the advanced-level English language competence of adult nonnative speakers of English. Its reading section consists of four passages designed to assess examinees' understanding of college-level reading texts.

It is conceptually appealing to understand L2 readers' performance in reading at the subskill level. However, it is very difficult to disentangle and report examinee performance regarding the subskills of reading using traditional psychometric tools. Reading is usually treated as a unidimensional construct by test developers, particularly by those who employ the common Item Response Theory (IRT, Lord & Novick, 1968) modeling for scaling and test calibration. Typically, a scaled score and/or a percentile rank on a common scale are provided as the result. Comparing the respective English reading ability of different language groups within the IRT framework does not yield subskill information, as all examinees are ranked on a single continuum. In contrast, the use of Cognitive Diagnostic Models (CDMs) would yield more detailed scores in that examinees are assigned a multidimensional profile for the subskills involved in the test (DiBello, Roussos, & Stout, 2007). This more fine-grained diagnostic information can be extracted from test responses and can be subsequently used to effectively support teaching and learning.

The present study compared the reading subskills of two native language groups. One group consisted of individuals whose native languages are Chinese, Korean, or Japanese. This group is referred to as the "East Asian" group. The other group consisted of individuals whose

native language is one of the Romance languages, referred to as the "Romance" group. Indo-European languages consist of many sub-language families, such as Albanian, Armenian, Balto-Slavic, Celtic, Germanic, Hellenic (Greek), Indo-Iranian, Romance, and Tocharian (Fortson, 2004). The English language belongs to the Germanic languages, but it is very close to the Romance languages — thanks to the influence of Latin and French (Crystal, 2004). Therefore, only Romance language speakers were considered in this study due to the similarities between these languages and English.

Comparisons between the East Asian and Romance groups were made using a Differential Skills Functioning (DSF) approach, which is technically adapted from Differential Item Functioning (DIF, Camilli & Shepard, 1994). DSF occurs when examinees from different groups show different probabilities of success with a certain skill underlying the measured construct, after being matched on the underlying ability the test is intended to measure (Milewski & Baron, 2002). Directly comparing the English reading subskills of these two groups may not yield much information, because examinees whose first language is Romance usually perform better on English reading tests than examinees whose first language is an East Asian one. Therefore, the comparison was conducted under the condition that the two groups had the same overall English reading ability (i.e., overall English reading ability was controlled for).

Overall, the study used a sequential mixed-methods design that combines the strengths of quantitative and qualitative methods. In quantitative research, researchers try to confirm hypotheses or answer research questions focusing on assessing the generalizable relationships among variables or testing a treatment variable. In qualitative research, however, the inquiry is more exploratory, emphasizing the description and understanding of a central phenomenon. Researchers analyze the data for a rich description of the phenomenon as well as for themes,

which, in turn, leads to new questions and interpretations (Creswell, Clark, Cutmann, & Hanson, 2003). With a sequential mixed-methods design, qualitative data are collected first; next, quantitative data are used to explain or confirm relationships suggested by the qualitative data (Creswell, 2002).

The first stage of the study generated hypotheses about the reading subskill differences between East Asian and Romance ESL learners. Studies on the transfer of L1 to L2 reading provide the theoretical framework wherein East Asian and Romance ESL learners are expected to show different patterns in their reading processes and skills. However, to date, evidence regarding possible differences at the subskill level is insufficient. Grounded theory is a widely used qualitative method that builds a theory based on data when the theory is not available or insufficient (Glaser & Strauss, 1967; Strauss & Corbin, 1990). It is usually used to explore and understand how complex phenomena occur. Therefore, data were collected via think-aloud protocols from ESL learners with an East Asian language background and ESL learners with a Romance language background and analyzed with a grounded theory approach, i.e., through constant contrastive comparison (Glaser, 1978). This stage resulted in explicit hypotheses regarding how East Asian ESL learners and Romance ESL learners differ in terms of reading subskills.

The second stage of the study quantitatively tested the hypotheses using cognitive diagnostic modeling and the DSF approach with a large-scale dataset of the MELAB reading test. The Fusion Model (Hartz, 2002) was used to estimate examinee profiles on each reading subskill, i.e., examinees were each identified as masters or non-masters of each reading subskill underlying the MELAB reading test. However, one critical input for the Fusion Model is a Q-matrix which represents the subskills required by each item. Therefore, prior to developing and

testing the hypotheses for this study, it was necessary to build and validate a Q-matrix underlying the MELAB test in a series of pilot studies. With data collected from multiple sources, such as the think-aloud protocol (Pressley & Afflerbach, 1995) and expert rating (Leighton & Gierl, 2007) in the pilot study, a tentative Q-matrix was initially developed. This Q-matrix was then validated via an application of the Fusion Model using data from the MELAB program. Subsequent to the Q-matrix validation and Fusion Model calibration, a logistic regression DSF approach was used to test the research hypotheses by comparing the reading subskill differences between the two language groups, when their overall English reading ability was controlled for.

1.4 Overview of the Organization

The focus of this study was to test the theory that ESL learners from different linguistic/cultural backgrounds who are otherwise equal in overall English reading ability evince important differences in English reading subskills. Such differences have important pedagogic implications for teaching English reading to ESL learners from different parts of the world. In practice, this theory cannot be tested with currently existing instruments. This is because most available large-scale English reading tests for ESL learners today have a focus on a unidimensional construct of overall English reading ability and the scaling of these instruments generally aims toward such a singular overall English reading ability. In order to test the theory about subskills, it was necessary to retrofit an existing ESL test to determine the implicit latent subskills underlying the otherwise unidimensional test.

For the purpose of this study, the test retrofitted in a series of pilot studies was the MELAB test administered by the English Language Institute (ELI) at the University of Michigan. In order to extract the examinees' latent reading subskill profiles, it was necessary to

identify the subskills required by each item in the test, known as the Q-matrix. Chapters 2 and 3 provide the background, the theories, and the methodology used in a series of pilot studies involving the use of think-aloud protocols, expert rating, and statistical analysis of MELAB data via an application of the Fusion Model to identify and validate the Q-matrix for the MELAB.

The specific hypotheses on the reading subskill differences were generated in another series of pilot studies via a grounded theory approach based on think-aloud reports from a sample of ESL students. The hypotheses focused on the differences between East Asian learners and learners with a Romance L1 linguistic background. The final hypotheses and the process involved in generating these hypotheses are described in Chapter 4.

The hypotheses were tested using item response data from 669 examinees drawn from the overall MELAB dataset with L1 linguistic backgrounds in Chinese, Korean, or Japanese, or a Romance language. The subskill profile of each examinee was identified by applying the Q-matrix, as developed in Chapters 2 and 3, to a Fusion Model of cognitive diagnostic modeling. The specific hypotheses developed in Chapter 4 were then tested by comparing the subskill profiles of the East Asian examinees against the subskill profiles of those with a Romance L1 background via a series of DSF analyses through logistic regression techniques. A detailed description of the design and procedures to test the hypotheses is provided in Chapter 5. Finally, Chapter 6 discusses the overall findings and their implications for the second-language reading instruction and cognitive diagnostic assessment of reading. Limitations of the dissertation and important areas for future research are also addressed.

CHAPTER 2

COGNITIVE DIAGNOSTIC MODELING IN READING RESEARCH

With traditional Item Response Theory (IRT) (Lord & Novick, 1968) modeling, examinees' abilities are ordered along a continuum. Typically, a scaled score and/or a percentile rank are provided as the result. Results of scoring via Cognitive Diagnostic Models (CDMs) are different, however, in that examinees are assigned multidimensional skill profiles by being classified as masters versus non-masters of each skill involved in the test (DiBello, Roussos, & Stout, 2007). Fine-grained diagnostic feedback can thus be provided to teachers and students to facilitate teaching and learning. The No Child Left Behind Act (2002) has brought increased emphasis to providing more detailed diagnostic feedback to examinees and other stakeholders. Therefore, even though cognitive diagnostic analysis first appeared almost two decades ago (e.g., Tatsuoka & Tatsuoka, 1982; Tatsuoka, 1983), it has become the subject of broad and intensive attention in recent years.

In this chapter, the techniques of CDMs and especially the Fusion Model are introduced. Then a review of component skills of reading and reading taxonomies used in cognitive diagnostic analysis is presented. Finally, methods used in Q-matrix construction in reading research are summarized.

2.1 Overview of Cognitive Diagnostic Models

With a CDM, examinees are assigned multidimensional skill profiles that classify them as either masters or non-masters of each skill involved in the test (Dibello & Stout, 2007). However, currently researchers have different opinions regarding what counts as a CDM. Fu and Li (2005) loosely defined CDMs as "all explicitly and implicitly multidimensional (at test level)

psychometric models" (p. 4). As a result of this broad definition, they listed as many as 62 CDMs in their review. DiBello, Roussos, and Stout (2007), however, defined a narrower scope. They focused on "psychometric models that explicitly contain multiple examinee proficiency variables corresponding to the skills or attributes that are to be diagnosed" (p. 984).

Despite the disagreement over the definition and scope of CDMs, Rupp and Templin's (2008) review is regarded as the most detailed and comprehensive one in recent years. In this review, CDMs are defined as:

> probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modeling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables. The predictor variables are combined in compensatory and noncompensatory ways to generate latent classes. (p. 226).

This definition is even narrower than that given by DiBello, Roussos, and Stout (2007), as the latent variables are specified as categorical. To be consistent, the definition offered by Rupp and Templin is used throughout this dissertation.

The family of CDMs can be traced back to Fischer's Linear Logistic Test Model (LLTM), which models how the difficulty parameter is influenced by the cognitive operation by decomposing item difficulty parameters into discrete cognitive skill-based difficulties (Fischer, 1973, 1983). However, the difficulty parameter is not item-specific for each skill; rather, this parameter only indicates the difficulty of a skill across the whole test. Therefore, LLTM is regarded as a unidimensional IRT model and does not provide cognitive skill diagnosis for each examinee (DiBello, Roussos, & Stout, 2007). Still LLTM is usually included in the review of CDMs due to its innovative use of the Q-matrix.

One of the earliest methods for cognitive diagnostic analysis, Tatsuoka's (1983) groundbreaking work on the Rule Space Model classifies examinee item responses into categories of cognitive skill patterns. The Attribute Hierarchy Method (AHM) (see Leighton, Gierl, & Hunka, 2004) is an updated version of the Rule Space Model. It specifies the hierarchical relationships among the attributes (or skills), whereas the Rule Space Model assumes a linear relationship. Besides these two models, which are mostly regarded as classification algorithms, most of the other CDMs are IRT-based latent class models (Roussos, Templin, & Henson, 2007). These include the DINA (deterministic input noisy and) model of Haertel (1984, 1989, 1990), the NIDA model of Junker and Sijtsma (2001), the HYBRID model of Gitomer and Yamamoto (1991), the Reparameterized Unified Model (RUM) or Fusion Model of Hartz (2002), the DINO (deterministic input noisy or) model of Templin and Henson (2006), and the NIDO (noisy input deterministic or) model of Templin, Henson, and Douglas (2006). In the following section, some of the important characteristics of CDMs are discussed based on the definition given by Rupp and Templin (2008).

To begin with, one salient characteristic of CDMs is multidimensionality. In unidimensional IRT models, examinee ability is modeled by a single general ability parameter. CDMs make it possible to investigate the mental processes underlying observed responses by breaking the overall ability down into different components. The number of dimensions depends on the number of skill components involved in the assessment. In the area of language testing, it is not clear whether a unidimensional approach or a multidimensional approach is most appropriate. For instance, based on related literature, expert judgment, and examinee verbal reports, Jang (2005) identified nine subskills involved in the TOEFL reading test and provided examinee performance profiles on each of the nine subskills as a result of using a CDM.

However, reading ability involved in the TOEFL has been found to be unidimensional by a confirmatory factor analysis (Sawaki, Stricker, & Oranje, 2009). It seems that the subskill dimensions in Jang's study were likely to be substantive dimensions rather than statistical dimensions. Substantive dimensions are usually supported by test specifications, content analysis, and expert judgment, and subsets of items are arbitrarily assigned on each dimension. However, the substantive dimensions may not be consistent with the results of statistical analyses (Walker, Azen, & Schmitt, 2006). Language tests in particular assess a variety of skills for examinees from diverse educational, linguistic, and cultural backgrounds and thus may show more than one dimension (Henning, Hudsan, & Turner, 1985). Haberman and von Davier (2007) have commented on this as a dilemma. They stated that on the one hand, using a multidimensional model for a unidimensional test is probably unnecessary and less accurate; on the other hand, the market demands a richer diagnostic feedback provided by a multidimensional analysis. As suggested by Junker and Sijtsma (2001, p. 271), "even when the fit is good, standard unidimensional IRT modeling might not be as relevant as some discrete attributes models, if the goal of testing is cognitive assessment or diagnosis." Therefore, Jang's work contributed toward goals of remediation by assessment through her fit of the multidimensional CDM with the TOEFL reading test.

Second, CDMs are inherently confirmatory. The loading structure of a CDM is the Q-matrix, i.e., a particular hypothesis about the examinee's response process using 1s or 0s that indicate which skills are associated with which items. Attributes are defined as "a description of the procedures, skills, or knowledge a student must possess in order to successfully complete the target task" (Birenbaum, Kelly, & Tatsuoka, 1993, p. 443). For the purpose of this discussion, "attributes" are used interchangeably with "skills." We will let k stand for the number of skills

being measured, *i* stand for the number of items, and *j* stand for the number of examinees. $Q = \{q_{ik}\}$, where $q_{ik} = 1$ when skill *k* is specified as being required by item *i*, and $q_{ik} = 0$ when skill k is not required by item *i.* As shown in Table 2.1, skill A is required by item 1, and skill B and skill C are required by item 2. The simplest structure for a Q-matrix only has one skill associated with each item, but more complex models have more than one skill associated with each item. Ideally, the relationship demonstrated in a Q-matrix is specified *a priori*. As described by Gierl and Cui (2008, p. 265), "a cognitive model would be developed first to specify the knowledge and skills evaluated on the test and then items would be created to measure these specific cognitive skills." However, currently very few large-scale tests are designed with a cognitive diagnostic purpose; therefore, in most application studies, the Q-matrices have been constructed retrospectively with existing tests. Haberman and von Davier (2007) have cautioned about the danger involved in retrofitting, as it is difficult to identify the skills involved in the test items. Still, successful studies (e.g., Jang, 2005; Klein, Birenbaum, Standiford, & Tatsuoka, 1981) have shown that using CDMs with existing tests can extract richer diagnostic information.

Table 2.1

Sample Q-Matrix

	Skill A	Skill B	Skill C
Item 1	1	0	0
Item 2	0	1	1
…	…	…	…

Third, CDMs allow for both compensatory and non-compensatory (or conjunctive) relationships among subskills, although non-compensatory models are currently more popular (Roussos, Templin, & Henson, 2007). With a compensatory model, a high level of competence on one skill can compensate for a low level of competence on another skill in performing a task.

In contrast, with a non-compensatory model, a high level of competence on one skill cannot offset a low level of competence on another skill. Some of the most well-known non-compensatory models are the Rule Space Model, the Attribute Hierarchy Method, the DINA, the NIDA, the HYBRID model, and the Reparameterized Unified Model (RUM), also known as the Fusion Model. The DINO and NIDO models, however, are compensatory. The question of whether we should use non-compensatory or compensatory models does not have a clear-cut answer, and the answer mainly depends on the theory of the construct, the diagnostic setting, and how the Q-matrix is specified.

Finally, unlike traditional IRT models, which generally model continuous latent variables, the latent variables in CDMs are discrete. At present, most CDMs and the associated estimation procedures only allow for dichotomous latent variables (e.g., mastery vs. non-mastery), though theoretically the models can be extended to polytomous/ordinal levels, such as a rating variable with the values of "outstanding performance," "good performance," "fair performance," and "poor performance." The MDLTM software (von Davier, 2006) for the General Diagnostic Model (Xu & von Davier, 2006) allows for dichotomous or polytomous latent variables; however, in practice most application studies using this software to date have modeled dichotomous latent variables in order to reduce the complexity of estimation.

In conclusion, the purpose of using CDMs is to provide a fine-tuned examinee performance profile relating to multiple skills. A typical procedure of the cognitive diagnostic analysis is as follows: (i) identify a set of skills involved in a test; (ii) construct a Q-matrix demonstrating the relationships among the skills and the test items; (iii) estimate the profiles of skill mastery for individual examinees based on actual test performance data; and (iv) provide score reporting and or diagnostic feedback to examinees and other stakeholders (Lee & Sawaki,

2009b). However, due to the relative newness of CDMs and their requirements for data from diagnostic tests, the application of CDMs is still limited.

2.2 Fusion Model

2.2.1 Introduction to the Fusion Model

Among the large number of CDMs, the Fusion Model (Hartz, 2002; Roussos, DiBello, et al., 2007) is particularly promising for cognitive diagnostic analysis with reading tests. Also known as the Reparameterized Unified Model (RUM), the Fusion Model is an IRT-like multidimensional model that expresses the stochastic relationship between item responses and underlying skills as follows:

$$P\left(X_{ij}=1\middle|\bar{\alpha}_j,\theta_j\right)=\pi_i^*\ \prod_{k=1}^{k} r_{ik}^{*(1-\alpha_{jk})\,q_{ik}}\,p_{ci}\,(\theta_j) \qquad [2.1]$$

Where,

X_{ij} is the response of examinee j to item i (1 if correct; 0 if incorrect); and

q_{ik} specifies the requirement of mastery of skill k for item i (q_{ik} = 1 if skill k is required by item i; q_{ik} = 0 otherwise).

There are two ability parameters, $\bar{\alpha}_j$ and θ_j:

$\bar{\alpha}_j$ refers to a vector of cognitive skill mastery for examinee j for skill k specified by the Q-matrix ($\bar{\alpha}_{jk}$ = 1 if examinee j has mastered skill k; $\bar{\alpha}_{jk}$ = 0 if examinee j has not mastered skill k); and

θ_j represents a residual ability parameter of potentially important skills unspecified in the Q-matrix in the range of -∞ to ∞.

There are three item parameters, π_i*, r_{ik}*, and c_i:

π_i* is the probability that an examinee, having mastered all the Q-matrix required skills

required for item i, will correctly apply all the skills to solving item i. π_i* can be interpreted as the Q-matrix-based difficulty level of item i, ranging from 0 to 1; and

r_{ik}* = $P(Y_{ijk} = 1|\alpha_{jk} = 0)/P(Y_{ijk} = 1|\alpha_{jk} = 1)$ is an indicator of the diagnostic capacity of item i for skill k, ranging from 0 to 1. The more strongly item i requires mastery of skill k, the lower is r_{ik}*. r_{ik}* can be interpreted as the discrimination parameter of item i for skill k; and

c_i is an indicator of the degree to which the item response function relies on skills other than those assigned by the Q-matrix, ranging from 0 to 3 (the bounds were chosen for convenience). The lower the c_i is, the more the item response function depends on residual ability θ_j. When c_i is 3, $p_{ci}(\theta_j)$ is very close to 1, which means that the item response function is practically uninfluenced by θ_j; when c_i is 0, $p_{ci}(\theta_j)$ will dramatically influence the item response probability. Therefore, c_i is regarded as a Q-matrix completeness index.

The number of item parameters specified by the model is dependent on the Q-matrix; each item has $2+k_i$ parameters: π_i*, c_i, and r_{ik}*. When an item is only related to one skill in the Q-matrix, each item would only have one r_{ik}* parameter (Roussos, DiBello, et al., 2007).

The Fusion Model has been intensively studied during the past several years, and some new developments have emerged. For instance, Roussos, Xu, and Stout (2003) studied how to equate with the Fusion Model using item parameter invariance; Bolt, Li, and Stout (2003) explored linking calibrations based on the Fusion Model, and Fu (2005) extended the Fusion Model to handle polytomously scored data using a cumulative score probability function (referred to as PFM-C). Templin (2005) developed a generalized linear mixed model for the proficiency space of examinee abilities (GLMPM) using the Fusion Model. And, Henson and

Templin (2004) developed a procedure for analyzing National Assessment of Educational Progress (NAEP) data with the framework of the Fusion Model. Overall, these studies show the great potential of the Fusion Model for cognitive diagnostic analysis.

2.2.2 Estimation Methods and MCMC Convergence Checking

A Bayesian hierarchical structure was developed to increase the capacity of model-data fit and also to simplify and improve the estimation procedure. Hartz (2002) and Roussos, DiBello, et al. (2007) gave detailed descriptions of the Bayesian framework, which is summarized as follows.

The prior used for the residual θ parameter is simply set to a standard normal distribution. However, the Bayesian framework for other ability parameters and the item parameters are much more complicated. The dichotomous α_{kj} ability parameters are modeled as Bernoulli random variables with probability of success p_k, the population proportion of masters for skill k. The prior for the α_{kj} consists of the p_k parameter for each skill and the tetrachoric correlations between all skill mastery pairs. These parameters are modeled as hyperparameters in a hierarchical Bayesian model. The tetrachoric correlations between the dichotomous skills assume that continuous normal random variables underlie the dichotomous α_{kj} mastery variables. It is assumed that the continuous variables have been dichotomized by cut-point parameters. The continuous variables are denoted as $\tilde{a}_{kj}$, which determines the mastery status of a skill when dichotomized by cut points. The cut-point parameter k_k is related to p_k parameters by the relation $P(\tilde{a}_{kj} > k_k) = p_k$, where $\tilde{a}_{kj}$ is assumed to follow a standard normal distribution. The correlations between θ_j and the underlying $\tilde{a}_j$ variables are modeled as nonnegative correlations. These correlates are estimated as hyperparameters and are given a Uniform prior, U (0.01, 0.99).

Both π_i^* and r_{ik}^* range from 0 to 1, whereas c_i approximately ranges from 0 to 3, and thus

$c_i/3$ is used in the following discussion. Because the values of π_i*, c_i, and r_{ik}* can vary greatly for a given data set, the priors (distribution functions) for these three types of item parameters are each chosen to be a Beta distribution. As shown in Hartz (2002, p. 24), the hyperparameters $a_{\pi*}$, a_c, and a_{r*} are given uniform priors of U (0.1, 0.9), and hyperparameters $b_{\pi*}$, b_c, and b_{r*} are given uniform priors of U (0.5, 10.0):

π_i*~β ($a_{\pi*}$, $b_{\pi*}$)

c_i/3~β (a_c, b_c)

r_i*~β (a_{r*}, b_{r*})

Each of ($a_{\pi*}$, a_c, a_{r*},) ~ U (0.1, 0.9)

Each of ($b_{\pi*}$, b_c, b_{r*},) ~ U (0.5, 10.0)

The Arpeggio program (Bolt et al., 2008) incorporates the required flexibility in the relationships between the item parameters and simplifies the estimation procedures by using a Bayesian approach with a Markov Chain Monte Carlo (MCMC) algorithm. The MCMC estimation provides a jointly estimated posterior distribution of both the item parameters and the examinee skills parameters, which may provide a better understanding of the true standard errors involved (Patz & Junker, 1999). Furthermore, MCMC routines adapt easily to produce posterior predictive model diagnostics, and they also provide a ready capability for comparing model parameter prior and posterior distributions as a measure of parameter identifiability (Sinharay, 2006). Therefore, the MCMC algorithm has become popular with many skill diagnostic models, such as the Fusion Model, the NIDA, and the DINA.

The MCMC process converges to a posterior distribution instead of some specified tolerance as in an Expectation Maximization (EM) algorithm. Each time point (or step) in the chain corresponds to a set of simulated values for the parameters. After a long enough number of

steps, i.e., the burn-in phase of the chain, the remaining simulated values will approximate the desired Bayesian posterior distribution of the parameters. In particular, "the values generated by an MCMC algorithm will vary even after convergence, together with the usual analytical intractability of the posterior distribution of interest" (Sinharay, 2004, p. 462). This makes the convergence of MCMC especially difficult to evaluate.

Sinharay (2004) has classified convergence diagnostics into four categories. The first category is the simple graphic method, which works for single or multiple chains. A time–series plot provides a graphical check of the stability of the generated parameter values, whereas a mean plot checks graphically if the mean of a parameter has stabilized. The second method uses ratio of dispersions, useful for multiple chains. For instance, the Gelman–Rubin R ratio uses parallel chains with dispersed initial values to test whether they all converge to the same target distribution. An R value close to 1 indicates convergence. The third method is based on spectral analysis, useful for single chain. For instance, the Geweke Z takes two non-overlapping parts (usually the first 0.1 and last 0.5 proportions) of the Markov chain and compares the means of both parts, using a difference of means test to see if the two parts of the chain are from the same distribution. Parameters with $|z| > 2$ indicate non-convergence. However, because of the conventional Type I error rate used in classical significance tests for multiparameter models, 5% of the calculated Zs are allowed to fall outside the range (Ntzoufras, 2009). The Heidelberger–Welch diagnostic method examines the last part of a chain to evaluate the null hypothesis that the generated Markov chain has stabilized. A one-sided test based on a Cramer-von Mises statistic is used, and small *p*-values (such as < 0.05) indicate non-convergence. The fourth method is based on the theory of Markov chains, useful for single chains. This method uses such indices as the Raftery–Lewis diagnostic, which gives the number of iterations required to attain accuracy *r* with

probability s in estimating quantile q of interest. When the total samples needed are fewer than the MCMC sample, this indicates a lack of convergence.

However, although each method described above provides some check of convergence, none of the methods applied can guarantee convergence of an MCMC algorithm. Therefore, it is generally advisable to apply as many of them as possible. A practical solution suggested by Sinharay (2004) is to choose one or two diagnostics belonging to a number of different types and conclude convergence only when all the chosen diagnostics indicate convergence.

Based on their experience of convergence checking with the Fusion Model, Roussos, DiBello, et al. (2007) concluded that visual inspection of plots is very effective while the Gelman–Rubin R is not very powerful at detecting a lack of convergence. They also cautioned that non-convergence may frequently occur when the c parameter is included in the full-length Fusion model. This is probably because most of their applications were with unidimensional tests; thus, the continuous θ parameter may "soak up" most of the variance in the item responses. If that is the case, a reduced Fusion model without the residual part is probably more practical. If non-convergence still occurs, an extremely long chain can be run in order to make sure the burn-in phase is long enough to reach the posterior distribution phase. Finally, if the longer chain length still does not lead to convergence, one may revisit the model building steps and reconsider the Q-matrix to determine the changes that may be needed.

2.2.3 Model Fit Statistics

Just as with any other statistical models, to evaluate the fit between the model and the data is of crucial importance. Because of the involvement of multiple latent skills, methods to evaluate fit in CDMs are more complex than those used in typical unidimensional IRT applications (Rupp, Templin, & Henson, 2010).

When a Bayesian approach is used for parameter estimation, posterior predictive model checking (PPMC, see Levy, Mislevy, & Sinharay, 2006, 2007; Sinharay, 2005; Sinharay, Johnson, & Stern, 2006) is a typical approach to evaluating model fit that compares observed and model-predicted statistics. In PPMC, the posterior predictive distribution based on the data (i.e., the distribution of new data predicted from the model under a Bayesian framework) is used to simulate a large number of data sets, and a test statistic of interest is computed for each data set. The observed value of the test statistic from the sample data is then compared to the empirical sampling distribution, so that critical values and credible intervals can be computed. Based on these values, the likelihood of the observed values is evaluated to determine whether there is sufficient evidence for item or respondent misfit. The statistics typically calculated in the Fusion Model system are the proportion-correct scores on the items, the item–pair correlations, and the examinee raw score distribution (see Henson, Roussos, & Templin, 2004, 2005).

Another type of fit evaluation is one referred to as an internal validity check (Roussos, DiBello, et al., 2007). The Fusion Model produces two types of such internal validity check statistics: IMstats for item mastery statistics and EMstats for examinee mastery statistics. IMstats compares the observed item scores for masters (masters of all required skills for the item) and non-masters (non-master of at least one required skill for the item). If the model fit is good, a strong difference in performance between masters and non-masters should be expected. An item by item plot performance for masters and non-masters can be examined to help judge model fit. There is no formal hypothesis testing approach for IMstats, because inconsequential differences may be statistically significant due to a large number of examinees. EMstats produces evaluation statistics on an examinee-by-examinee basis. Examinees are expected to have a high probability of answering items correctly if they have mastered the required skills for the items. The

examinees will be flagged if their scores are uncharacteristic of their skill mastery profiles. When there are too many examinees that have aberrant responses, this may indicate a lack of model fit. However, the use of the EMstats index is limited at present probably due to the lack of flexibility of changing the preset criteria in Arpeggio to tailor to data from different tests. Most studies using Arpeggio for Fusion model calibration (e.g., Jang, 2005; Romàn, 2009; Schrader, 2006) did not report this index.

2.2.4 Applications of the Fusion Model with Reading Tests

The Fusion Model, due to its relatively new status, has not yet been widely used. The most exemplary study using the Fusion Model is by Jang (2005), who studied the reading comprehension part of the IBT TOEFL. Based on think-aloud protocols, expert rating, and content analysis, Jang identified nine primary reading skills involved in TOEFL reading and created a Q-matrix demonstrating the specific skills required by each item. Then she fitted the Fusion Model with the LanguEdge field test data of IBT TOEFL to estimate the skill mastery probability for 2,703 test takers. Another accomplishment of the study was profile reporting and the use of diagnostic reports. Before teaching a summer TOEFL class, Jang assessed some students via the Fusion Model and provided diagnostic feedback to each student. Following the class, each student was assessed again, with overall gains in skill mastery shown on the score report. The average change in posterior probability of mastery was an approximate gain of about 0.12, and approximately 85% of the students improved their performance on average over the skills. All the participating teachers reported that the diagnostic feedback was useful for increasing students' awareness of their strengths and weaknesses in reading skills. Jang's study has shown the great potential of using CDMs with existing language tests.

A similar application of the Fusion Model was conducted by Lee and Sawaki (2009a).

Data from a large-scale field test of IBT TOEFL reading and listening were used. Different from Jang's study, only four skills were identified as underlying the TOEFL reading test. Two other CDMs, the General Diagnostic Model, and the latent class model (Gitomer &Yamamoto, 1991) were also used for the analysis. It was found that the three models yielded similar diagnostic results. In addition to reading tests, the Fusion Model has been applied to other tests, such as the Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT, Hartz, 2002), the ACT math (Hartz, 2002), an end-of-course high school geometry examination (Montero et al., 2003), a math test on mixed-number subtraction problems (Yan, Almond, & Mislevy, 2004), the Iowa Tests of Educational Development (ITED, Schrader, 2006), and the Concept Assessment Tool for Statistics (CATS, Román, 2009).

The Fusion Model appears to be very promising for cognitive diagnostic analysis with reading tests. The biggest advantage of the Fusion Model is that it acknowledges the incompleteness of the Q-matrix and compensates for this incompleteness by including the residual parameter c_i, which represents all the other skills that have been used by examinees but not specified in the Q-matrix. As we do not have a full understanding of the cognitive processes underlying reading, it is impossible to be certain that we have identified all the skills necessary to correctly answer an item. The inclusion of the residual parameter admits this practical limitation.

Furthermore, the Arpeggio program helps to modify the Q-matrix by removing insignificant item parameters, thereby facilitating the process of building a valid Q-matrix. As demonstrated by Hartz (2002), the Fusion Model takes a stepwise reduction algorithm to increase the estimation accuracy of the item parameters by eliminating non-informative parameters. Therefore, the Q-matrix can be refined iteratively. For instance, if the best possible r_{ik}* is 0.9, which indicates a lack of diagnostic capacity for discriminating the masters from the non-masters

for skill *k* for item *i*, the corresponding Q-matrix entry can be dropped. Also, a c_i parameter above 2 indicates that the skills required to successfully answer the item are completely specified by the Q-matrix, and thus c_i can be dropped in this case.

Another advantage of the Fusion Model is that it not only evaluates examinee performance on the cognitive skills, but it also evaluates the diagnostic capacity of the items and the test. For instance, the r_{ik}* parameter indicates how strongly an item requires mastery of a skill. The more strongly item *i* requires mastery of skill *k*, the lower is r_{ik}*. If all the r_{ik}* values are very small, the test is considered to have a "high cognitive structure" (Roussos, Xu, & Stout, 2003). Overall, given the complexity of reading comprehension, the Fusion Model has great potential for conducting cognitive diagnostic analysis with reading tests.

2.3 Reading Comprehension Skills

2.3.1 Component Skills of Reading

Regarding the question of "whether separable [reading] comprehension subskills exist, and what subskills might consist of and how they might be classified" (Alderson, 2000, p. 10), researchers hold different positions on a continuum. At one end of the continuum are holistic general-factor theories (Goodman, 1976; Thorndike, 1917a, 1917b, 1917c; Thurston, 1946; Vacca, 1980). At the other end are multiple-factor models (Davis, 1944; Gray 1919; Spearritt, 1972). A popular multi-factor model was proposed by Munby (1978), who argued that 19 micro-skills are required for reading comprehension, such as recognizing the script of a language, deducing the meaning and use of unfamiliar lexical items, understanding conceptual meaning, understanding the communicative value of sentences, recognizing indicators in discourse, having basic reference skills, skimming, scanning to locate specifically required information, and so

forth. This taxonomy has been very influential in language instructional materials and in test development despite a lack of consensus among reading researchers as to whether these dimensions exist. Furthermore, between these two positions is the argument that two factors underlie reading comprehension. These include either "vocabulary," "decoding," or "literal reading" as the first factor and "comprehension" or "inferential reading" as the second factor (Johnson & Reynolds, 1941; Pettit & Cockriel, 1974; Stoker & Kropp, 1960; Vernon, 1962).

A large number of studies have examined the factor structure of reading tests, mainly with factor analyses. According to Lennon's (1962) summary, half of those factor analytic studies found a single general factor and the remainder found two or more factors. As reported by Carver (1992), four factor analyses of the data from several reading tests resulted in an efficiency level factor when there was one factor; and when there were two factors, one was interpreted as an accuracy level factor and the other as a rate level factor. Rost's (1993) factor analysis yielded either the single broad factor of "general reading competence" or, at most, the two factors of "inferential reading comprehension" and "vocabulary" (p. 79). His findings also suggested that high correlations among the subtests made it difficult to differentiate reading subskills; thus, he doubted the possibility of conducting a reliable and valid diagnostic assessment of reading comprehension. To summarize, different results have been generated from different studies, as data from different tests were used, different examinees (native or non-native) took the tests, and different statistical procedures were applied.

Weir and Porter (1994), however, doubted the validity of some one-factor studies. For instance, most studies only targeted native English speakers and/or the factor analysis methods were flawed. When only native speakers were involved in the study (e.g., Lunzer, Waite, & Dolan, 1979; Rost, 1993), a linguistic factor did not emerge as a separate factor because the

native speakers were not likely to experience linguistic problems. In addition, in Rost's study, when rotation was used in the factor analysis, the second factor, which Rost believed to be vocabulary, did emerge. Weir and Porter (1994) thus concluded that although "it may not be consistently possible to identify multiple, separate reading skill components, there does seem to be a strong case for considering vocabulary as a component separate from reading comprehension in general" (p. 5).

Qualitative methods have also been adopted in studies of reading skill components. For instance, in Alderson (1990a, 1990b), a group of experts were presented with a long list of reading skill components and asked to identify which items measured which skills on the list. The results showed a lack of agreement on assigning particular skills to test items and also in regard to whether an item tested a "higher-level" or "lower-level" skill component. Alderson regarded this as evidence against the divisibility of reading skills. However, Weir and Porter (1990) criticized Alderson's study because it lacked clear definitions of "higher-level" and "lower-level" skills and because the raters did not receive appropriate training. Alderson's study was unable to determine what constitutes reading comprehension, but it led to many debates on the divisibility of reading comprehension in the field of second language research.

Some researchers have suggested that there are hierarchical relationships among reading skill components. For instance, Gray (1960) distinguished the skills of reading the lines (the literal meaning of the text), reading between the lines (inferred meaning), and reading beyond the lines (critical evaluations of the text). This leads to an implicit hierarchy of levels of understanding: the literal level may be lower than the level of inferred meaning, which is again lower than the level of critical understanding. Corresponding to this hierarchy is the assumption that it is more difficult to attain the higher level of understanding. Another classification

distinguishes between literal comprehension and inferential comprehension. Literal comprehension is based on lower-level cognitive processes of reading, such as lexical access and syntactic parsing. In contrast, inferential comprehension involves using higher-level cognitive processes to construct a text base (what the text says) and a situation model (understanding what it is about) (Alptekin, 2006).

According to Alderson and Lukmani (1989), "memory," "translation," and "interpretation" are related to "lower-level" skills, whereas "analysis," "synthesis," and "evaluation" are related to "higher-level" skills. They speculated that the lower-level questions might measure language abilities and the higher-level questions might measure cognitive skills, reasoning ability, etc. However, Alderson (1990a) found that lower-level skills were not prerequisites for the high-level skills. In other words, readers with poor performance on lower-level questions did not necessarily fail to answer the higher-level questions correctly. In a response to Alderson's study, Matthews (1990) claimed that lower-level items would probably be more difficult than the higher-level items, because the higher-level items relate to a long passage of text and thus may be easier for poor readers to understand.

It is theoretically and statistically difficult to establish whether there are distinct component skills in reading comprehension; however, identifying reading component skills can provide a useful framework to help in course design, teaching, and test and materials development (Lumley, 1993). Moreover, reading tests designed with a clear subskills structure can provide more fine-grained diagnostic information.

2.3.2 Reading Taxonomies Used in Cognitive Diagnostic Studies

A widely studied reading test involved in cognitive diagnostic analyses is the TOEFL reading test (Jang 2005; Kasai, 1997; Lee & Sawaki, 2009a; Sawaki, Kim, & Gentile, 2009;

Scott, 1998). The Q-matrices used in those studies are typically based on literature, content expert judgment, and/or examinee verbal reports.

An exemplary cognitive diagnosis study of reading skills is provided by Jang's (2005) dissertation. Based on students' verbal protocols along with the analysis of text and items, she identified nine reading subskills involved in IBT TOEFL reading test, including (a) context-dependent vocabulary, (b) context-independent vocabulary, (c) syntactic and semantic linking, (d) textually explicit information, (e) textually implicit information, (f) inferencing, (g) negation, (h) summarizing, (i) mapping contrasting ideas into mental framework. These nine skills with their descriptions were presented to five experts who then identified which skills were involved in each of the 37 items. Overall, 26 of 37 items showed a moderate degree of agreement on skills identified by the experts. Jang found that the experts had difficulty distinguishing between "textually implicit information" and "inferencing" skills. Also, the experts tended to identify and assign both "context dependent" and "context independent vocabulary" skills to the same items. In the final Q-matrix, 12 out of 37 items each required one skill, 20 items each required two skills, and only five items each required three skills.

However, Sawaki, Kim, and Gentile (2009) reported a different set of reading subskills in the same IBT TOEFL reading test. In this study, the expert team initially identified six subskills as potential categories for the TOEFL reading test: (a) understanding word meaning, (b) identifying information: search and match, (c) understanding information within sentences, (d) understanding and connecting information within a paragraph, (e) understanding and connecting information across paragraphs, and (f) understanding the relative importance of information and relationships among ideas. The draft Q-matrix with the above-mentioned skills was analyzed with a Fusion Model. The expert team then refined the skills based on multiple rounds of

discussions as well as estimates of the Fusion Model item parameters. It was decided that skills (d) and (e) should be combined into one category called "Connecting information." Also, skills (b) and (c) were combined into one category called "Understanding specific information." Therefore, the final list involved only four skills. Across the two test forms, with 20 items in each form, only 12 in Form A and 10 in Form B were coded for two or three skills.

Table 2.2

TOEFL Reading Skills Identified by Kasai

Category	Skills
Whole passage	1) Low-frequency vocabulary
Locating information	2) Location explicitly indicated
	3) Location indicated by lexical overlap
	4) Location not obvious
Obtaining a correct answer	5) Low-frequency vocabulary
	6) Lexical overlap
	7) Beyond passage
	8) Plausibility of distracters
	9) Understanding the relationship between sentences
	10) Knowledge of rhetorical organization
	11) Time constraint
	12) Lexical overlap (incorrect options)
	13) Complex sentence structure
	14) Infrequent sentence structure
Test-taking strategies	15) Making use of options to obtain the correct option
	16) Long correct option

Kasai (1997) and Scott (1998) used Rule Space Models to analyze TOEFL reading test data. The studies by Kasai and Scot each included significantly more skills than were used by Jang (2005) and Sawaki et al. (2009), and the former studies also included some interactions between different skills. In Kasai (1997), initially 16 skills were identified in four categories, which are summarized in the above Table 2.2. Based on preliminary data analysis results, Kasai decided to further include interactions among the skills. However, it was not clear how to interpret those interactions to examinees and other stakeholders. The process of coding items

with such a large number of subskills is tremendously complex, and to communicate the results to non-expert audience is extremely difficult (Buck & Tatsuoka, 1998).

A different type of reading taxonomy is used with the Attribute Hierarchy Method (AHM), which is an updated version of the Rule Space Model. The AHM assumes that cognitive skills (or attributes) are hierarchically related, which is thought to better reflect the characteristics of human cognition. Wang and Gierl (2007) analyzed SAT critical reading data with the AHM. The final hierarchy is represented in Figure 2.1 and Table 2.3.

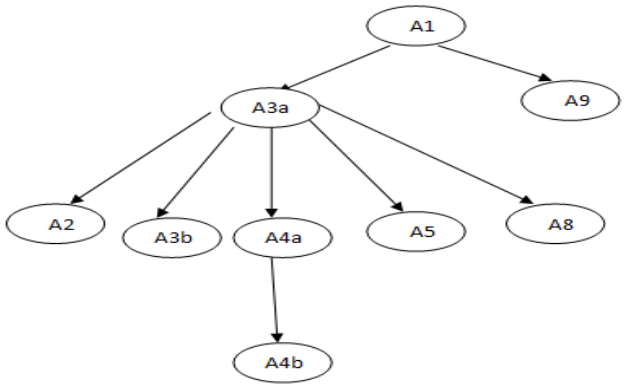

Figure 2.1. Hierarchical relationships among the subskills of SAT critical reading.

Table 2.3

List of Subskills of SAT Critical Reading

Skill	Description
A1	Basic language knowledge, such as word recognition and basic grammar
A2	Determining word meaning by referring to context
A3a	Literal understanding of sentences with minimal amount of inferences
A3b	Understanding sentences by making inferences based on the reader's experience and background knowledge
A4a	Literal understanding of larger sections of text with minimal amount of inferences,
A4b	Understanding larger sections of text by making inferences based on the reader's experience and world knowledge; building coherence across, summarizing, and evaluating larger sections of text
A5	Analyzing author's purposes, goals, and strategies
A8	Using rhetorical knowledge
A9	Evaluating response options

As "basic language knowledge" (i.e., A1 in Table 2.3) is fundamental, all other skills require A1. Skill A3a is the prerequisite of skills A2, A3b, A4a, A5, and A8 in Table 2.3, because the readers must possess A3a before they can use other skills to process the text. For similar reasons, skill A4a is also the prerequisite of A4b. Conducting a cognitive diagnostic analysis with the AHM could generate rich diagnostic information. Unfortunately, so far almost no major test has been designed with the AHM framework. To retrofit an AHM analysis with an existing test is extremely challenging due to the difficulty of identifying the hierarchical relationships between the subskills (Gierl & Cui, 2008).

As discussed, there are divergent views regarding the component skills of reading, and even for the same reading test, such as TOEFL, different taxonomies have been used in cognitive diagnostic analyses. To build a well-validated Q-matrix indicating the relationships between skills and items in a reading test is, therefore, very challenging. The following provides a review of the practices of Q-matrix construction and validation in reading research.

2.4 Q-Matrix Construction and Validation in Reading Research

2.4.1 Terms and Definitions

The very first step in building and using a CDM is to construct an appropriate Q-matrix. Different terms have been used in the literature when discussing the dimensions of the cognitive construct, such as *latent traits, processes, attributes, skills,* and *strategies*. It is, therefore, necessary to first clarify some of these terms and to identify the target of the Q-matrix for this study before delving into the options and methods for the construction of a Q-matrix. *Latent traits* refer to mental components of interests that are thought to be stable across time in contrast to latent states that change over time (Rupp, Templin, & Henson, 2010). *Attributes* are defined as

"a description of the procedures, skills, or knowledge a student must possess in order to successfully complete the target task" (Birenbaum, Kekky, &Tatsuoka, 1993, p. 443). Though *attributes* and *skills* may have different connotations and show different beliefs, they are mostly used as synonyms in the measurement literature. *Skills* are more frequently used in this dissertation as aligned with the language used in the reading literature.

S*trategies* are conceptually distinguishable from *skills*. *Strategies* are specifically used to refer to actions that readers select and control to achieve desired goals or objectives (Johnston & Byrd, 1983; Paris, Lipson, & Wixson, 1983; van Dijk & Kintsch, 1983). As stated by Paris, Wasik, and Turner (1991, p. 611):

> *Skills* refer to information-processing techniques that are automatic, whether at the level of recognizing grapheme–phoneme correspondence or summarizing a story. Skills are applied to a text unconsciously for many reasons including expertise, repeated practice, compliance with directions, luck, and naive use. In contrast *strategies* are actions selected deliberately to achieve particular goals. An emerging skill can become a strategy when it is used intentionally. Likewise, a strategy can "go underground" (cf. Vygotsky, 1978) and become a skill. Indeed strategies are more efficient and developmentally advanced when they become generated and applied automatically as skills. Thus, strategies are "skills under consideration."

Similar to other cognitive diagnostic analyses with reading tests (e.g., Jang, 2005; Lee & Sawaki, 2009a), this dissertation tends to focus on reading *skills* rather than *strategies*. Only the skills regarded as essential to correctly answering the items are coded for building the Q-matrix. However, the difference between *skills* and *strategies* may be blurred, and sometimes it is difficult to make the distinction. Therefore, the use of *skills* in this dissertation does not exclude

the potential involvement of *strategies* in some circumstances.

2.4.2 Methods Used in Q-Matrix Construction and Validation

To construct a Q-matrix is the critical first step for the cognitive diagnosis analysis. If a test is developed with a cognitive diagnostic purpose in mind, the structure of the Q-matrix should be specified beforehand. However, currently most large-scale tests are not designed with diagnostic purposes; therefore, most diagnostic studies reported in the literature retrofitted the models with existing tests, which makes constructing Q-matrices especially challenging.

Though various methods have been used to build Q-matrices, the following procedure described by Buck et al. (1998) is typical: (i) develop an initial list of skills, (ii) code each item based on whether or not the item requires each skill to construct an initial Q-matrix, (iii) analyze data using an appropriate CDM with the developed Q-matrix, and (iv) modify the initial Q-matrix based on statistics on each skill along with the theoretical importance of the skill. Steps (iii) and (iv) are repeated until a well-defined Q-matrix is achieved. Common approaches used in steps (i) and (ii) are described in detail next.

First, it is considered economical and efficient to produce an initial Q-matrix using existing test specifications. Test specifications are usually a two-way matrix that shows relationships between content and skill, anticipating the cognitive skills that might be used in each content area (Bloom, 1956; Gierl, 1997). For instance, Xu and von Davier (2008) analyzed the grade-12 data from the 2002 NAEP Reading Assessment using the General Diagnostic Model. A simple Q-matrix was constructed with three subscales—reading for literary experience, reading for information, and reading to perform—taken as three subskills. For example, if an item was measured in the reading for information scale, then the item had 1 on this skill and 0 on the other two subskills (i.e., reading for information and reading to perform a task). Despite the

low cost and convenience of relying on test specifications, the categories indicated in the test specifications are usually very general. To rely entirely on test specifications to explain cognitive processes is, therefore, unwarranted (Leighton & Gierl, 2007).

An important empirical source of information, in addition to test specifications, is that of observing students' underlying cognitive processes through think-aloud protocols (Leighton, 2004). That is, the items under consideration are presented to a sample of students who are probed about the thinking processes they engaged in when responding to the items. During an interview, a student must perform the task while continuously reporting the thoughts passing through his/her memory. Then a researcher attempts to deduce the underlying thinking processes based on the student's verbal reports. Although there are concerns about the validity of think-aloud verbal reports, they are regarded as fairly reliable and useful for reading research (Ericsson & Simon, 1993; Pressley & Afflerbach, 1995). The think-aloud protocols have been commonly used to build Q-matrices and to detect cognitive structures in reading research (e.g., Gao, 2006; Jang, 2005; Wang & Gierl, 2007). A concurrent think-aloud approach is more frequently used, whereas retrospective think-aloud has also been widely used. As stated by Pressley and Afflerbach, there may be little difference between concurrent reports of reading processes and briefly delayed ones, as concurrent reports have been observed to involve reporting of the reading process that is just completed. The closer in time the retrospective reports are to when the actual processing occurs, the more likely that traces of the processing that occurred would still be retrievable and thus reportable (Ericsson & Simon, 1993).

Yet another approach is to ask a panel of experts to describe the underlying cognitive processes needed to respond to each question, based on their prior experience in the domain. Despite their potential arbitrariness, expert ratings have been widely used in test development,

standard setting, and for many other educational measurement purposes. The key to using this approach successfully is the composition of the panel, the members of which must deeply understand both the domain and the contexts in which students acquire and use the skills specified (Rupp, Templin, & Henson, 2010). For instance, in Sawaki, Kim, and Gentile (2010), a team of content experts, including three IBT TOEFL assessment development specialists and three language assessment researchers, built the Q-matrix for the TOEFL reading test. One major concern with this approach is that the experts' ability is typically substantially higher than that of the students, and there is no empirical evidence showing that the identified skills and processes are truly used by the students (Leighton & Gierl, 2007).

In addition to approaches discussed thus far, statistical and psychometric models are also used to construct Q-matrices. Factor analyses are a traditional method of detecting the cognitive structure of large-scale tests. For example, based on prior literature suggesting that the three dimensions of morphosyntactic form, cohesive form, and lexical form are measured by the grammar section of the Certificate of Proficiency in English (ECPE) test (Liao, 2007), Henson and Templin (2007) used a three-factor exploratory model to identify basic clusters of items that might measure similar abilities. However, factor analyses are not very effective for Q-matrix construction, especially when retrofitting the CDMs with existing tests; this is because most current large-scale tests are unidimensional.

After an initial Q-matrix is built, large-scale empirical response data can be used for Q-matrix validation based on preliminary results of cognitive diagnostic modeling. For instance, as a result of analyzing the SAT verbal test data with the Rule Space Model, Buck et al. (1998) decided to delete one skill because nearly all the examinees had mastered the skill, its low correlation with the total score, and/or a high correlation with other skills. The Arpeggio

software used for the Fusion Model calibration has a built-in iterative algorithm that indicates the non-significant Q-matrix entries or skills that can be removed. However, this recommendation is solely based on statistical concerns, and it is important to make decisions based on both statistical and substantive input. For example, in Jang (2005), the initial run of the Arpeggio resulted in more than 16 r_{ik}* parameters that were relatively too high (> 0.9), indicating that they were statistically insignificant. However, dropping all of them would have drastically altered the item-by-skill specifications in the Q-matrix and might have made the cognitive structure theoretically less justifiable. She finally decided to drop the c_i parameter and tried to keep the r_{ik}* parameter. In general, using initial parameters and fits of different models implied by different Q-matrices is useful in determining an appropriate Q-matrix structure.

To summarize, Chapter 2 reviews the general background of cognitive diagnostic modeling and especially the Fusion Model. The component skills of reading, reading taxonomies used in cognitive diagnostic analysis, and the methods used for building Q-matrices are also summarized. The following Chapter 3 focuses on the application of these processes to construct and validate a Q-matrix for the MELAB reading test, which is used as the key instrument in this dissertation.

CHAPTER 3

Q-MATRIX CONSTRUCTION AND VALIDATION FOR THE MELAB READING TEST

A Q-matrix, which indicates the skill(s) required by each item in a test, is a crucial input for cognitive diagnostic modeling. However, specifying the skill–item relationship is not an easy task, especially when a cognitive diagnostic model (CDM) is retrofitted to an existing test, such as the MELAB reading test. This chapter describes the procedures used to build and validate the Q-matrix for the MELAB reading test in order to prepare for the follow-up cognitive diagnostic analysis aimed at extracting the examinees' reading subskills as represented in the test.

3.1 Introduction to the MELAB Reading Test

The MELAB evaluates the advanced-level English-language competence of adult nonnative speakers of English. Many educational institutions in the United States, Canada, the United Kingdom, and other countries accept the MELAB as an alternative to the TOEFL (ELI-UM, 2003). It consists of three parts: composition, a listening test containing 50 multiple-choice items, and a grammar/cloze/vocabulary/reading test containing a total of 100 multiple-choice items.

The reading section of the MELAB is designed to assess examinees' understanding of college-level reading texts. All passages are expository texts, and the language is representative of English for academic purposes. The readability of the passages, as measured by a standard readability formula, suggests that their vocabulary and structural complexity are at the college level. To counter any possible bias toward examinees of a particular educational or cultural

background, test developers select texts on a range of topics and include different genres of passages in each test form (ELI-UM, 2003).

The reading section consists of four passages, each of which is followed by five multiple-choice items. Each item consists of a question stem and four options (one key and three distracters). According to the item-writing guidelines provided by the English Language Institute of the University of Michigan (ELI-UM), the organization which is responsible for developing the MELAB, the questions following each passage are intended to assess a variety of reading abilities, including recognizing the main idea, understanding the relationships between sentences and portions of the text, drawing text-based inferences, synthesizing, understanding the author's purpose or attitude, and recognizing vocabulary in context (ELI-UM, 2003). The items with good discrimination and difficulty levels are maintained after some initial field testing.

3.2 Initial Q-Matrix Construction

One important input for the Fusion Model is a Q-matrix, which indicates the skills required by each item in the MELAB reading test. I constructed the initial Q-matrix with a series of procedures as described in the following sections.

3.2.1 Initial Cognitive Framework for the MELAB Reading Test

An initial cognitive framework for the MELAB test was first proposed based on second-language reading theories and related literature. Gao (2006) developed a model of the cognitive processes used by examinees taking the MELAB reading test based on verbal reports from Chinese ESL students and content experts. The model involves 10 general categories of processing components as follows: (a) recognize and determine the meaning of specific words or phrases; (b) understand sentence structure and sentence meaning using syntactic knowledge; (c)

understand the relationship between sentences and the organization of the text; (d) speculate beyond the text; (e) analyze the function/purpose of communication using pragmatic knowledge; (f) identify the main idea, theme, or concept, and skim the text for gist; (g) locate the specific information requested in the question and scan the text for specific details; (h) draw inferences and conclusions based on information implicit in the text; (i) synthesize information presented in different sentences or parts of the text; and (j) evaluate the alternative choices. The relationship between the proposed cognitive processes and empirical indicators of item difficulty was further investigated using the tree-based regression (TBR). The results of Gao's study informed the construct validation of the MELAB reading and laid "a foundation for the MELAB reading as a diagnostic measure" (Gao, 2006, p. 1).

Both the MELAB and the TOEFL are English-language proficiency tests used by North American universities in admission decisions regarding international students. They have very similar content areas and cognitive structures, and a concordance table is available to convert MELAB scores to TOEFL scores and vice versa. Therefore, I also referred to the taxonomies for TOEFL reading used in cognitive diagnostic analyses (e.g., Jang, 2005; Kasai, 1997; Lee & Sawaki, 2009a; Sawaki, Kim, & Gentile; 2009; Scott, 1998). Among the TOEFL taxonomies, Jang's taxonomy is especially detailed; therefore, I examined it the most closely. As reviewed in Chapter 2, based on students' verbal protocols along with content analysis and expert judgment, Jang identified nine reading subskills involved in the IBT TOEFL reading test, including context-dependent vocabulary, context-independent vocabulary, syntactic and semantic linking, textually explicit information, textually implicit information, inferencing, negation, summarizing, and mapping contrasting ideas into a mental framework.

Table 3.1 summarizes Gao's and Jang's cognitive models. The subskills appear to fall

into five categories. The first category is vocabulary. Gao had one subskill for vocabulary, whereas Jang's study listed two: context-dependent and context-independent. The second category is syntax, for which Gao had one subskill and Jang had a separate subskill referred to as negation. The third category is explicit information at the local level for which both researchers had one subskill. For the fourth category—connecting and synthesizing information—Gao listed understanding the relationships between sentences, synthesizing information, and identifying main ideas, whereas Jang listed mapping contrasting ideas into a mental framework and summarizing. The last category is making inferences beyond the text (reading beyond the lines). Gao included speculating beyond the text and making inferences based on implicit information. Similarly, Jang listed inferencing and textually implicit information.

Table 3.1

Summarizing Cognitive Models of Reading as Designated by Gao and Jang

Category	Gao (2006)	Jang (2005)
Vocabulary	• Recognize and determine the meanings of specific words or phrases using context clues or phonological/orthographic/vocabulary knowledge	• Context-dependent vocabulary • Context-independent vocabulary
Syntax	• Understand sentence structure and sentence meaning using syntactic knowledge	• Syntactic and semantic linking • Negation
Extracting explicit information	• Locate the specific information requested in the question and scan the text for specific details	• Textually explicit information
Connecting and synthesizing	• Understand the relationship between sentences and organization of the text using cohesion and rhetorical organization knowledge • Synthesize information presented in different sentences or parts of the text • Identify the main idea, theme, or concept and skim the text for gist	• Summarizing • Mapping contrasting ideas into a mental framework
Making inferences	• Speculate beyond the text; e.g., use background/topical knowledge • Draw inferences and conclusions based on information implicit in the text	• Inferencing • Textually implicit information

Two of the original subskills in Gao's model are not included in Table 3.1. The skill of analyzing the function/purpose of communication using pragmatic knowledge was not found to be associated with any of the items in the current test form of the MELAB. The skill of evaluating alternative choices to select the best answer seemed to be involved with all items and, thus, may have little diagnostic value. Therefore, based on the literature and a brief content analysis of the MELAB reading passages used in this dissertation, I hypothesized the initial framework for MELAB reading as consisting of five categories represented in the first column of Table 3.1. This initial framework was further revised and validated with evidence from students' verbal reports, expert ratings, and the literature.

3.2.2 Think-Aloud Protocol

To supplement the initial framework shown in Table 3.1, think-aloud protocols (Ericsson & Simon, 1993; Pressley & Afflerbach, 1995) were used to gather information about possible cognitive processes involved in responding to the MELAB items. A preliminary pilot was conducted with two ESL students to fine tune the data collection method. I found that highly advanced ESL students only produced minimal verbal reports, as reading the passages was not challenging for them and many processes were automatic. Therefore, I decided to recruit participants from students currently enrolled in ESL classes. The initial pilot of the procedures also showed the superiority of using both concurrent and retrospective verbal reports rather than solely concurrent verbal reports. Therefore, I adopted both concurrent and retrospective think-aloud protocols. Finally, to read and think aloud about four passages in one session was exhausting for participants; thus, I conducted two sessions with two passages per session.

Participants. In the spring and summer of 2010, I contacted two major ESL training centers—the Mid-State Literacy Council in State College, PA, and the Intensive English

Communication Program (IECP) at the Pennsylvania State University—to obtain permission to recruit participants (please see Appendix A for the consent form and Appendix B for the email invitation for the think-aloud activity).

Table 3.2

Background Characteristics of Think-Aloud Participants

Name	Gender	First language (native country)	Highest degree (where obtained)	Major or field of study	TOEFL score	Self-rating of English reading ability
Jin	M	Chinese (China)	Bachelor (China)	Engineering	65*	Basic
Ted	M	Chinese (China)	Master (China)	Education	85	Excellent
Fei	F	Chinese (China)	Bachelor (China)	Philosophy	N/A	Between basic and good
Yao	F	Chinese (China)	Bachelor (China)	Educational technology	85	Basic
Ming	M	Chinese (China)	Bachelor (China)	Computer science	83*	Good
Hon	M	Korean (Korea)	Bachelor (Korea)	Biochemical engineering	N/A	Basic
Chika	F	Japanese (Japan)	Bachelor (Japan)	Social welfare	N/A	Basic
Afsar	F	Persian (Iran)	Master (Iran)	Textile engineering	88	Good
Sabina	F	Spanish (Colombia)	Master (US.)	Agricultural engineering	110	Very good
Katia	F	Portuguese (Brazil)	Master (US.)	Environmental engineering	N/A	Very good
Dora	F	French (Morocco)	High school (Morocco)	N/A	85	Good
Leon	M	Spanish (Colombia)	High school (Colombia)	N/A	N/A	Basic
Eva	F	Spanish (Spain)	Master (Spain)	History and musicology	N/A	Basic

*Note.** Jin and Ming took the paper-based TOEFL, and their original scores were converted to the IBT TOEFL scores.

Given that the dissertation is about the reading skill differences between East Asian ESL learners and ESL learners whose primary languages are Romance languages, participants with those language backgrounds were especially targeted. In April 2010, data were collected from 10 participants: Ted, Chika, Hon, Jin, Sabina, Dora, Leon, Eva, Katia, and Afsar. And, in June 2010, data were collected from three more participants: Fei, Yao, and Ming. In total, 13 ESL students participated in the study, and their background information is shown in the above Table 3.2. Pseudonyms have been used to protect the participants' privacy.

Instrument and Procedures. In order to familiarize participants with think-aloud protocols, a brief training session was provided prior to the formal think-aloud activity. I explained and demonstrated the think-aloud procedure for each participant, and then the participant practiced thinking aloud using the provided training task (please see Appendix C for the verbal script and Appendix E for the think-aloud training material). In order not to distract the participant, I sat at the other end of the desk during the think-aloud session. Only when a long silence occurred, such as 10 seconds, I prompted the participant with questions such as "What are you thinking now?" After the participant had answered all five questions following a passage, he/she would start to recall the processes. At this retrospective stage, I asked some questions for clarification and further inquiry. Each session lasted approximately an hour and was recorded using a digital voice recorder.

All the participants had intermediate to advanced English proficiency and were enrolled in ESL classes taught in English; therefore, they did not have difficulty either understanding or expressing themselves in English. The participants were told to use whichever language they felt comfortable with during the think-aloud activity. Except for a few Chinese participants who used Chinese intermittently, all the other participants used English exclusively.

Data Analysis. I transcribed recordings of all the participating students. The first stage of the coding was open coding, during which I read through the verbal reports line-by-line, underlining any meaningful and interesting parts and commenting on the skills the students had used. The purpose of this initial coding was to understand the skills involved and to revise and validate the initial cognitive framework of the MELAB reading test.

The initial framework was mostly confirmed by the data. First, it was difficult to distinguish whether students determined the meaning of specific words using contextual clues or phonological/orthographic/vocabulary knowledge. For instance, recognizing the word *attenuated* was the key to answering item 16[1]. Students who knew this word beforehand could easily pick the answer containing the word *reduced*. For those students who did not know the word *attenuated*, some successfully guessed the meaning by relying on context. Therefore, I decided to have one vocabulary skill as Sawaki, Kim, and Gentile (2009) did in their diagnostic analysis of the TOEFL reading.

Second, syntactic knowledge was critical for responding to some items. In particular, long and complicated sentences with relative clauses, inversion of subject and verb, passive voice, subjunctive mood, and/or pronoun references seemed to be difficult for students.

Third, in many cases, students needed the skill of understanding explicit information at the local level in order to find answers to the items. Most often, students read the items and then scanned the text searching for specific information relevant to the item. Comprehension usually inhered in a literal understanding of a sentence at the local level.

The fourth category, which focused on connecting ideas from multiple sentences, appeared to involve different levels of elements. In some cases, students only needed to read and

[1] All the MELAB test items demonstrated in this dissertation were paraphrased or adapted for test security purpose.

connect information from adjacent sentences in a single paragraph, such as item 13. However, in other cases, students had to read across different paragraphs or the whole passage in order to find the main idea or the gist of the passage, such as item 3. It would be meaningful to have two separate skills: (a) connecting ideas from multiple sentences; and (b) summarizing for main ideas. Still only one item was found to test main ideas. It was not practical, though, to have a separate skill for main ideas. Therefore, a final decision was made to use the generic skill described as connecting and synthesizing.

The final category was making inferences. The answers to the items were usually implicit in the text, which meant that students had to draw on their background knowledge to answer the questions. For instance, item 11 asked about the validity of the study, and the students with advanced degrees, such as Eva and Ted, picked the answers more easily than did those with only a high school education, such as Leon and Dora. This is probably because those with advanced degrees had received some education in research methods. In general, making inferences appeared to be a distinct skill, as students needed to go considerably beyond the text in order to draw conclusions based on implicit information in the text.

Metacognition and test-taking skills, regarded by some researchers (Baker & Brown, 1984; Ehrlich, 1996; Perfetti, Marron, & Foltz, 1996) as reading strategies, were noticeable in the think-aloud verbal reports. For instance, some students read the questions before reading the passage, and/or they skipped questions that they were not able to answer immediately. Also, some students consistently answered the questions by guessing or by eliminating alternative choices. A typical example in this regard was Chika, a Japanese female, who used eliminating alternative choices for all the items. When asked, she said, “This is my personality. I am never confident about my choices. So I have to make my decision by eliminating other choices.”

Based on the think-aloud verbal reports in conjunction with reading theories, the initial cognitive framework of MELAB reading was revised, as shown in Figure 3.1. It can be seen that the reading comprehension construct underlying the MELAB consists of two major categories: linguistic skills and comprehension skills. Linguistic skills refer to vocabulary and syntax, whereas comprehension skills refer to extracting explicit information, connecting and synthesizing, and making inferences. A residual skill category was also added to the model, which may include metacognition, test-taking, guessing, or any other skills (or strategies due to the potential overlapping between skills and strategies) not specified in the cognitive framework.

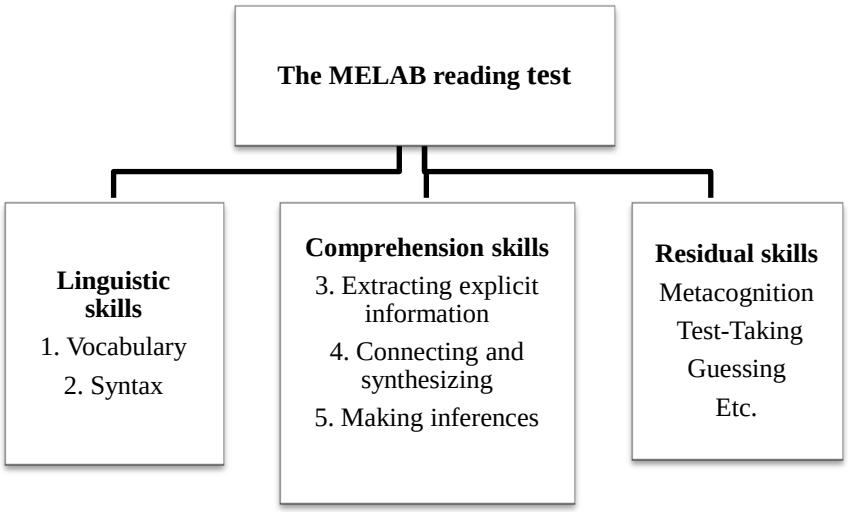

Figure 3.1. Modified cognitive framework of the MELAB reading.

In addition to a substantive concern in building and revising the cognitive framework, another important consideration is the grain size of the subskills of reading. The more categories identified, the closer the cognitive model is to the actual processes underlying reading. However, a Q-matrix representing a large number of subskills may lead to a poor model fit from a statistical perspective. Hartz (2002) suggested that one skill should be assigned to at least three items in order to have sufficient information to estimate that skill. Given the fact that the MELAB reading test has only 20 items, the number of skills was expected to be small.

Table 3.3

Think-aloud Protocols Coding Scheme

Skills	Elaboration	Coding guide
1. Vocabulary	• Recognize and determine the meanings of specific words or phrases using context clues • Recognize and determine the meaning of specific words or phrases using phonological/orthographic/vocabulary knowledge	•Understanding the word is critical for comprehension. •The words are usually infrequently used.
2. Syntax	• Understand sentence structure and sentence meaning using syntax, grammar, punctuation, parts of speech, etc.	• Understanding the sentence is critical for comprehension, and its structure is complex (for instance, inversion, relative clauses, passive voice, pronoun references).
3. Extracting Explicit information	• Match lexical and/or syntactic information in the question to those in the relevant part of the text • Identify or formulate a synonym or a paraphrase of the literal meaning of a word, phrase, or sentence in the relevant part of the text	• Information is explicitly stated at local level, usually in one sentence. •The items usually ask for specific details, and only literal understanding is necessary to answer the question.
4. Connecting And synthesizing	• Integrate, relate, or summarize the information presented in different sentences or parts of the text to generate meaning • Understand the relationship between sentences and organization of the text using cohesion and rhetorical organization knowledge • Recognize and evaluate relative importance of information in the text by distinguishing major ideas from supporting details	•The information is stated in different places of the text. • Answering the question involves connecting two or more ideas or pieces of information across sentences or paragraphs, but it is not necessary to go much beyond the text.
5. Making inferences	• Speculate beyond the text, e.g., use background/topical knowledge • Draw inferences and conclusions or form hypotheses based on information implicitly stated in the text	• Information is implicitly stated. • It is necessary to make further inferences based on other information from text and/or on background knowledge.
6. Residual skills	Including but not limited to: • Metacognitive skills (e.g., adjusting reading speed, decision to skip/skim/carefully read materials, decision to reread materials, attempt to pinpoint confusion, etc.) • Test-taking skills • Guessing	• All the skills (or strategies) not explicitly specified in the cognitive framework belong to this category. • Residual skills are affiliated with all the items, and thus it is not necessary to code.

As suggested by Pressley and Afflerbach (1995), a clear description for scoring think-aloud protocols should be established for the analysis of verbal reports. Synthesizing information from the revised cognitive framework as shown in Figure 3.1, the cognitive models suggested by Gao (2006) and Jang (2005), the classification scheme recommended by Pressley and Afflerbach, and also the students' think-aloud verbal reports, I constructed a coding scheme as shown in the above table 3.3. The think-aloud verbal reports helped to build the coding scheme which afterwards, guided the coding of the data. This coding scheme was also verified by experts as shown in Table 3.5.

I read through all the participants' verbal reports and synthesized the commonality by referring to the coding scheme. Table 3.4 is a brief snapshot of several participating students' activities while answering item 2. All four students managed to pick the correct answer but in different ways. The word *minute* (sound [mainju:t]) means *small* in the context, which was difficult for some students. For instance, Hon, a Korean student, read the word as [minit], which made me suspect that he did not understand its meaning in the context. Therefore, during the retrospective think-aloud session, I asked him what the word meant, and he said, "[minit]? I think it means *time*. No, here it is different, *minute absorption of elements*. I think it could be kind of amount. But I don't know many or less." In addition, students needed to locate the sentence containing the word *device* in order to extract the information to answering the question. As shown in Table 3.4, both Hon and Afsar did not seem to know the word *device* but simply matched the sentence containing the word *device* with the option containing the same word, whereas Jin and Katia successfully recognized the word *device* and thus were able to directly pick the right answer. To summarize, the skills identified as essential for item 2 were skill 1 (vocabulary) and skill 3 (extracting explicit information).

Table 3.4

Sample Participants' Reading Activities with Item 2

Student	Summary of reading activities
Hon	He first tried to eliminate alternative options, but he did not seem to know the word *minute* and failed to recognize its meaning. He then tried to guess *device* from the context but still failed. Finally after an extended period of time, he got the right answer by matching the sentence containing the word *device* with the options. He was not sure about his choice at all.
Jin	He read the item and then compared alternative options. He successfully located the sentence containing the answer and then picked the right answer immediately. It seems that he knew the words *minute* and *device* beforehand.
Katia	She did not know the word *minute,* but was able to guess its meaning based on a cognate in Portuguese and also English vocabulary knowledge. Then she quickly picked the right answer.
Afsar	She seemed to be quite confused by the word *device* in the item's stem. Then she went back to the passage to search for the word *device*. By matching the sentence containing the word *device* and the options, she successfully picked the right answer.

The whole process of data collection and data analysis was iterative. Coding started after the think-aloud verbal reports had been collected from the first several participants. The identified item–skill relationship was further compared across participants when more verbal reports became available. Verbal reports from the first 10 participants seemed to have yielded adequate information. Even though data were collected from three more participants, Ming, Fei, and Yao, no new information emerged regarding the skills identified for each item. Therefore, coding ceased because data saturation had been achieved.

3.2.3 Expert Rating

Participants. Four experts were invited to identify the reading skills required by each item. All were advanced doctoral students in education or applied linguistics and had rich experience teaching English reading to ESL students (please see Appendix F for the consent form). Their qualifications and experience are summarized in Table 3.5.

Table 3.5

Experts' Background Information

	Beck	Elena	Lucy	Adriana
Native language	Uzbek	Spanish	English	Spanish
Education	Master's degree in TESL, PhD candidate in applied linguistics	Master's degree in TESL, PhD candidate in applied linguistics	Master's degree in TESL, PhD candidate in educational psychology	Master's degree in educational psychology, PhD candidate in adult education
ESL teaching experience	5 years	8 years	3 years	3 years

Instrument and procedures. Each expert was provided with the four MELAB reading passages, a one-page introduction to the MELAB, a MELAB cognitive framework (see Figure 3.1), a coding scheme (see Table 3.3), and a coding form (see Appendix H). In order to acquaint the experts with the rating task, a half-hour training session was held prior to the formal rating. During the training session, the MELAB reading test was introduced to the experts, and also the cognitive model of reading and the coding scheme were presented for their review and critique. The experts reached a common understanding of the nature of the task and did not suggest any changes to either the cognitive model of reading or the coding scheme.

After training, experts read the passages and conducted the rating task independently. They identified the skills for each item and also made annotations about the evidence based on which they made the decision. When they had finished rating each passage, the experts convened and compared their ratings. Specifically, if the experts thought an item required a certain skill, they wrote 1 in the cell, otherwise 0.

Spearman rho was calculated to indicate the agreement between the ratings given by each expert. As shown in Table 3.6, the correlations between the four experts were all statistically significant at the 0.01 level. The values of spearman rho were all higher than 0.30, indicating

moderate agreement. I also observed that the experts showed more agreement as they proceeded with the rating task.

Table 3.6

Inter-Rater Agreement

	Beck	Elena	Lucy	Adriana
Beck	1.000	0.319**	0.393**	0.561**
Elina		1.000	0.396**	0.465**
Lucy			1.000	0.332**
Adriana				1.000

Note. ** Indicates significant at the 0.01 level (2-tailed).

3.2.4. Initial Q-Matrix

With reference to the coding scheme, I constructed an initial Q-matrix based on evidence from the think-aloud verbal reports and the expert ratings. However, a frequently encountered problem here is that students' verbal reports may not agree with the expert ratings (Gierl, 1997; Jang, 2005; Zappe, 2007). When this discrepancy occurred in the present study, the think-aloud verbal reports were regarded as the primary evidence, because the verbal reports more or less captured the real-time reading process and thus were regarded more reliable and authentic. The value of the expert rating, however, should not be underestimated, as it provides important evidence from a different perspective. Furthermore, when it was difficult to determine whether a certain skill should be retained for an item, the skill was usually retained. This is because the follow-up Fusion Model calibration would provide evidence concerning the importance of the skill for the item; that is, if the calibration showed the skill to be inconsequential, it could be dropped at this later point.

The initial Q-matrix for the MELAB items is shown in Table 3.7. The number 1 indicates that the skill was required by the item, whereas 0 indicates that the skill was not required by the

item. The residual skills were thought to be affiliated with all the items; their coding, therefore, is not listed in the table.

Table 3.7

Initial Q-Matrix

Item	Skill 1 (vocabulary)	Skill 2 (syntax)	Skill 3 (extracting explicit information)	Skill 4 (connecting and synthesizing)	Skill 5 (making inferences)
1	1	1	0	1	0
2	1	0	1	0	0
3	0	0	0	1	0
4	0	0	1	0	0
5	1	1	0	0	1
6	1	0	1	0	0
7	0	1	1	0	0
8	1	0	0	1	0
9	0	0	1	0	0
10	1	0	0	0	1
11	0	0	1	0	0
12	1	1	1	0	0
13	0	0	0	1	0
14	1	0	0	1	0
15	1	1	0	0	1
16	1	1	1	0	0
17	0	1	0	1	0
18	0	1	1	0	0
19	1	0	0	1	0
20	0	0	1	0	0

3.3 Empirical Validation

Response data from 2,019 examinees to each MELAB reading item were used for the empirical validation of the initial Q-matrix (Please see Appendix I for the item statistics, including the mean, standard deviation, item–total correlation, and Cronbach's alpha if item deleted.). The response data were provided by the ELI-UM MELAB via Spaan Fellowship from the ELI-UM. There were no missing data because data from examinees skipping one or more of

the items (about 3% of the total number of examinees) had been excluded. They were excluded because these examinees may have simply been guessing and thus were not instigating the processes required by item solution (Gao, 2006). The data set was analyzed with the Arpeggio software, and the following procedures were used.

3.3.1 MCMC Convergence Checking

MCMC convergence is difficult to achieve and also difficult to judge (Sinharay, 2004). In the present study, MCMC convergence was mainly evaluated by visually examining the time–series chain plots and density plots. Other criteria, such as the Heidelberg–Welch diagnostic and the Geweke Z, were also examined. The Gelman–Rubin R was not used because it has been found to be insensitive to non-convergence checking with Fusion Model calibration (Roussos, DiBello, et al., 2007). The Raftery–Lewis diagnostic was not used, as the required precision of the quantiles has to be adjusted according to the scaling of each variable (Ntzoufras, 2009), and subsequently the parameters of the resulting Fusion Model would not be on the same scale.

With the Fusion Model, MCMC chains of simulated values are generated to estimate all the parameters. Each time point (or step) in the chain corresponds to a set of simulated values for the parameters. After a sufficient number of steps, i.e., the burn-in phase of the chain, the remaining simulated values approximate the desired Bayesian posterior distribution of the parameters. Typically, the results of the initial thousands of steps or values are thrown out, and these thrown-out values are called those of the "burn-in" period. A critical issue in implementing the MCMC is to determine the number of steps or runs until the Markov chain converges to the posterior stationary distribution. As suggested in the Arpeggio manual, in the present study, two Markov chains were run with a chain length of 40,000 with burn-in steps of 20,000. For comparison, a chain length of 60,000 with burn-in steps of 30,000 and an extremely long chain

length of 100,000 with burn-in steps of 50,000 were also run in order to rule out the possibility of an insufficient chain length.

It was found that the convergence of chain length of 60,000 was better than that of a chain length of 40,000 for many parameters, whereas a chain length of 100,000 did not noticeably improve the convergence. An example is illustrated in Figure 3.2, when every 10th observation was used to draw the time–series chain plots and density plots. With the chain length of 40,000, the chain plots showed jumping, which indicates a lack of convergence. With the longer chain length of 60,000, the chain plots were stabilized, which indicates good convergence. Running an extremely long chain of 100,000 did not seem to be necessary. In general, a chain length of 60,000 with burn-in steps of 30,000 was found to be appropriate.

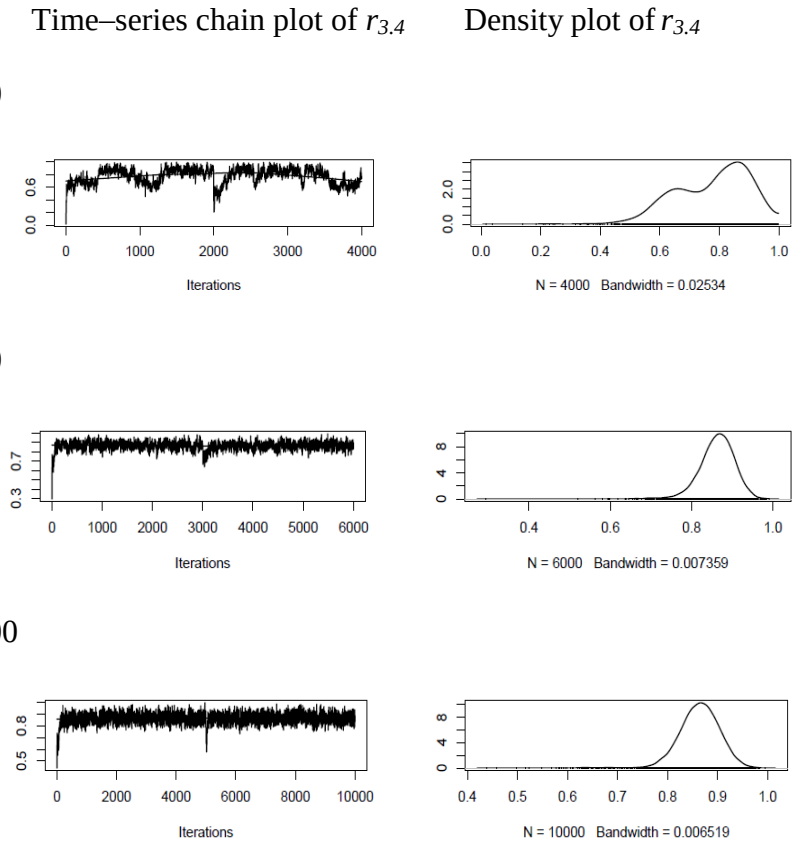

Figure 3.2. Time–series chain plots and density plots of $r_{3.4}$ with different chain lengths.

With a chain length of 60,000, the majority of parameters achieved excellent convergence. However, the time–series chain plots and density plots for some parameters, such as p_{k5} (proportion of masters of skill 5 in the population), $r_{5.1}$ (diagnostic capacity of item 5 to skill 1), $r_{5.5,}$ $r_{8.1,}$ $r_{10.1,}$ $r_{10.5,}$ $r_{15.1,}$ $r_{15.2}$, and $r_{19.1}$, showed moderate fluctuation. As shown by the examples in Figure 3.3, the time–series chain plots for $r_{5.1}$ showed some fluctuations that may indicate non-convergence, whereas the time–series chain plots of $r_{4.3}$ were smooth and stable, indicating excellent convergence.

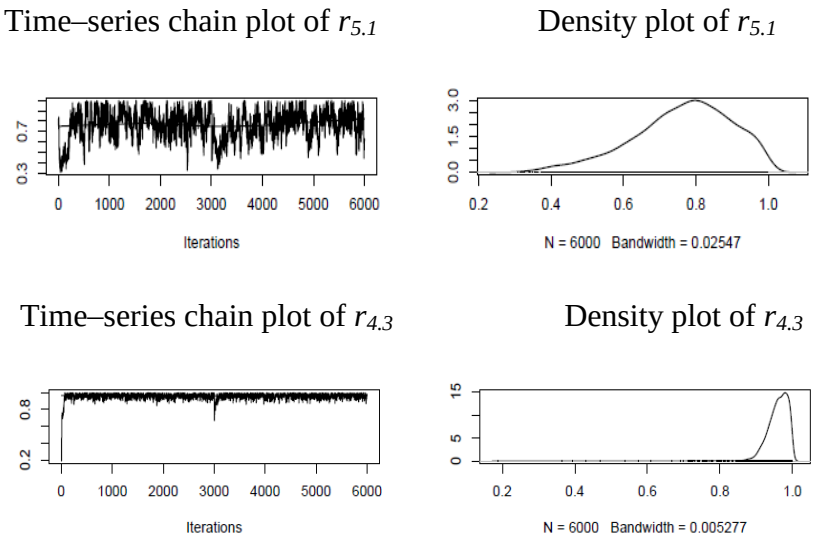

Figure 3.3. Time–series chain plots and density plots of $r_{5.1}$ and $r_{4.3}$.

Some numerical criteria were also used to help judge MCMC convergence. As shown in Table 3.8, the Heidelberg–Welch diagnostic indicated that all the parameters except $c_{8.1}$ achieved good convergence. However, the Geweke Z showed that 18 of the 79 parameters had a z value out of the range of -2 and 2, indicating non-convergence. Among those parameters, eight had z values out of the range of -3 and 3.

The potential non-convergence of p_{k5} was worrisome, because the p_k parameters are one of the priors in the Fusion Model calibration and thus the non-convergence of p_{k5} may influence

the estimation of other parameters. It is also important to note that many of the potentially problematic parameters here are affiliated with skill 5 (making inferences) or items that require skill 5, namely items 5, 10, and 15. In conclusion, the overall MCMC convergence for all parameters was acceptable but not excellent, and the validity of explicating skill 5 from the MELAB reading test seems to require further examination.

Table 3.8

Summary of MCMC Convergence Check

	Criteria	Problematic parameters		
Time–series chain plots and density plots	Obvious trends indicate non-convergence	$p_{k5,}$ $r_{5.1,}$ $r_{5.5,}$ $r_{8.1,}$ $r_{10.1,}$ $r_{10.5,}$ $r_{15.1,}$ $r_{15.2,}$ $r_{19.1}$		
Heidelberg–Welch Diagnostic	$p < 0.05$ indicates non-convergence	$c_{8.1}$ ($p = 0.0475$)		
Geweke Z	$\|z\| > 2$ indicates non-convergence	p_{k5} *(z = 6.09)*	$\pi_{5.1}$ (z = - 3.75)	$\pi_{10.1}$ (z = - 3.18)
		$r_{2.1}$ (z = 2.22)	$r_{2.3}$ (z = - 2.26)	$r_{5.1}$ (z = - 3.1)
		$r_{5.5}$ (z = 2.56)	$r_{10.1}$ (z = - 4.6)	$r_{12.1}$ (z = 2.77)
		$r_{12.3}$ (z = - 2.47)	$r_{14.1}$ (z = 4.05)	$r_{14.4}$ (z = -2.19)
		$r_{15.1}$ (z = 2.16)	$r_{15.5}$ (z = - 3.24)	$r_{20.3}$ (z = 3.16)
		$c_{14.1}$ (z = - 2.56)	$c_{16.1}$ (z = - 2.23)	$c_{20.1}$ (z = - 2.75)

3.3.2 Refining the Initial Q-Matrix

In the initial Q-matrix, only three items were assigned to skill 5, whereas 11 items were assigned to skill 1, 8 items to skill 2, 10 items to skill 3; and 7 to skill 4 (see Table 3.7 for details). Thus, the information pertaining to skill 5 was probably insufficient for estimation.

It has been recommended that an *r* parameter bigger than 0.9 should be removed from the Q-matrix, as the affiliated skill is not significantly important for the item (Hartz, 2002). When the item parameters were examined, it was found that $r_{15.5}$, i.e., the discrimination capacity of item 15 to skill 5, was 0.913. The stem of item 15 (*One can infer from the passage that*) uses the word *infer*, indicating that item 15 is about making inferences. However, upon closer

examination of the item and also the think-aloud verbal reports, it was found that items 5 and 10 required speculating considerably beyond the passage; yet, the answer to item 15 was in fact embedded in different places of the text, despite the use of the word *infer* in the item stem. Item 15 was, therefore, reclassified as requiring skill 4 (synthesizing and connecting) because information from different places of the text was needed to answer it.

After item 15 had been reassigned to skill 4, only items 5 and 10 required skill 5. This resulted in too little information for the Fusion Model to estimate skill 5–related parameters, as at least three items for a certain skill have been recommended for Fusion Model calibration (Hartz, 2002). Therefore, it was decided that skill 5 should be dropped from the Q-matrix.

High values of *r* and *c* are indicative of a possibility for model simplification (Hartz, 2002; Roussos, DiBello, et al., 2007). If *c* is, say, bigger than 2, this indicates that the skills required to successfully answer the item are completely specified by the Q-matrix, and thus *c* can be dropped. However, whether to drop a certain Q-matrix entry depends on both statistical criteria and substantive knowledge. First, the six large *c* parameters were dropped from the Q-matrix one at a time, as they did not drastically change the Q-matrix structure. Then four of the large *r* parameters, namely $r_{1.2}$, $r_{2.3}$, $r_{7.2}$, and $r_{12.3}$ were dropped from the Q-matrix one at a time. The remaining three large *r* parameters, namely $r_{4.3}$, $r_{9.3}$, and $r_{13.4}$ were kept because their affiliated skill was the only skill identified for the item.

The convergence of the Fusion Model calibration using the Q-matrix thus refined was reevaluated. The time–series chain plots and density plots of the parameters did not show noticeable trends or fluctuation. All the parameters met the Heidelberg–Welch diagnostic and Geweke Z convergence criteria. Therefore, after skill 5 had been removed, the current Fusion Model calibration achieved excellent convergence.

3.3.3 Model Fit

There are two main approaches to assessing model fit with the Fusion Model: comparing the model-predicted values to the observed values and evaluating the characteristics of the skill mastery classification. In the following, the model fit of using the initial Q-matrix and the refined Q-matrix were compared based on different evidence. However, for most of the model-fit judgment discussed below, there are no commonly agreed cut-off criteria, and thus only descriptive model fit evidence is presented.

Observed Versus Predicted *P*-Values across Items. The first index is the residual between the observed and model-predicted *p*-value across items. A *p*-value refers to the proportion of examinees who respond correctly to the item. The predicted *p*-value of each item was derived based on the result of the Fusion Model calibration.

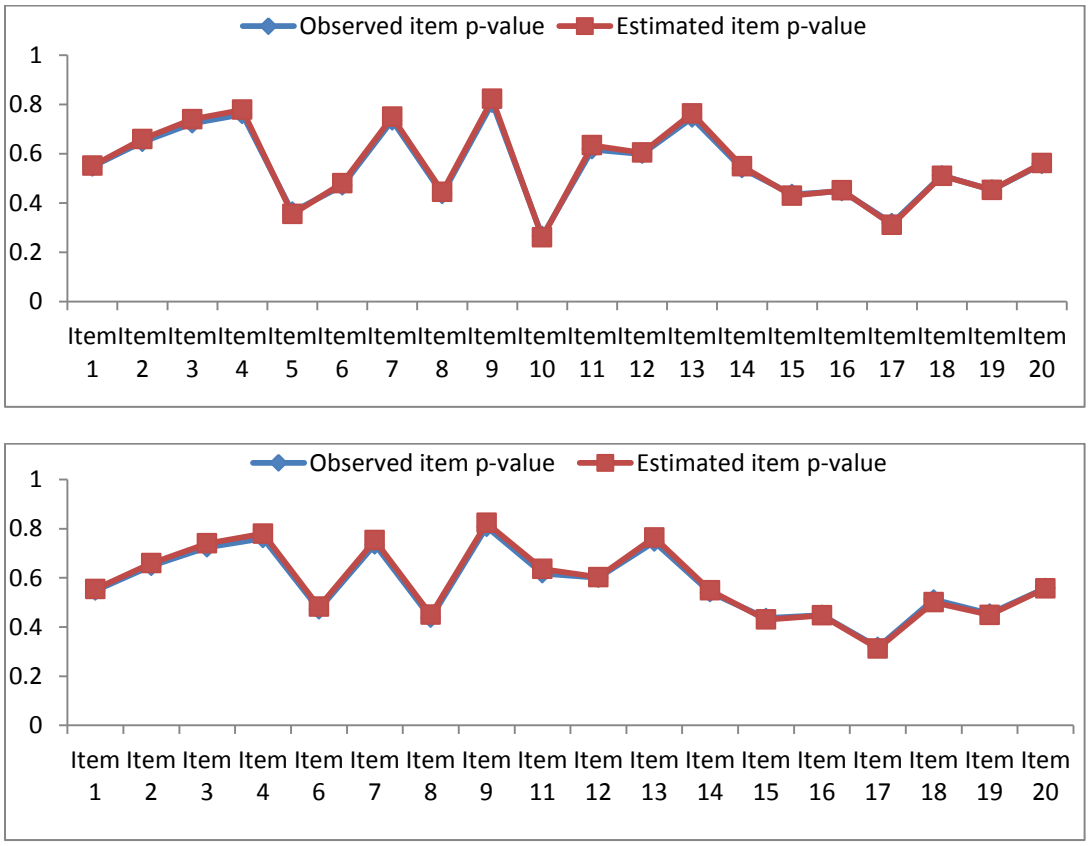

Figure 3.4. Observed versus predicted *p*-values across items.

The chart at the top of Figure 3.4 shows the observed *p*-value versus the estimated *p*-value for each item when the initial Q-matrix was used, whereas the chart at the bottom shows the observed *p*-value versus the estimated *p*-value when the refined Q-matrix was used. The two lines were very close or overlapped for most of the items. Table 3.9 also shows that the mean and mean square error of the difference between the observed and predicted *p*-value were negligible. This small difference provides evidence for good model fit.

Table 3.9

Comparison of Observed and Predicted P-Values across Items

Difference between observed and predicted *p*-values	Initial Q-matrix	Refined Q-matrix
Mean	-0.006	-0.007
Mean square error	0.000	0.000

Observed Versus Predicted Total Scores across Examinees. The observed and predicted total scores across examinees were also compared to further judge model fit. The observed total scores were calculated by adding up all the item scores for each examinee, whereas the predicted total scores were provided as a result of the Fusion Model calibration.

Table 3.10

Comparison of Observed and Predicted Total Scores across Examinees

Difference between observed and predicted total scores	Initial Q-matrix	Refined Q-matrix
Mean	0.000	0.000
Mean square error	0.070	0.085

As shown in Table 3.10, when both the observed and predicted total scores for the individual examinees were standardized as z-scores, the mean of the difference between the observed and predicted total scores was zero whether the initial or the refined Q-matrix was

used. The mean square error was a little bit smaller when the initial Q-matrix was used, probably because the initial Q-matrix had more parameters than did the refined one. In general, the difference between the observed and predicted total scores was not big.

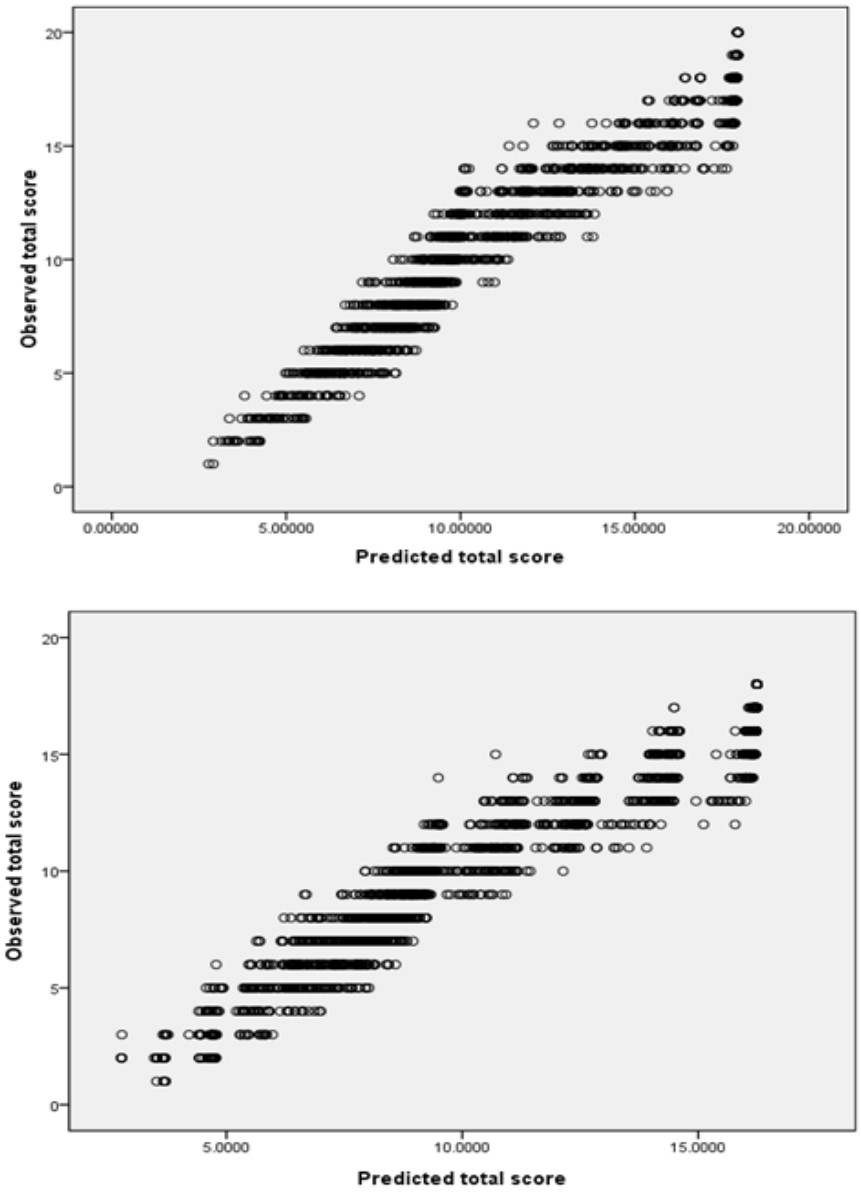

Figure 3.5. Scatter plots of the observed and predicted total scores.

The scatter plots of the observed and predicted total scores for all 2,019 examinees are shown in Figure 3.5. The top chart refers to the initial Q-matrix, and the bottom chart refers to the refined Q-matrix. The observed and predicted total scores correlated very well in both the

charts. When the initial Q-matrix was used, the correlation between the observed and predicted total scores was 0.960, and the correlation was almost the same, 0.957, when the refined Q-matrix was used. However, both charts indicate that examinees at the higher end appeared to have been underestimated in terms of their total scores. This misfit has also been observed in previous studies (e.g., Jang, 2005; Romàn, 2009), as the categorical CDMs may overestimate the scores for the lowest-scoring examinees and underestimate the scores of the highest-scoring examinees. Because the purpose of the Fusion Model calibration is to estimate categorical skill mastery status, the slight underestimation of total scores at the higher end may not substantively influence the classification results (Roussos, DiBello, et al., 2007).

Item Mastery Statistics. IMstats computes the observed proportion-correct scores for item masters and item non-masters on an item-by-item basis. An item master is an examinee who has mastered all the skills required by the item, and an item non-master is an examinee who has not mastered at least one of the skills required by the item. Informally, a substantial difference between the proportion-correct scores of these two groups indicates a high degree of model fit or internal consistency, as the membership of item masters or non-masters is based on the examinee's skill classification. Therefore, IMstats is also used as internal validity evidence.

In Figure 3.6, the top chart shows the proportion-correct scores of item masters and non-masters when the initial Q-matrix was used, and the bottom chart shows the proportion-correct scores of item masters and non-masters when the refined Q-matrix was used. Despite a lack of consensus on the criteria according to which the difference should be measured, both charts show substantial difference between the proportion-correct scores of the item masters and those of the non-masters. As indicated in Table 3.11, the average proportion-correct scores of the item masters were over 0.9 in both cases, whereas the average proportion-correct scores of the item

non-masters were less than 0.45. To summarize, the differences, as shown in Figure 3.6 and Table 3.11, provided important evidence for good model fit.

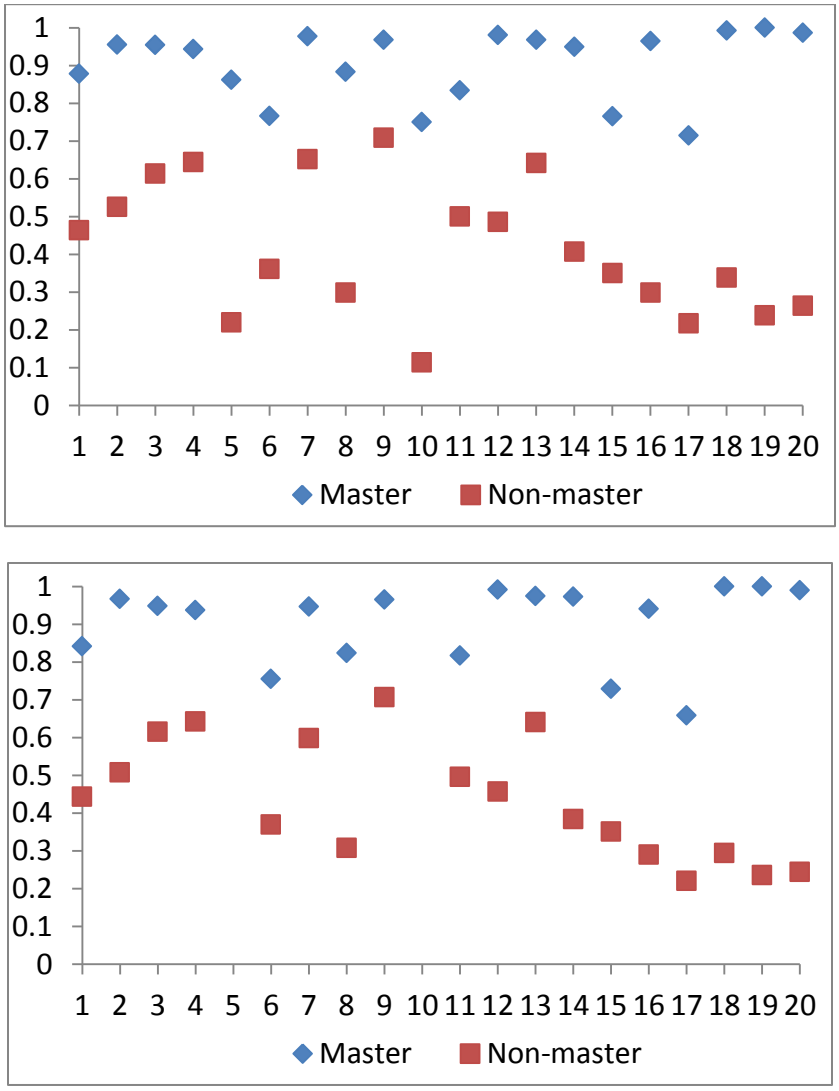

Figure 3.6. Proportion-correct scores of item masters and non-masters.

Table 3.11

Comparison of Average Proportion-Correct Scores of Item Masters and Non-Masters

Statistic	Initial Q-matrix	Refined Q-matrix
Mean proportion-correct score of item masters	0.905	0.904
Mean proportion-correct score of item non-masters	0.417	0.434

3.3.4 Final Q-Matrix

Based on the model-fit evidence presented in the previous section, the model fit the data reasonably well regardless of whether the initial Q-matrix or the refined Q-matrix was used. In other words, when the more parsimonious refined Q-matrix was used, the model fit was not noticeably worse. For the sake of model parsimony, the refined Q-matrix, therefore, was adopted as the final Q-matrix for the MELAB reading test.

The item parameters are shown in Table 3.12, and the shaded cells indicate the entries or parameters that have been dropped. The remaining cells describe the item parameters that give detailed information about the cognitive structure and the diagnostic capacity of the MELAB reading test.

Table 3.12

Item Parameters of the Final Calibration

Item	π*	r*1	r*2	r*3	r*4	r*5	c
1	0.842	0.732			0.843		1.647
2	0.985	0.779					1.028
3	0.971				0.862		1.280
4	0.993			0.959			1.116
5							
6	0.765	0.862		0.653			1.397
7	0.983			0.853			1.275
8	0.872	0.689			0.763		0.838
9	0.985			0.918			1.528
10							
11	0.868			0.854			1.077
12	0.973	0.654	0.878				1.547
13	0.995				0.911		1.165
14	0.985	0.671			0.891		0.893
15	0.727	0.710	0.748		0.827		
16	0.920	0.898	0.631	0.432			
17	0.616		0.417		0.748		
18	0.969		0.424	0.727			
19	0.975	0.879			0.249		
20	0.958			0.311			

The π parameter is the probability that an examinee, having mastered all the Q-matrix-required skills for item *i,* will correctly apply all these skills to solving item *i*. The average π parameter in Table 3.12 was 0.910, indicating that the identified skills for the items were generally adequate and reasonable. However, the π parameter for item 17 was as low as 0.616. This indicates that the probability of an examinee correctly answering item 17 was only 0.616, given that he/she had acquired the required skills of syntax and connecting and synthesizing information. Item 17 was a rather difficult item. As shown in Appendix I, the proportion-correct score for item 17 was only 0.318, whereas the average proportion-correct score across all the items was 0.550. This probably explains why the π parameter for item 17 was low. In general, the overall values of the π parameters are reasonable and satisfactory regarding the quality of the Q-matrix.

The *r* parameter is an indicator of the diagnostic capacity of item *i* for skill *k,* ranging from 0 to 1. The more strongly the item requires mastery of skill *k,* the lower is *r*. The *r* parameters, as shown in Table 3.12, were generally large, indicating that the diagnostic capacity of the MELAB reading test is low. For instance, $r_{7,3}$ was 0.853. This indicates that the probability of correctly answering item 7 when skill 3 (extracting explicit information) has not been mastered is 0.853 times the probability of correctly answering item 7 when skill 3 has been mastered. In other words, it does not matter much whether examinees have mastered skill 3 or not. As shown in Appendix I, item 7 was a rather easy item with a proportion-correct score of 0.736. This is probably why its diagnostic capacity was limited. Overall, the MELAB reading test is an English proficiency test that is not designed for cognitive diagnostic purposes, which may explain why its diagnostic capacity is not very high.

The *c* parameter is an indicator of the degree to which the item response function relies

on skills other than those assigned by the Q-matrix. The lower the *c*, the more the item depends on the residual ability. Some researchers (e.g., Jang, 2005; Roussos, DiBello, et al., 2007) have reported that when *c* parameters are included, the residual part of $p_{ci}(\theta_j)$ might dominate the model. If that occurs, most of the p_k parameters will be very large, which artificially makes nearly everyone a master of most of the skills. In addition, the *c* parameters themselves sometimes cannot converge. This, however, was not found to be the case in the present study. All the p_{ks} were less than 0.5, which indicates that fewer than half the examinees were masters of the skills. Also, all the *c* parameters had good convergence. The only concern is that *r* parameters were generally large. In order to examine whether this was because the *c* parameters had "soaked up" the variance, the Fusion Model was run with all *c* parameters fixed. It was found that the convergence was poor when *c* was fixed. And, the values of the *r* parameters were not noticeably smaller as a result of fixing *c*. In addition, the cognitive framework built for the MELAB reading test involves a residual part. Therefore, keeping the *c* parameter and using the full Fusion Model is statistically and theoretically sound. As a result, only six large *c* parameters were dropped for model parsimony, while the rest of the *c* parameters were maintained in the Q-matrix.

3.3.5 Calibration Results

With the recommended Q-matrix and the item response data of the 2,019 examinees, the Fusion Model calibration was conducted using Arpeggio. The calibration results are as follows.

Continuous Posterior Probability of Mastery (PPM) indicates the probability that an examinee is a master of the skill being studied. As shown in Figure 3.7, most of the examinees had either a very high or very low PPM, so that they could easily be classified as masters or non-masters of the skills. The mean PPM for skill 1 (vocabulary) was 0.31, which indicated that on average the probability that an examinee would be a master of skill 1 was 0.31. The mean PPMs

for skill 2 (syntax), skill 3 (extracting explicit information), and skill 4 (connecting and synthesizing) were 0.33, 0.40, and 0.34, respectively.

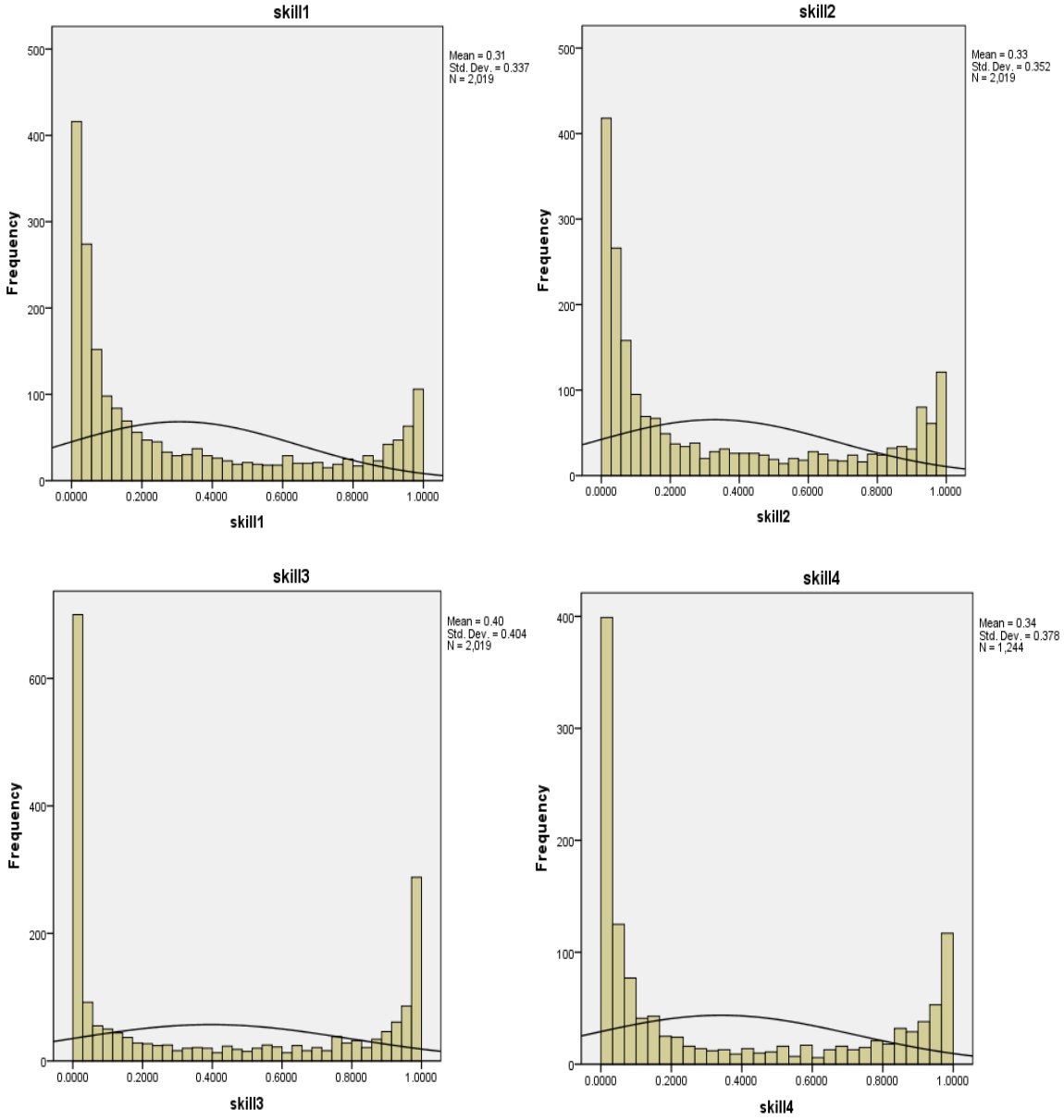

Figure 3.7. Continuous posterior probability of mastery *(PPM).*

In addition, a dichotomous mastery/non-mastery status can be accomplished by using a cut-off point of 0.5 (Hartz, 2002). If PPM > 0.5, the examinee is a master of the skill, and if PPM < 0.5, the examinee is a non-master of the skill. As shown in Figure 3.8, about 26.8% of

examinees were masters of skill 1 (vocabulary), 29.4% were masters of skill 2 (syntax), 39.6% were masters of skill 3 (extracting explicit information), and 32.2% were masters of skill 4 (connecting and synthesizing). Sometimes, a more refined polytomous status can be determined by using 0.4 and 0.6 as cut-off points (Jang, 2005). However, in the present study, less than 7% of examinees had PPMs between 0.4 and 0.6 for all four skills; therefore, a more refined polytomous status would not have changed the classification results much.

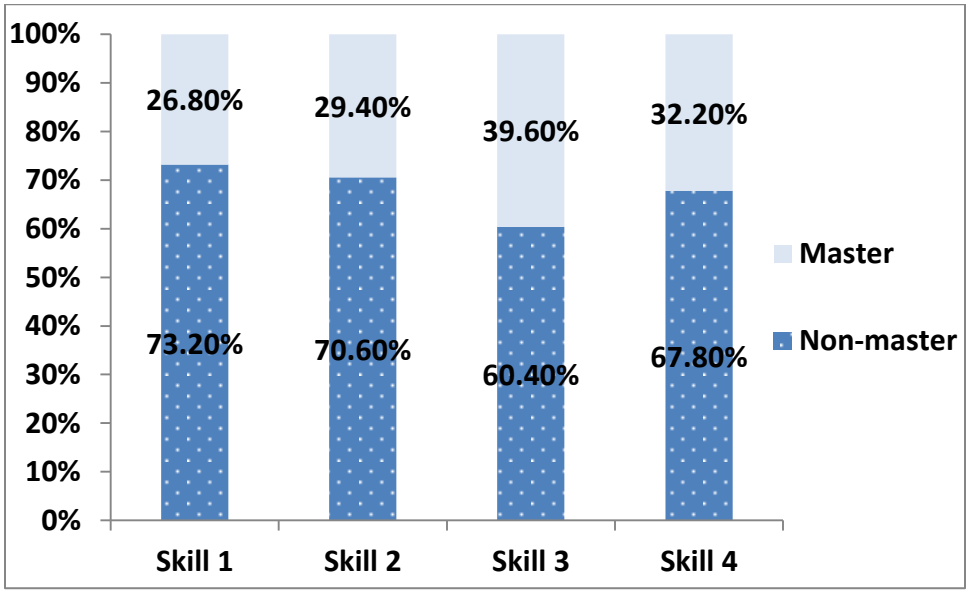

Figure 3.8. Categorical skill mastery status.

These results are relatively congruent with my expectations. First, it seems that the MELAB reading test is rather difficult, as fewer than half the examinees were found to be masters of each skill. This was to be expected as the average total score in the overall dataset was only 11 out of 20 (please see Appendix I for item statistics). Second, skill 1 (vocabulary) seemed to be the most difficult skill for the examinees, which agreed with the consistent finding that lack of vocabulary is the major obstacle in reading comprehension (Grabe, 2009). Third, in accord with the literature on reading, the present study found that skill 4 was more difficult than skill 3. Skill 3 is that of extracting explicitly stated information at a local level, whereas skill 4 refers to

connecting and synthesizing information from different places of the text. Thus, skill 4 required more cognitive processes and was more challenging than skill 3.

With four skills involved in the test, examinees could have as many as 16 (i.e., 2^4) skill profile patterns, as shown in Table 3.13, in which 1 indicates mastery of the skill and 0 indicates non-mastery of the skill. For instance, a skill profile of 0101 indicates that the examinee is a non-master of skill 1 (vocabulary), a master of skill 2 (syntax), a non-master of skill 3 (extracting explicit information), and a master of skill 4 (connecting and synthesizing). As illustrated in Table 3.13, about half the examinees were non-masters of all four skills, i.e., their profiles showed 0000, and about 17% were masters of all four skills, i.e., their profiles showed 1111. The third largest number of examinees had a profile of 0010, indicating that they were only masters of skill 3 (extracting explicit information). This was to be expected, as skill 3 was found to be the least challenging in the think-aloud protocols.

Table 3.13.

Skill Mastery Patterns

Skill mastery pattern	Frequency	Percentage
0000	1069	52.95%
1111	352	17.43%
0010	124	6.14%
0111	76	3.76%
0011	66	3.27%
1011	60	2.97%
0110	51	2.53%
0001	48	2.38%
1110	46	2.28%
0100	25	1.24%
1010	24	1.19%
1000	20	0.99%
0101	19	0.94%
1001	15	0.74%
1101	15	0.74%
1100	9	0.45%

3.4 Discussion

Successful cognitive diagnostic modeling depends to a large extent on the robustness of the Q-matrix, and a sound Q-matrix relies on evidence from multiple sources (Jang, 2009). Qualitative methods such as the think-aloud protocol and expert rating have proven very useful in understanding examinees' cognitive processes (Gierl, Alves, Roberts, & Gotzmann, 2009; Leighton & Gierl, 2007). A sequential combination of the two sources was adopted in this study. Students' verbal reports seemed to be especially important at the initial stage when the construct of reading needed to be deeply understood and when the subskills were identified and categorized. After the initial exploration into the cognitive structure, experts' judgment was very helpful in cross-validating the initial coding. Experts were also important in critiquing and auditing the coding process. In general, it seemed very helpful to use both the think-aloud protocol and expert rating for Q-matrix construction.

The initial Q-matrix was further validated with the Fusion Model calibration process. Statistical evidence, such as convergence, fit indices, and parameter values, gave clues regarding how the initial Q-matrix could be modified. However, as with any other model-modification procedure, blindly following statistical evidence may compromise the validity of the Q-matrix. For instance, Alderson (2010) criticized Sawaki, Kim, and Gentile's study (2009) for not including vocabulary as a subskill in its Q-matrix for the TOEFL listening test. Sawaki et al. made the decision not to include vocabulary as a subskill based on model comparison. Specifically, they compared the results of using three different Q-matrices for Fusion Model calibration, one with vocabulary keyed to each item, one without the subskill of vocabulary at all, and one with vocabulary only keyed to the items as suggested by experts. The outcomes of using the three different Q-matrices were very similar, and thus they decided to drop vocabulary from

the entire Q-matrix for model parsimony. As observed by Alderson, vocabulary is an important language skill needed for the TOEFL listening test, and excluding such an essential skill based solely on statistical concerns risks losing the meaning of the cognitive diagnostic analysis.

Both substantive and statistical factors were considered in this study. However, a question, as raised by Jang (2005), is how much should be determined by substantive knowledge and how much should be determined by statistical modeling. Though there is no definite answer to this question, in this dissertation substantive knowledge and statistical evidence agreed more than they disagreed. Whenever the Fusion Model calibration gave a suspicious signal, the substantive knowledge of reading could help to identify and interpret the potential issue. Therefore, it is very important that Q-matrix modification decisions based on statistical modeling receive substantive support.

CHAPTER 4

HYPOTHESES GENERATION ON READING SUBSKILL DIFFERENCES –A GROUNDED THEORY STUDY

The present study compares the reading subskills between East Asian and Romance ESL learners as represented in the MELAB reading test. However, there is no sufficient theoretical and empirical evidence on how exactly the two groups are different at the subskill level. This chapter explains the process of establishing hypotheses regarding how East Asian and Romance ESL learners may differ on those subskills with a grounded theory approach.

4.1. A Grounded Theory Approach

ESL learners from different native language groups may show different patterns in their reading processes and skills (Koda, 2005). The particular teaching and learning styles in East Asian countries may also shape their ESL learners' reading in different ways than do the particular teaching and learning styles experienced by ESL learners with a Romance language background. However, there is insufficient evidence regarding exactly how these groups differ at the subskill level.

A grounded theory approach is an appropriate method for exploring this issue. First developed by Glaser and Strauss (1967), grounded theory has been widely used to build theory from data when the theory is not available or is insufficient. Generally speaking, the grounded theory approach comprises reading (and re-reading) a textual database (such as field notes and interview transcripts) and "discovering" or labeling variables (e.g., categories, concepts, and properties) and their interrelationships (Bryant & Charmaz, 2007; Charmaz, 2006; Glaser 1992; Strauss & Corbin, 1990, 1998). This approach has the advantage of being systematic and

creative, identifying, developing, and relating concepts that are the building blocks of theory (Strauss & Corbin, 1998, p. 13).

However, there is much variation and disagreement in the application of the grounded theory approach (Charmaz, 2006). Though Glaser and Strauss initially developed grounded theory together, they subsequently disagreed about how to actually carry out a grounded theory study. Glaser has approached grounded theory based on the purist view that researchers should be naïve of the phenomenon, without any preconceived expectations, conceptual framework, or prejudices. The researcher must have an open attitude to the research question, so that the generation of theory is not compromised by a researcher's preexisting views but directly emerges from the data. Therefore, Glaser has insisted that the researcher should not review the literature until the emerging theory has developed sufficiently based on the data. In contrast, Strauss and Corbin (1990, 1998) have taken a more pragmatic approach. They have advocated reviewing literature for five reasons: (a) to stimulate theoretical sensitivity, as the literature can provide concepts and relationships that can be checked against actual data; (b) as the secondary source of data; (c) to stimulate questions; (d) to direct theoretical sampling, as ideas may arise that suggest where to go next to uncover phenomena important to the development of emerging theory; and (e) as supplementary validation: the researcher can refer to the appropriate literature to validate the accuracy of findings. Both these approaches have pros and cons, and Glaser (1998) has referred to this split of methodology as a "rhetorical wrestle."

In this study, a literature review was only conducted on general topics such as the theoretical framework of reading skills transfer from L1 to L2, and English instruction and assessment in East Asian countries. This approach was taken based on the idea that a preliminary review of some general findings helps to build the researcher's theoretical sensitivity and also

provides justification for the study (Strauss & Corbin, 1990). A more extensive literature review was conducted after the data analysis in order to validate the generated hypotheses.

4.2 Literature Review

4.2.1 Theoretical Framework of Reading Skills Transfer from L1 to L2

L2 reading shares a similar process with L1 reading but is also unique in many ways (Grabe, 2009). Alderson (1984) raised the often-asked question of whether poor L2 reading is a reading problem or a language problem. This inquiry considers two primary premises: (a) poor reading in a foreign language is due to poor reading ability in the first language, and (b) poor reading in a foreign language is due to inadequate knowledge of the target language (Koda, 2005). The following two hypotheses underlie Alderson's above-mentioned speculations.

According to the Developmental Interdependence Hypothesis (Cummins, 1979, 1986), reading performance in a second language is largely shared with reading ability in a first language—a position that has been supported by many empirical studies. For example, it has been shown that school-age English learners' L1 and L2 reading abilities are highly correlated (Cummins & Mulcahy, 1978; Legarretta, 1979; Troike, 1978). Some researchers supporting this view have argued that higher-level processing skills may be transferred to a second language and may in fact compensate for inadequacies in lower-level linguistic skills (Coady, 1979; Hudson, 1982). Similar arguments are that the process of learning to read and the resulting ability to read is undergone once, and that the awareness of the reading process can be transferred to L2 reading; thus, the reading process does not need to be learned again in acquiring another language (Gamez, 1979; Goodman, Goodman, & Flores, 1979; Rigg, 1977).

The Linguistic Threshold Hypothesis (Clarke, 1980; Cziko, 1980; Yorio, 1971), on the

other hand, contends that in order to read in a second language, a level of second-language linguistic abilities must first be achieved. According to this view, first-language reading skills cannot help learners read text in a second language. Empirical studies have also shown that L2 knowledge explains 30 to 40% of the variance in L2 reading scores (Bernhardt & Kamil, 1995; Bossers, 1991; Carrell, 1991). Thus, limited L2 knowledge does compromise ESL learners' ability to use their L1 skills.

Without referring to the characteristics of the L2 readers and the context, it is probably not meaningful to claim that one hypothesis is more correct than the other. Each hypothesis explains L2 reading and the transfer of L1 to L2 from a distinct perspective. The Developmental Interdependence Hypothesis is more successful in explaining the reading of young ESL learners, whose L1 literacy is still emerging, though it has been cautioned that other background factors may confound the L2 reading variance explained by L2 knowledge (Koda, 2005). A general rule is that the extent to which the L1 is similar to the L2 is an important factor influencing the transfer of reading skills. The more similar the L1 and L2 are, the easier it is for the transfer to occur.

4.2.2 English Instruction and Assessment in East Asian Countries

Although communicative language teaching (Savignon & Berns, 1984; Swan 1985a, 1985b) has been advocated for many years, the teaching of English in East Asian countries, such as China, Japan, and Korea, is still greatly influenced by a teacher-centered, grammar-translation method (Liu & Littlewood, 1997). Depicted as quiet and shy, East Asian students take copious notes and seldom ask questions or participate in discussions (Cortazzi & Jin, 1996; Jones, 1999; Rao, 2001). Teachers are regarded as the ultimate authority, and knowledge is transmitted by the teacher rather than discovered by the learner. Moreover, instruction focuses primarily on

vocabulary and grammar. In general, East Asian ESL students' English learning emphasizes rote memorization (Rao, 2001). For example, in a study by Hu (2001, cited in Zhong 2006), many Chinese students memorized all the words in the vocabulary handbook in order to prepare for English exams; a few even tried to memorize words in a Chinese–English Dictionary.

The teaching and learning of English in East Asian countries is also intensively test-oriented. The Chinese civil service exam, which started around the year 606 and officially ended in 1905, has exerted a great influence, too, on neighboring countries such as Korea, Japan, and Vietnam (Suen & Yu, 2006).The current practice of English language assessment in three East Asian countries, namely, China, Korea, and Japan is briefly introduced as follows.

In China, English is a compulsory subject in the national college entrance examination for all types of universities and colleges. At most universities, students have to show good performance on the College English Test (CET) in order to receive their academic degrees. The national graduate school entrance examination also requires an assessment of English ability. Apart from English as an academic requirement, English skills are tested for all those seeking employment or promotion in governmental, educational, scientific research, medical, financial, business, and other government-supported institutions (He, 2001). According to Cheng (2008), excelling in English tests in China is the key to success in Chinese society. Therefore, many English learners in China learn English not for using the language but for passing the tests to foster their success in Chinese society.

The role of English tests is similar in Korea. Secondary school students invest tremendous amounts of time preparing for English tests in order to gain admittance to universities. Even after graduation, a student seeking employment may still need to submit his/her Test of English for International Communication (TOEIC) score report to companies as

part of the application process. Choi (2008) cautioned that a large number of English learners in Korea are under great pressure to perform well on English tests and that the whole education system as it pertains to teaching English puts a great emphasis on test preparation. As a result, many Korean ESL learners are experienced at taking English tests, but may not be proficient with English language skills.

English-language tests are of great importance in Japan as well. Entrance into universities in Japan has been described as a hierarchical system of exam halls (Cutts, 1997; Poole, 2003; Yoneyama, 1999). For instance, in January 2006, 492,586 students, 40% of the high school graduates that year, took the English test given by the National Center for University Entrance Examination in Japan, as 60% of the universities in Japan required a student to pass this English test for admission (Sasaki, 2008). The standardized English tests have propelled English learning in Japan but have also brought negative effects.

To summarize, tests play an important role in English teaching and learning in East Asian countries (Ross, 2008). It is expected that the particular educational and test-taking experiences of ESL learners in East Asian countries and also their native languages may make them distinct from learners with a Romance language background regarding reading skills. A grounded theory study was thus conducted to explore the differences based on think-aloud verbal reports from ESL learners in both groups.

4.3 Methods

4.3.1 Sampling and Participants

To implement the grounded theory approach to generating research hypotheses, some of the data collected via the think-aloud protocols during the Q-matrix development stage of this

dissertation were re-examined. Further, the data were supplemented with think-aloud protocols from additional participants. The general orientation was one in which "the analyst jointly collects, codes and analyzes his data and decides what data to collect next and where to find them, in order to develop his theory as it emerges" (Glaser & Strauss, 1967, p. 45). Data collection ceased when saturation had been achieved, i.e., when no more new themes were emerging.

As the overall research question was to compare the reading skills of East Asian and Romance ESL learners, these two groups were especially targeted in the initial data collection. Participants were mainly recruited from the Mid-State Literacy Council in State College, Pennsylvania, and from a Level-2 ESL class in the Intensive English Communication Program (IECP) at the Pennsylvania State University. In April 2010, data were collected from 10 participants, namely, Ted, Chika, Hon, Jin, Katia, Dora, Leon, Eva, Sabina, and Afsar. After some initial data analyses, I found that more East Asian participants were needed. Thus, data were collected from Yao, Fei, and Ming in June 2010. However, for the final analysis, Sabina was excluded due to her high reading ability, and Afsar was also excluded as her native language was Persian. Furthermore, Ming was excluded from the final analysis because he produced few usable verbal reports. To summarize, data from 10 participants were examined for the purpose of generating hypotheses. The participants' demographic information is shown in Table 4.1.

One important factor was participants' English reading ability, as the research question focused on comparing the two groups' reading skills when their overall English reading ability was controlled for. A background information sheet on their English learning experiences and scores from the TOEFL were collected (see Appendix D). However, less than 1/3 of the participants had taken the TOEFL, and thus the TOEFL score was not sufficient as a sole

criterion. Another source of information, albeit post hoc, was the participants' performances in the MELAB reading test as used in the think-aloud activity. Although assessing their English reading ability was not the goal of the think-aloud activity, the number of items they answered correctly could be used as an indicator of their English reading ability. Their self-evaluation of English reading ability was also used. A final approach was to refer to their overall profiles, including their TOEFL scores (if any), their performances on the MELAB test, their English learning backgrounds, and their self-evaluations. As it was not possible to match them on an individual basis, participants were grouped into high and low levels of English reading ability based on their overall profiles, as shown in Table 4.1.

Table 4.1

Participants' Background Information

Language group	Reading ability	Name	Gender	Native language	Native country
East Asian	High	Ted	M	Chinese	China
		Yao	F	Chinese	China
		Chika	F	Japanese	Japan
	Low	Fei	F	Chinese	China
		Hon	M	Korean	Korea
		Jin	M	Chinese	China
Romance	High	Katia	F	Portuguese	Brazil
		Dora	F	French	Morocco
	Low	Leon	M	Spanish	Colombia
		Eva	F	Spanish	Spain

4.3.2 Data Collection

The data used to generate hypotheses were the students' think-aloud verbal reports, as described in Chapter 3. Both concurrent and retrospective think-aloud activities were conducted in order to collect as much information as possible. I did not communicate with the participants during the concurrent think-aloud sessions. However, during the retrospective think-aloud

sessions, I did ask some clarification questions. Following that, I also briefly interviewed the participants about their English language learning experiences, test-taking experiences, and any other relevant experiences. The questions were usually formulated based on the participants' background information and their observed reading processes, such as "How did you learn English in your home country?" and "Are you familiar with multiple-choice reading tests? Why?"

4.3.3 Data Analysis

A constant comparison method of data analysis was used for the grounded theory study. This method involves comparing incidents or events in the data to develop categories. As recommended by Glaser (1978), this approach includes looking for key issues, recurrent events, or activities in the data that become categories of focus and then writing about the categories. As a result of the Q-matrix construction exercises reported in Chapter 3, major categories of reading subskills were established based on the previous literature and the think-aloud verbal reports. As the categories had already been established, the present data analysis focused on patterns and relationships between the existing categories to explore the differences between East Asian and Romance ESL learners in terms of the subskills of reading.

First, the transcripts were open-coded, and the purpose was to identify incidents and understand processes. The questions that guided this initial coding were "What is happening here? How did he/she get this item right or wrong? How are the reading processes related to his/her native language?" I added brief comments in the wide margins of the transcript. For instance, Eva had said, "I know g*ranite* because it's very similar in my language. It's a kind of rock." And, I commented that the participant had "used cognates in understanding English words."

Second, I reviewed the portions of the transcript on which I had written comments again in a more general way in order to understand any patterns that the participant had shown in answering the questions. For instance, one salient pattern in Eva's reading process was a heavy reliance on her native language (Spanish) for vocabulary recognition. During this second review, I wrote additional comments, and all the comments together with relevant portions of the transcript were cut and pasted into an Excel spread sheet for each participant across all items.

Third, constant comparisons were made within and across the native language groups. For instance, the transcripts by Eva (Spanish) and Jin (Chinese) were compared to see if their different native languages might have caused some differences in their respective reading skills and processes. The transcripts by Jin (Chinese) and Hon (Korean) were compared to see if there was any commonality, as they both belonged to the East Asian group. The transcripts of Jin (Chinese) and Ted (Chinese) were compared to see if they shared any commonality despite the differences in their overall English reading ability. As the study proceeded, I wrote memos and journals in order to capture, define, and summarize the differences between and the commonalities among the participants.

Finally, I reviewed the memos and related transcripts once again with reference to the research question, and generated the hypotheses. The procedures described in this section were repeated iteratively until I was confident about the hypotheses that had emerged from the data.

4.4 Results

4.4.1. Participant Profiles

Qualitative studies largely rely on richness of information and context. Understanding each participant is crucial to understanding how his/her reading might have been influenced by

his/her native language and educational experiences. The following is a detailed description of the participants' profiles.

East Asian High-Level Group. Ted was a male Chinese doctoral student majoring in education. He had been in the US for almost two years and had spent 17 years learning English in total. He thought his English reading was "very good." He had received a score of 85 on the TOEFL three years previously; however, he thought this score underestimated his real English ability, because he had spent only one month preparing for the TOEFL but a lot of time during the period leading up to preparing for the GRE instead. Back in China, he had taken TOEFL preparation classes in Beijing and was very experienced with multiple-choice reading tests. He observed that "the difficulty level" of the MELAB reading test "is similar to TOEFL's, but TOEFL has many items that ask you to infer. MELAB mainly asks specific details. "

Yao was a female Chinese student. She had been in the US for 22 months accompanying her husband who was a doctoral student. She had a bachelor's degree in education from a university in China and had studied English for 10 years. She had taken the TOEFL half a year before and achieved a score of 85, higher than the minimum requirement of 80 for acceptance into a graduate program. However, she still thought her English reading ability was only basic. She seemed to be quite familiar with English tests, and the first question she asked before taking the MELAB test was "What is the level of this test? Is it TOEFL level or the College English Test (CET) level?" Regarding test-taking experiences, she said "I practiced a lot for TOEFL.… I spent about one year reciting TOEFL words every day.… I practiced all the previous TOEFL tests, including the paper-based ones. The most practiced parts were reading and listening."

Chika was a Japanese female taking ESL classes with the Mid-State Literacy Council. She had been in the US for three years, accompanying her husband who was a postdoctoral

researcher. She had a bachelor's degree in social welfare from a Japanese university, and she had studied English for 13 years, 10 years in Japan and 3 years in the US. She thought her English reading ability was good. She had no intention of going to graduate school and thus was not planning to take any English-language proficiency tests in the US. While at the high school, in order to prepare for the college entrance exams, she had to practice English tests, and thus was very familiar with multiple-choice reading tests.

East Asian Low-Level Group. Jin was a male Chinese student attending the IECP Level-2 class. He had been in the US for three years and had a bachelor's degree in engineering from a university in China. He was planning to enter graduate school but had not achieved a good enough TOEFL score to do so. He had taken TOEFL four times in the previous three years, but his highest score of 65 was lower than the minimum requirement of 80. In his own words, he was "very experienced with taking multiple-choice reading tests," and thought the MELAB "was just like TOEFL." He had studied English for 11 years, and thought his English reading ability was basic. He said he never enjoyed reading in English, and he mostly only read in order to prepare for the test.

Fei was a Chinese female taking ESL classes with the Mid-State Literacy Council. She had been in the US for 20 months accompanying her husband who was a doctoral student. She had started to learn English since she was in the 6th grade, and had a bachelor's degree in philosophy from a university in China. She was planning to apply for graduate school, and had devoted herself to preparing for the TOEFL during the past year. "I spent four hours a day preparing for TOEFL for the past year," she said, but she thought her preparation was still not sufficient. She seemed to be quite worried about the coming TOEFL test, and thought her English reading ability was between basic and good.

Hon was a male Korean student taking ESL classes with the Mid-State Literacy Council. He had been in the US for one year, having come to join his new wife, who was American. He had a bachelor's degree in biochemistry from a university in Korea, but his real interest was music. He had studied English for 13 years and thought his English reading ability was basic. He was very critical of English education in Korea, as the teachers only focused on analyzing grammar and sentences instead of providing training on critical thinking skills. Now he had abandoned all the old ways of learning English and was enjoying taking English with the Mid-State Literacy Council, as he truly appreciated reading English stories. He had no intention of taking the TOEFL.

Romance High-Level Group. Katia was a Brazilian female taking ESL classes with the Mid-State Literacy Council. She was a native speaker of Portuguese and had been in the US for five years and seven months. She had earned a master's degree in environmental engineering at an American university the year before, and now was accompanying her husband who was a professor. She thought her English reading was "very good." When asked how she had prepared to take the TOEFL five years earlier, she said "two days before the test, I went to the ETS website to look at what the test looked like. And my test result was OK."

Dora was a female ESL student attending the IECP Level-2 class. From Morocco, she was a native speaker of French and could also speak Arabic. She had just graduated from high school and had been in the US for one year to improve her English. She was interested in science and technology. She had taken the TOEFL two months previously and had obtained a score of 85. She considered her English reading ability to be very good. She learned English at high school for three years and then for one year in the US. Back in Morocco, she had not taken any multiple-choice reading tests, because "it's the French system. For reading, they will ask you

questions and you will have to write and respond." When asked how she had prepared for the TOEFL, she answered, "I practiced a little bit."

Romance Low-Level Group. Eva was a female ESL student attending the IECP Level-2 class. She had been in the US for three months to accompany her husband for a one-year academic visit. She had a master's degree in history and musicology and was a music teacher in Spain. She had grown up bilingual in Barcelona, Spain, speaking both Spanish and Catalan. She also spoke French and a little Italian. She had only learned English for a short while, for a year in middle school, for four months at an American English school in Spain before coming to the US, and for another three months at the IECP. Regarding her experience learning English in Spain, she said, "I used English in class, and I also used English outside because I found people for exchange language. I think it's interesting to practice." She also mentioned that she had studied English at the American school mainly because she wanted to learn the language. She thought her English reading ability was good.

Leon was a male student attending the IECP Level-2 class, a native speaker of Spanish from Colombia. He had just finished high school, and he had been in the US for six months on an exchange program. He had plans to become a professional ESL instructor. He also reported that he had not done many reading exercises: "We didn't focus on this kind of passages, but different kind of reading test.… I didn't focus on reading, just in speaking, listening, and culture…. In my English class in Colombia, you can find different kinds of activities, but it's focusing on conversation, role plays. You can watch a movie, or video clips, you try to take the general ideas, some details. In this process, you can get some information but not technically following the grammar." He planned to take the TOEFL in several months, and thought his English reading ability was basic.

4.4.2 Exhibited Group Differences

Vocabulary. The skill of vocabulary refers to recognizing and determining the meaning of specific words and phrases using phonological/orthographic/vocabulary knowledge and/or using contextual clues. An overwhelming theme from the think-aloud verbal reports was that given a similar level of overall English reading ability, members of the East Asian group showed lower performance on the skill of vocabulary. The following are some typical incidents and phenomena found in the think-aloud verbal reports.

Eva, the native Spanish speaker with relatively low English reading ability, frequently referred to Spanish cognates or Latin roots, stems, suffixes, and prefixes for the purpose of recognizing English words. Sometimes, the English word was very similar to its Spanish equivalent. For instance, she said "I know *granite* because it's very similar in my language. It's a kind of rock." Many other times, she was able to recognize a word based on her knowledge of morphology and Latin. For instance, she commented on the word *unpalatable*: "It's from Latin. *Un-* means *non-*. This is the word for this part about mouth in my language. *Palate* is very similar in Spanish. It means *eat*. So I guess this word means *not possible to eat*. All of the technical words are always very similar, because they come from Latin." Eva also realized that the strategy did not work all the time. For example, she said "I can't find *enable* from Latin. Probably this is an Anglo-Saxon word or German word." To summarize, Eva was very proficient at resorting to her native language for word recognition, and this skill did give her an advantage in the reading process.

Ted, the Chinese male, was very troubled by some unknown technical words in passage 3, despite the fact that his overall English reading ability was high. While reading passage 3, he constantly complained about the unknown words. "I hate this passage, too many new words, I

wasn't really sure." In contrast to Eva who could get the meaning of *granite* from her native language, Ted had to rely on his memory: "Maybe it is kind of rock. I may have recited this word before for GRE. Anyway, I am not sure." It seemed that he was drawing the meaning of the word from his memory instead of constructing the meaning during the reading process. The meaning of *granite* was critical to answering item 15, and thus Ted was very hesitant about the answer. Compared to Eva, who drew extensively from her native language, Ted was at a disadvantage with word recognition, though his English reading ability was higher than Eva's.

Katia also occasionally encountered some unknown words in reading despite her relatively high English reading ability. For instance, item 2 required readers to understand the word *minute* [mainju:t], which meant *small* in the context. She pronounced the word as [minit] when reading the passage, which made me suspect that she might not understand it. When asked if she knew the word, she said "I just guessed, small…. The idea of the test makes me feel it. I also think it's because in Portuguese, even in English, *minute* is similar to *minimal*. In Portuguese, it's *mínimo*. I didn't know I had it in my mind when I was reading." When asked to underline the words in the passage that had cognates in Portuguese, she underlined almost half the words in the first two sentences. "Can I stop here? I think there are too many English words that are similar in Portuguese" she said laughing.

A different case was Hon, the male Korean student with low English reading ability. He also pronounced minute as [minit]. Therefore, during the retrospective think-aloud activity, I asked him what the word meant, and he said "[minit]? I think it means time. No, here it is different, *minute absorption of elements.* I think it could be kind of amount. But I don't know many or less." It seemed that Hon was clueless about what *minute* meant in that particular context. Leon, the Spanish-speaking male with low level English reading ability, in fact,

pronounced the word *minute* as *minimal* when reading the passage. When asked how he determined the word's meaning, he said, "I don't know. I keep translating between Spanish and English. Sometimes I can find some Spanish words similar. For instance, *traditional* is very similar in English and Spanish. When you say *traditional*, my mind is *tradicional*. It happens, it's processing in my mind, but I don't know." Still, compared to Hon, who had almost nothing to draw on in his Korean vocabulary, Leon did seem to benefit from his native language of Spanish in recognizing English words.

The above discussion shows how the Romance group benefited from cognates and the shared prefixes, suffixes, and word roots in their efforts to recognize English words. The following case offers an example of how a Chinese speaker was disadvantaged in word recognition due to her unfamiliarity with English pronunciation. Fei, the female Chinese student, seemed to be very bothered by the pronunciation of people's names. Data from her concurrent think-aloud protocols showed that Fei did not seem to recognize the theory mentioned in the passage as Darwin's theory of evolution. I, therefore, asked her (pointing to the word *Darwin*) "Have you heard of this person?" She hesitantly said [dewin]?" I then pronounced the name as [da:win]. She immediately recognized it: "Oh! No! Now I know. It's *Darwin*." And, she added, "When I read a foreigner's name, I just automatically filter it, because I couldn't pronounce it. This happens to me a lot. I usually didn't read names. I just ignore foreigners' names." It seems that she had difficulty pronouncing people's names in English and also was not sensitive to names. "I never learned spoken English," she said, "Although I've learned English for so many years, I couldn't open my mouth to speak at all when I first came here."

A similar case was that of Jin, the Chinese male. He mispronounced *Darwin* as [dra:win] during the think-aloud. I asked him (pointing to the word *Darwin*), "Have you ever heard of this

person" "No." "Then have you heard of the person [dɑ:win]." I pronounced the word in English this time. "Oh, *Darwin*, yeah, I know, all people are from monkey." Now he seemed to connect the name with the Darwin he knew. Due to the mispronunciation of this single word, he failed to connect the theory in the passage to Darwin's theory of evolution even though he knew the theory very well. Jin admitted that if he had realized this, he would have found the passage much easier to understand, as he could have drawn on his prior knowledge.

To summarize, this section described some of the typical incidents and phenomena identified in the two groups' performance on the subskill of vocabulary. They indicated that given similar English reading ability, the Romance group seemed to be more competent on the vocabulary subskill than the East Asian group was.

Syntax. The skill of syntax refers to understanding sentence structure and sentence meaning using syntax, grammar, punctuation, and parts of speech. An item is coded as requiring syntax when understanding the sentence structure is critical for comprehension, and its structure is complex, for instance, inversion of subject and verb, passive voice, two or more clauses connected by a subordinate conjunction, subjunctive words, and pronoun references. The think-aloud verbal reports showed that given similar English reading ability among the Romance-language and East Asian ESL students, English syntax was more challenging for the latter. The following are some typical incidents and phenomena found in the think-aloud protocols.

Ted, the male Chinese student with high English reading ability, expressed the trouble he had with some of the sentence structures. In talking about why he hated paragraph 3, he said, "I don't know many words, and also many long sentences…. I don't feel comfortable with the attributive clauses behind nouns." Yao, the Chinese female with relatively high English reading ability, also seemed to have difficulty with grammar. When asked if she found any particular

sentence structure difficult, she said, “Yeah, sentence structures are very messy to me.” One example she gave showed that she had difficulty understanding the use of passive voice.

Jin, the male Chinese student with low English reading ability, seemed to have difficulty with prepositions and passive voices. He wrongly understood “far more than” as “not more than” in item 7, and thus picked the wrong answer. In passage 4, he could not understand the sentence “The need might be met.” He knew the original verb *meet,* but it seemed that he did not realize that *met* was the past participle of *meet*. He noted that he found vocabulary to be more difficult than grammar, but he was confused by long sentences especially those with attributive clauses, in the perfect present tense and/or in the passive voice. He told me that he was considering taking a grammar class to improve his TOEFL score.

Katia, the female Portuguese speaker with high English reading ability, did not show or report any difficulty with syntax. Dora, the native French speaker with high English reading ability, noted that she had “no problem in writing and grammar.” Eva, the Spanish speaker with low English reading ability, said “Grammar and sentence structure are not difficult for me. English grammar is not really difficult,” but she also admitted that “English grammar has a lot of tricky questions for us, because we have a lot of verb tenses, more than English. Also the prepositions are very different than our prepositions.” Leon, the Spanish male, reported that “I feel OK with grammar”; however, like Jin, he did not recognize the passive voice used in “be met with,” and thought “met” meant “meeting or appointment.” Though both had relatively low English reading ability, Leon did not feel as troubled by long and complex sentences as Jin did.

Another distinctive feature was that East Asian students seemed to explicitly analyze syntax during reading, while Romance ESL students did not seem to engage in a similar analytical process in this regard. Fei, the Chinese female with low English reading ability,

showed a marked tendency to analyze grammar when reading, probably due to her intensive test preparation and English learning experience. For instance, when reading "scientists have started …" she explicitly pointed out the "present perfect tense." Regarding the sentence "While it is hoped it might be…," she was very analytical about the grammar: "*might* means *not yet*, they hope so, but it may not happen." When asked how she felt about the grammar, she said, "Grammar is Ok with me, if I read slowly." She admitted that she usually "analyze the sentence," because that was her habit and personality.

Hon, the male Korean student gave a similar account of his own reading practice: "In Korea, we just analyze the sentence when we read. I don't like that. I like the short story class now, because I really enjoy reading. I don't want to analyze sentences anymore." He also commented that the word order of Korean is opposite to that of English, and he had to read or speak slowly because of this reversed word order.

To summarize, even though members of the Romance group were occasionally bothered by English syntax, their East Asian counterparts seemed to be more challenged by it during the reading process. Therefore, given similar English reading ability, the Romance group seemed to have better performance on syntax than did the East Asian group.

Extracting Explicit Information. When extracting explicit information, readers attempt to match lexical and/or syntactic information in the question to the relevant part of the text, and they may also identify or formulate a synonym or a paraphrase of the literal meaning of a word, phrase, or sentence in the relevant part of the text. An overall pattern shown in the think-aloud verbal reports was that given similar English reading ability East Asian students were better at extracting explicit information than were the Romance students.

One characteristic of Leon's reading process was his extremely low speed in responding to the items. Even for explicitly stated information, he had to read the passage over and over searching for the information necessary to respond to the items. The answer to item 11 was exactly stated in the opening sentence of paragraph 2, such that a reader who matched the question to this sentence would obtain the correct answer. However, Leon spent about five minutes reading the whole passage word for word in order to search for the information. He actually read the sentence containing the answer three times without realizing that was the answer. He seemed to be very troubled by the word *lava*: "They say *lava*. They say *lava*. I don't know this word." In contrast, Chinese speaker Jin, whose overall English reading ability was similar to Leon's, quickly picked the correct answer to item 11. Immediately after reading the item, he pointed to the sentence containing the answer: "It's right here. It told you just now." He also said that he did not understand the word *lava*, but that it was "not necessary to know it. It's just a word, some stuff."

In these examples, neither Leon nor Jin knew the word *lava*, but Jin did not seem to be bothered at all. He simply matched the item to the answer. Leon might have needed more time to process information due to his less advanced English reading ability, but another possible reason for requiring more time was his lack of experience in taking multiple-choice reading tests. When answering the first item, he asked this question: "Can I choose two or just one?" He also reported that he had not completed many multiple-choice reading exercises in the past. Jin, though, was very experienced at taking tests: "You know, the order of the item is the same as the order where the answer appears in the passage. So if it's the first item, I will go to the beginning of the passage to search information." I asked Jin how he had learned this strategy. He said that he had learned it from the New Oriental TOEFL preparation class in Beijing.

Fei, another Chinese student, was even more purposeful. As Fei worked on passage 1, she only read the first paragraph and then proceeded to answer the questions. During the retrospective think-aloud session, I asked: "Would you worry that not reading the second paragraph may influence your decision making?" "It didn't influence me for the items 1 to 2. Then for item 4, I thought it was influenced, because the question kept mentioning *poison*, and I didn't read about it. So I returned to paragraph 2." She was also very clear about the implicit rule that the answers to the first several items could usually be found at the beginning part of the passage. By going back and forth frequently between the passage and the items, Fei was quite successful at locating the necessary information.

Yao, another Chinese female, even started to work on the questions without reading the passage at all. After reading item 2, which contained the word *liquid,* Yao started to search for information in passage 1: "Let me see, *liquid*. Where's *liquid*?… Let me find *liquid*.… *Liquid* is out as some *lava cool*. Oh, *lava cool*. Oh, there's *cool* here.… So I will choose a *cooling lava*. It's there." Even though she did not read the passage at all, simply by focusing on the word *liquid* in the item stem she had been able to quickly find the right answer. However, she probably did not understand the sentence at all, because she thought that *incidental product* meant *impetus*, while *incidental product* was the key to understanding the question. During the retrospective think-aloud session, Yao said, "This is similar to TOEFL. It's in order, and it doesn't need you to carefully understand the passage. I didn't understand it at all. I found the sentences and then compare, and then I used logical analysis."

Katia, the Portuguese speaker with high English reading ability, managed to pick the right answer for items requiring explicit information most of the time, but usually at very low speed, and she seemed to be effortful in finding the right answer. For instance, the answer to item

9 was explicitly stated in the passage, but Katia had to read the passage word for word from beginning to end searching for the information. After quite a while, she said "Well, both *A* and *B* fit here, I don't know." It seemed that she was focusing on understanding the passage without much awareness of the fact that it was a multiple-choice test. On the contrary, Jin, the Chinese male with much lower English reading ability than Katia had, pointed at the right answer several seconds after reading the item: "You see, the study asks question about the interviewees." He mainly matched the word *interviewees* as it appeared in the item stem and in the sentence. It seems that despite Katia's high English reading ability, she was not used to taking multiple-choice reading tests, which may have resulted in her relatively lower performance in terms of extracting explicit information for the items. Jin, however, was well trained in extracting explicit information for multiple-choice reading tests, and thus he was even more efficient and successful.

To summarize, this section described some of the typical incidents and phenomena evident in the two groups' performances on the subskill of extracting explicit information. They indicated that given similar English reading ability, the East Asian group seemed to be better at extracting explicit information than the Romance-language group was.

Connecting and Synthesizing. The skill of connecting and synthesizing involves higher-level cognitive processes. Readers may integrate, relate, or summarize the information presented in different sentences or parts of the text to generate meaning; they may understand the relationships between sentences and the organization of the text by using cohesion and rhetorical organization knowledge; and they may recognize and evaluate the relative importance of information in the text by distinguishing major ideas from supporting details.

Hon, the male Korean ESL student, struggled with many of the items. He felt that passage 1 was "confusing" and that the options provided were "distracting." His difficulty in most cases seemed to be due to his lack of English vocabulary. For instance, item 2 required understanding the explicit information within a single sentence. As long as he could find the sentence containing the answer, he should have been able to pick the right answer by matching. Unfortunately, he did not understand five words in that single sentence, i.e., *assist, device, perceive, minute,* and *extract.* As a result, it took him a long time to choose the right answer for item 2, which was in fact an easy item. Yet, item 3, which asked about the central idea of the passage, seemed to be fairly easy for him. When asked about item 3, he said, "You know this (pointing to the correct answer) is the big picture. This one explains everything." It seems that although bothered by particular words at the individual sentence level, he was still able to grasp the main idea of the passage based on scattered pieces of information.

The same phenomenon was also observed with Jin, the Chinese male with low English reading ability. He was not able to answer item 6, for which it was necessary to understand the word *diligence*. He spent quite a while and then wrongly thought *diligence* meant *rich*. He was also troubled by item 7 mainly because he was confused by the "more … than" structure. However, he efficiently found the answer to item 8, for which it was necessary to synthesize information from different parts of the passage in order to grasp its overall theme. After reading the four options, he found that he did not understand some of the words in the options. Still, after working on items 9 and 10, he returned to item 8 and quickly selected the correct answer. When asked, he said that "The first paragraph and the last paragraph say a lot about survival. I think the overall passage should be about survival. It couldn't be something else. The answer should be about *selection and survival*." Although Jin seemed to be blocked by the lack of vocabulary or

insufficient syntax knowledge, his ability to understand information at different places in the passage helped him understand its overall idea.

Dora, the female French speaker with high-level English reading ability, however, reported that she found item 8 challenging. Seemingly troubled by the phrase "according to the passage," she stated that "They say according to the passage. I thought the answer must be explicitly specified in the passage." Constrained by this assumption, she only aimed at the explicitly stated information and thus picked a wrong answer. It seems that Dora was not very familiar with the idea of items asking about implicit information embedded in different places in the text. When asked about this point, she stated that "We don't do a lot of these kinds of multiple-choice questions for reading tests. In the French system, it's more about writing essays, for example, they will ask the question, and you will have to write the response. We don't do this kind of test at all." Her lack of experience with multiple-choice reading exercises appeared to contribute to her failure to pick the right answer to item 8.

Differences were also apparent in the speed with which the participants read and responded to the items. Both Leon and Katia picked the right answer for item 3, but only after an extended period spent reading and hesitating. Katia probably had the highest English reading ability among her peers in the think-aloud activity, but she considered item 3 to be "very difficult." The think-aloud activity was not timed, so that Leon and Katia could take the time reading the passages word by word repeatedly; however, it is unknown whether these two participants would have been able to pick the right answer to this item had the test been timed as it actually is.

Yao, the Chinese female, showed considerable expertise in using logical relations to find answers to items that required global understanding of the text. She appeared very purposeful,

targeting her efforts at picking the right answers rather than comprehending the passage. She focused on reading the items and then briefly explored the passage to find the required information. She also frequently skipped chunks of the passage and always read back and forth between the passage and the items. The stem of item 15 included the word *infer.* She was thus aware that the answer to this item involved an overall understanding of the text: "For this item I need to summarize. It is not easy to find it. *Slow cooling, slow cooling, slow cooling*, where did I see it? It should be about the whole text." Then she noticed the sentence "cooling more quickly may form a ramification." "Oh, I will eliminate option B *ramification,* because *ramification* is associated with *quick,* so B should be wrong." She continued to read "with high density, the same lava may cool more quickly." "Ok, option C *high density* is also wrong, because high *density* is also related to quick cooling…." She finally picked the right answer: "I really don't understand the sentences. I just used logic to figure out the relationship between them." Overall, Yao was very good at analyzing the relationships between sentences and synthesizing information from different places in the text, and she seemed to rely on this ability to compensate for her deficit in linguistic skills.

Leon, the male Spanish-speaker, seemed to be quite challenged by item 15, which required integrating information from different parts of the passage. He started to read the options one by one, and then went back to look for the keyword *cooling*. After reading the passage from beginning to end, he read the options again. Four minutes into this second reading, he was still unable to pick the right answer. It seemed that the information was too overwhelming for him, and he was not able to successfully manage and organize it. Furthermore, he seemed to be quite troubled by the unknown words: "I didn't know *lava*. They are technical words. I know this is research publication, but I am not very familiar with these vocabularies."

Jin, the Chinese male with similar English reading ability to Leon, though, did not seem to be troubled by the fact that he did not know the word *lava*. It appeared that Jin was used to reading in English with many unknown words and had found ways to circumvent difficult words by drawing on other skills.

In conclusion, given similar overall English reading ability, members of the East Asian group showed better performance with connecting and synthesizing information compared to members of the Romance group.

Interactions between Observed Group Differences and Overall English Reading Ability. The observed differences between the two language groups remained constant regardless of the students' respective English reading ability. For instance, as discussed in the previous section, both low-level and high-level East Asian readers appeared at a disadvantage in vocabulary and syntax compared to their Romance peers. Ted, the high-level Chinese male, who was a doctoral student in the US, was still troubled by some English syntax despite his very high English reading ability. Yet, Leon, whose overall English reading ability was much lower than Ted's, benefited from his native language Spanish in terms of vocabulary and syntax. Katia, the Brazilian female student, seemed to be slow and effortful in responding to items that required extracting explicit information, even though she held a master's degree from an American university and had very high English reading ability. However, Jin, the Chinese male with low English reading ability, seemed to be very efficient and accurate in searching for and locating information probably due to the training he had received on taking multiple-choice reading tests.

To summarize, the observed differences between the two language groups held for students of varying English reading ability. In other words, the exhibited group differences were not dependent on the students' overall English reading ability.

4.5. Post Literature Review Analysis

Based on the think-aloud verbal reports from 10 ESL students, six with East Asian language backgrounds, and four with Romance language backgrounds, hypotheses were generated regarding the differences between them on the subskills of reading as represented in the MELAB reading test. The specific hypotheses developed are as follows: (a) Given the same English reading ability, East Asian ESL learners do not perform as well as Romance ESL learners on linguistic skills, such as vocabulary and syntax, and (b) Given the same English reading ability, East Asian ESL learners perform better than Romance ESL learners on comprehension skills, such as extracting explicit information, and connecting and synthesizing information. These hypotheses are further justified in reference to related literature in the following section.

4.5.1. Transfer of Linguistic Skills from L1 to L2

The transfer of linguistic skills largely depends on how similar a learner's L1 is to the L2. Due to major differences between the first languages of East Asian students (i.e., Chinese, Korean, and Japanese) and the English language, the transfer of L1 linguistic skills to the study of L2 (.i.e., English) is more difficult compared to Romance-language speaking ESL students.

Regarding word recognition in reading, a certain writing system may lead to a different print-processing experience, and thus learners whose first languages have writing systems that are drastically different from the English writing system may have a disadvantage in English word recognition (Koda, 2005). For example, English and Spanish are alphabetic systems, and each letter represents a phoneme. However, each symbol maps into a morpheme in logographic systems, such as Chinese characters and Japanese Kanji. Native Chinese speakers may rely more on orthographic cues than phonological cues (Grabe, 2009; Hamada & Koda, 2010). Thus,

Chinese students may encounter difficulty in English-word recognition due to insufficient phonological awareness. This corresponds with the observations obtained during the think-aloud protocols, in which Chinese students Jin and Fei did not recognize the personal name *Darwin*. Japanese readers of ESL were also found to be less sensitive to phonological information in processing English words than those with an alphabetic L1 background (Brown & Haynes, 1985; Koda, 1990). Although Korean Hangul is also alphabetic, it constitutes basic graphic elements for forming words as it requires assembling individual symbols into syllable blocks (Taylor & Taylor, 1995). In addition, Korean orthography has been found to be a cause for the relative importance of morphological processing for Korean students (Cho & McBride-Chang, 2005a, 2005b). To summarize, the different writing systems of ESL learners' native languages may influence their English word recognition to different extents (Akamatsu, 1999; Biederman & Tsao, 1979; Perfetti & Zhang, 1991; Tzeng & Wang, 1983), and East Asian students are at a disadvantage compared to those with Romance language backgrounds in this regard.

In the case of sentence processing, though sentence processing mechanisms are likely to be universal (Inoue & Fodor, 1995), different languages may still have some specific features in this regard. When the sentence processing in L1 matches that in L2, the learning of L2 is facilitated, and vice versa. For instance, English and Spanish have a head-initial structure for relative clauses and adverbial clauses, whereas Japanese has a head-final structure. Flynn and Espinal (1985, p. 98) compared the different sentence structures in English, Spanish, and Japanese as follows:

English: The child [who is eating rice] is crying.

Spanish: El nino [que come arroz] llora.

"The child who eats rice cries"

Japanese: [Go han-o tabete-iru] ko-ga] naite-imasu

"Rice eating is child crying is"

In Japanese, the parsing decisions remain tentative until the final word of the sentence is processed; however, in English and Spanish, the parser makes some early commitment about its structural interpretation (Mazuka & Itoh, 1995). It is not surprising to find, therefore, that Japanese learners of English may spend more time on sentence processing than speakers of Spanish do. Korean also is a head-final language with the predominantly Subject–Object–Verb (SOV) word order, which makes it difficult for Korean ESL learners to process English sentences. In the think-aloud activity, Korean student Hon explicitly reported that the reversed word order in English made his processing speed slow.

Chinese has the same Subject–Verb–Object (SVO) word order as English. Though word order is the most important cue in sentence processing in English (Bates, Devescovi, & D'Amico, 1999), whereas in Chinese, the most important cue is the passive marker bèi (被), followed by noun animacy, word order, object marker bǎ (把), and indefinite marker yǐ (已) (Li, Bates, & Macwhinney, 1993). Therefore, comprehension of English relies heavily on word order, but comprehension of Chinese depends to a much greater extent on context and semantics. In other words, Chinese syntax focuses on meaning, and sentence structure is usually loose; therefore, Chinese ESL learners are challenged by many grammatical features that only exist in English. For example, Cheng (1993) found that some college students in Taiwan had significant difficulty in English reading due to confusion caused by the frequent use of prepositions and relative clauses. This agrees with my observation in the think-aloud activity, especially my observations of Chinese males Ted and Jin. Huang (2009) also noticed that Chinese students were especially troubled by long sentences with complicated structures, such that they had to

analyze the grammar carefully in order to aid comprehension. This echoes my observations of Fei who explicitly commented on the grammatical features of the sentences during reading. In addition, Hon, the Korean male student, frequently mentioned that he had to "analyze" when he was reading, and he did not like that.

Juffs (1998) conducted an experimental study to investigate the differences between the performances of Romance-language speakers (Spanish, Italian, Franco-phones, and Portuguese) and East Asian language speakers (Chinese, Korean, and Japanese) in terms of the accuracy and speed with which they processed sentences containing verbs that are temporarily ambiguous in interpretation between a main verb and a reduced relative clause. He concluded that the typological relationship between English and Romance afforded the Romance-speaking learners an advantage. Moreover, the similarity in accuracy between the Chinese and Japanese/Korean groups suggested that the head-final construction of relative clauses in these three languages put native speakers of them at a relative disadvantage in reading English.

To summarize, the reading literature supports the findings of the grounded theory study that given the same overall English reading ability, East Asian ESL learners as compared with Romance-language ESL learners showed disadvantages in linguistic skills such as vocabulary and syntax. This constitutes important evidence in support of the validity of the results of the grounded theory study.

4.5.2 Compensatory Nature of Reading

The think-aloud protocols showed that given the same overall English reading ability, East Asian ESL learners had advantages in comprehension skills such as extracting explicit information and connecting and synthesizing. It seems that East Asian ESL learners had to rely more on their comprehension skills and test-taking skills to compensate for linguistic deficiency.

This observed compensation is aligned with the prevailing reading theories. Stanovich (1980, 1986) proposed a compensatory-interactive model of reading. A major claim of the model is that "a deficit in any particular process will result in a greater reliance on other knowledge source, regardless of their level in the processing hierarchy" (p. 32). For instance, a person with poor word recognition skills may actually be prone to rely on contextual factors because these provide additional sources of information. This point is in accord with my observations of the think-aloud protocols with the East Asian ESL students. For instance, when Yao, the female Chinese student, encountered difficulty with the words and syntax of a passage, she habitually resorted to logical analysis in order to obtain a general idea of the text.

Other researchers have also discussed the compensatory nature of reading. Coady (1979) hypothesized that second language reading consisted of three interactive elements: conceptual abilities, background knowledge, and processing strategies. He also postulated that "a weakness in one area can be overcome by strength in another" (p. 11). Bernhardt (2005) also suggested a compensatory model of second language reading, according to which 20% of the variance of L2 reading is explained by L1 literacy, 30% is explained by L2 language knowledge, and 50% is explained by other elements. Bernhardt (2011) believed that her model reflects Stanovich's interactive-compensatory model in that "where knowledge sources at all levels contribute simultaneously to pattern synthesis and where a lower-level deficit may result in a greater contribution from higher-level knowledge sources" (Stanovich, 1980, p. 47). In addition, Walczyk's (1995, 2000) compensatory-encoding model argues that compensatory strategies are continuously used to counter inefficiencies and skill weaknesses during reading.

Some empirical studies have found evidence for compensatory reading processes. For instance, Stevenson, Schoonen, and Glopper (2007) found that readers compensate for language

difficulties by treating them with greater attention but without detracting from the global reading process. Similarly, it was found in the think-aloud protocols that some East Asian students explicitly analyzed the grammar, probably because they found this aspect difficult. Matthews (1990) also discussed the compensatory phenomenon in reading. For instance, reading items requiring understanding a large stretch of text might be easier for poor readers. Even though a student might not understand one or several words in a sentence, he/she can still resort to other parts of the text for information. As a result of this compensation, a poor reader might be more likely to correctly answer a global item than one requiring only local information. The think-aloud protocol has shown that Jin and Hon correctly answered the main idea question despite their lack of sufficient vocabulary. It seems that they tried to maximize the chances of success by relying on all the other available information to compensate for their linguistic disadvantages.

East Asian ESL learners seemed to greatly rely on their expertise in comprehension skills and test-taking skills acquired from test-oriented learning and intensive test preparation exercises. A Romance ESL student may have spent a few days acquainting herself/himself with an English test, whereas an East Asian ESL student may have invested several years preparing for it. The long time exposure to multiple-choice tests and also the intensive test preparation have trained the East Asian students to be more adept at locating information for multiple-choice items. For instance, in the think-aloud activity East Asian students were aware of the potential position of the sentence that might yield answers to the questions. Instead of following the flow of the passage and the order of the items, they jumped between the passage and the items, using their reasoning ability to synthesize and connect ideas from different places—a process that helped them to select the right answers.

To summarize, it seems that East Asian students showed higher performance in comprehension skills in order to offset their disadvantages in vocabulary and syntax, so that they could achieve the same reading performance with their Romance counterparts. In general, the findings of the grounded theory study make sense given the compensatory nature of reading.

4.6 Discussion

In this study, data were collected via think-aloud protocols from ESL students with an East Asian language background and ESL students with a Romance language background, and were analyzed using a grounded theory approach, i.e., constant contrastive comparison. The outcomes of this stage were explicit hypotheses regarding how East Asian ESL learners and Romance ESL learners may differ in terms of reading subskills.

External audits were consistently conducted during the study to safeguard its quality. Dr. Dorothy Evensen, Professor of Higher Education, is an expert on grounded theory methods and an active reading researcher. She offered considerable guidance and advice on methodology issues related to this study. Specifically, she guided and audited some of the coding procedures. For instance, I demonstrated the coding procedure with sample transcripts for Dr. Evensen to critique. Parts of the memo and narration were also provided to her for review as an element of the final project in her qualitative methods class. Dr. Hoi K. Suen, Distinguished Professor of Educational Psychology and a measurement expert, and Dr. Bonnie Meyer, Professor of Educational Psychology and an expert in reading comprehension, also audited my initial coding for the reading model and hypotheses generation. Peer auditing was consistently conducted with Julieta Fernandez and Aziz Yuldashev, both advanced graduate students in applied linguistics with rich ESL teaching experience who were also well-trained in qualitative methodology. They

each examined the data collection and analysis procedures.

Furthermore, the generated hypotheses were validated against the reading literature. A more extensive post literature review analysis was conducted to validate the hypotheses, and the important question “Do the hypotheses make sense?” was asked. Cross-linguistic studies have shown that East Asian ESL learners may have disadvantages with linguistic skills due to the fact that their native languages are very different from English. This agrees with the observation that East Asian students seemed to be more challenged than were the Romance-language students by vocabulary and syntax during the think-aloud activity. The compensatory-interactive model of reading (Stanovich, 1980) argues that readers try to use other resources to compensate for their low proficiency with a particular process. This is in alignment with the phenomena that East Asian ESL learners relied more on their comprehension skills and test-taking skills to offset their relatively deficient linguistic skills in order to achieve the same overall reading performance as their Romance-language counterparts. Overall, the generated hypotheses appeared sensible, workable, and trustworthy.

CHAPTER 5

HYPOTHESES TESTING ON READING SUBSKILL DIFFERENCES –DIFFERENTIAL SKILL FUNCTIONING

In this chapter, the specific hypotheses developed in Chapter 4 were tested by analyzing the subskill profiles of the East Asian examinees in comparison with those with Romance L1 backgrounds via a series Differential Skill Functioning (DSF) analyses. The DSF analysis method is an extension of the Differential Item Functioning (DIF) method. DIF is widely used to identify potentially biased test items, whereas DSF is used to investigate group differences at the level of cognitive skills.

5.1 Literature Review

5.1.1 Overview of DIF Techniques

DIF has been a widely used technique for item bias detection. DIF occurs when examinees from different groups show different probabilities of success on the item after being matched on the underlying ability the test is intended to measure (Camilli & Shepard, 1994). However, items showing DIF are not necessarily biased. Item bias occurs when examinees in one group are less likely to answer an item correctly than examinees in another group because of some characteristics of the test item or testing situation that are irrelevant to the test purpose. Therefore, DIF is regarded as a necessary but not sufficient condition for item bias (Zumbo, 1999).

There are two major types of DIF: uniform and non-uniform (Mellenbergh, 1982). Uniform DIF exists when the statistical relationship between item response and group membership is constant for all levels of the matching ability variable. An item may consistently

favor one group and against the other regardless of the underlying ability that is to be tested. Non-uniform DIF exists when this statistical relationship is not the same for all the matching ability levels. One group may have a relative advantage at one end of the ability level, whereas the other group may have a relative advantage at the other end of the ability level.

A variety of statistical procedures for detecting DIF have been developed, such as the delta-plot method (Angoff, 1972), the Mantel–Haenszel method (Holland & Thayer, 1988), the logistic regression method (Swaminathan & Rogers, 1990), the item response theory–likelihood ratio test (Steinberg, Thissen, & Wainer, 1990), and the SIBTEST method (Shealy & Stout, 1993). For the purpose of this study, the logistic regression method was used. The following gives a brief introduction to the two most commonly used methods: the Mantel–Haenszel method and the logistic regression method, and the advantages of using the logistic regression method over the Mantel–Haenszel method.

According to Zumbo (1999), the Mantel–Haenszel method treats the DIF detection process as one involving a three-way contingency table. The three dimensions of the contingency table are (a) whether the examinee gets an item correct or incorrect and (b) group membership when conditioning on (c) the total score discretized into a number of category score bins. The logistic regression method tests the statistical effect of the grouping variable and the interaction of the grouping variable and the total score after conditioning on the total score. The biggest difference between these two methods is that the Mantel–Haenszel method needs to discretize the continuous conditioning variable whereas the logistic regression method does not. In addition, the Mantel–Haenszel method assumes no interaction between the conditioning variable and the grouping variable, whereas the logistic regression method allows for an interaction. The Mantel–Haenszel method, therefore, only examines uniform DIF.

The logistic regression method has many advantages over the Mantel–Haenszel method. For example, the logistic regression method does not need to categorize a continuous conditioning variable; it can model uniform and/or non-uniform DIF simultaneously (Swaminathan, 1994); it can generalize the binary logistic regression model for use with ordinal item scores; and it provides flexibility in model specification and thus is especially efficient for simultaneous conditioning on multiple abilities (Zumbo, 1999). Furthermore, Rogers and Swaminathan (1993) have shown that the logistic regression method produces similar results to the Mantel–Haenszel method for uniform DIF detection, and it performs better than the latter in terms of type I error. Hence, the logistic regression approach was adopted for the DSF analysis in this study.

Logistic regression models the probability of a correct response to an item as a function of the observed total score X, group membership G, and the interaction between X and G. The general logistic regression equation is

$$Y = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 X * G \qquad [5.1]$$

where Y is a natural log of the odds ratio of a correct response. That is, the more precise equation is

$$\ln[\frac{p_i}{(1-p_i)}] = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 X * G \qquad [5.2]$$

where p_i is the proportion of individuals who endorse the item in the direction of the latent variable (or a correct response in the context of ability testing).

To test for significance of DIF, there is a natural three-step hierarchy of the entry of predictor variables as follows (Zumbo, 1999):

Model 1. The conditioning variable X (i.e., the total score) is entered:

$$Y = \beta_0 + \beta_1 X \qquad [5.3]$$

Model 2. The grouping variable G is entered:

$$Y = \beta_0 + \beta_1 X + \beta_2 G \quad [5.4]$$

Model 3. The interaction term X * G is entered:

$$Y = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 X * G \quad [5.5]$$

The -2 log-likelihood value of Model 2 is subtracted from the corresponding -2 log-likelihood value of Model 1. The resultant difference can be evaluated as a χ^2 value with 1 degree of freedom. A significant difference indicates the existence of uniform DIF.

Likewise, the -2 log-likelihood value of Model 3 is subtracted from that of Model 2. The resultant difference can be evaluated as a χ^2 value with 1 degree of freedom. A significant difference in this particular comparison indicates the existence of non-uniform DIF.

For an overall evaluation, the -2 log-likelihood value of Model 3 can be subtracted from that of Model 1. The resultant difference can be evaluated as a χ^2 value with 2 degrees of freedom. This is a simultaneous test of uniform and non-uniform DIF. However, as noted by Swaminathan and Rogers (1990), "the interaction term may adversely affect the power of the procedure when only uniform DIF is present because one degree of freedom is lost unnecessarily" (p. 366).

5.1.2 DSF in Cognitive Diagnostic Analysis

Though traditional DIF analyses focus on the functioning of individual items in a unidimensional test, the extension of these methods to investigate DSF would necessitate the consideration of multiple dimensions. Traditional unidimensional IRT models linearly order examinees in one dimension. With Cognitive Diagnostic Models (CDMs), each examinee receives a diagnostic profile indicating mastery or non-mastery of each of the skills required for the test. This profile is often called the examinee skill mastery pattern. Cognitive diagnostic

assessment has only emerged recently, such that using the DIF procedure in conjunction within a cognitive diagnostic framework has not been fully explored.

The origin of the idea for DSF can be traced back to Milewski and Baron (2002), who extended the DIF procedure to individual performance on skills measured by the Preliminary SAT/National Merit Scholarship Qualifying Test (PSAT/NMSQT) in order to compare aggregate groups, such as schools or states, to the total population matched on overall scores. A modified Rule Space Model was used to classify examinees into skill mastery patterns associated with different cognitive skills (DiBello, 2002; DiBello & Crone, 2001). The DIF procedure was applied to cognitive skills rather than to items, and thus Milewski and Baron proposed the name Differential Skill Functioning (DSF). The statistical method remains the same; however, item level differences are replaced by skill level differences.

Gierl, Zheng, and Cui (2008) described how the Attribute Hierarchy Method (AHM) has been used to evaluate differential group performances at the cognitive attribute (or skill) level. Similar to the Fusion Model, the AHM can estimate skill probabilities for each examinee and thus provide specific information about an examinee's skill mastery level. Gierl et al. proposed an approach called attribute-level differential functioning (ADF): "ADF occurs when examinees, with the same matching attribute pattern but from different groups, have unequal probabilities of responding to items that measure the studied [skill]" (p. 73). This is conceptually the same as the DSF analysis method proposed by Milewski and Baron, though the ADF follows a more structured confirmatory approach involving four steps. First, the ADF hypotheses are specified based on substantive knowledge. Second, the probabilities for the studied skill are estimated using the AHM model. Third, the matching skills are defined. Examinees are matched on only the skills that have no relationship to the studied skills, so that the matching skills are purified.

Last, the SIBTEST is used to evaluate the magnitude and direction of the group difference on the studied skill. This approach is structurally and theoretically sound. However, it is very challenging to specify hierarchical relationships among the skills; likewise, it is difficult to determine whether one skill is independent of another. It is also very difficult to specify the initial hypotheses due to the lack of understanding of the cognitive processes. Therefore, the ADF approach within the AHM framework is in its initial stages of development.

Some other comparison studies have been conducted at the skill level instead of the item level, though not necessarily using DIF techniques. Tatsuoka and her colleagues conducted a series of studies comparing mathematics achievement differences across countries, analyzing data from the TIMSS-R 1999 via the Rule Space Model. Tatsuoka, Corter, and Tatsuoka (2004) identified 23 skills as underlying the math test and compared the mean mastery level differences among the countries in their study. However, as no matching variable was controlled for, the differences demonstrated were only observable performance differences between the two groups. In another study, Dogan, Guerrero, and Tatsuoka (2005) combined traditional SIBTEST techniques and Rule Space Model to detect the strengths and weaknesses of 10 TIMSS-R countries. Examinees in different countries were matched on true scores instead of observed scores in the SIBTEST for DIF detection. This study showed that the different performances among the countries on certain skills led to DIF on the items that required these skills.

To summarize, the purpose of DSF is to compare group performance on certain subskills within the framework of cognitive diagnostic assessment in order to understand the relative weaknesses and strengths of examinees from different groups. DSF exists when examinees from different groups have different probabilities of successful performance with a certain subskill underlying the measured construct, given that their overall ability the test is intended to measure

is controlled for. Although DSF employs statistical techniques that are similar to those in DIF analyses, the target, or the unit of analysis, is a skill instead of an item.

5.1.3 Matching Criteria in DIF/DSF Studies

The traditional Mantel–Haenszel or logistic regression method for DIF analyses generally uses the total test score as the matching variable. However, including items that have been detected to have exhibited DIF in the total score may contaminate the matching criteria. Holland and Thayer (1998) proposed a two-step process in which the Mantel–Haenszel procedure is implemented. First, the total score is used as the matching criterion. Items identified as exhibiting DIF are then removed from the conditioning total score, and the Mantel–Haenszel procedure is re-implemented using this "purified" total score as the matching criterion. This process is referred to as purification of the matching criterion (Clauser, Mazor, & Hambleton, 1993; Zumbo, 1999). However, as the matching variable varies for different items, this process may result in less than optimal DIF detection and also makes the finding interpretation more complicated (Zenisky, Hambleton, & Robin, 2003).

A multiple-variable matching method has also been investigated in order to improve the matching mechanism in DIF detection (Clauser, Nungester, & Swaminathan, 1996; Kubiak, O'Neill, & Payton, 1992; Zwick & Ericikan, 1989). For instance, having included the variable Extra Lesson Hours After School (ELHAS) in their study on detecting DIF between Taiwanese students and American students in the Third International Mathematics and Science Study (TIMSS), Wu and Ericikan (2006) concluded that DIF would be identified more accurately when the external variables were also controlled for in addition to the ability being measured. Clauser et al. (1996) also confirmed that extra matching on an educational background variable improved the precision of DIF detection in the National Board of Medical Examiners' Part III examination.

To summarize, in addition to the internal ability matching variable, external matching variables can also be controlled for in order to improve the accuracy of DIF detection.

As cautioned by Zhang (2006), using the total score as the matching variable may not be feasible when the test is calibrated by a multi-dimensional cognitive diagnostic model. She proposed matching examinees on their skill profile patterns instead. However, when many skills are involved in a test, matching on profile patterns may not be practical. Take the MELAB reading test for example: with four skills involved in the test, examinees could have as many as 16 (i.e., 2^4) skill profile patterns, such as 1111, 1101, 1011, 1001, 0111, 0011, 0001, 0101, 1110, 1100, 1010, 1000, 0110, 0010, 0000, 0100, in which 1 indicates mastery of the skill and 0 indicates non-mastery of the skill. In addition, some skill profile patterns may have far fewer examinees than others due to the different difficult levels of the skills (Lee & Sawaki, 2009a). Given that the sample size of the current DSF study was only 669, it is not practical to match examinees on 16 skill profile patterns. To summarize, using the skill profile pattern as a matching variable may be conceptually appealing but impractical. I, therefore, decided to use the total score as the matching variable for the follow-up DSF analysis.

Milewski and Baron (2002) used the total score as the matching variable in their DSF study. Gierl, Zheng, and Cui (2008), however, used the skill profile pattern as the matching variable, and they also purified the matching variable in their ADF procedure. Their matching skills included only those skills independent of the studied skill. Though theoretically appealing, this approach is problematic in that the matching variable itself may change dramatically as the number of skills is usually small compared to the number of items. For instance, when four skills are involved in a test, to compare different groups' performances on skill 1, examinees are matched on skills 2, 3, and 4. Then, for skill 2, they are matched on skills 1, 3, and 4. There is an

obvious and substantial change of the matching variable from one analysis to another, which makes interpretation difficult. Because the DSF analysis in this study aims to test reading subskill differences of native language groups under the condition that they have the same overall English reading ability, it is important to have a stable proxy for overall English reading ability. Therefore, I decided not to purify the matching variable.

To summarize, although it is intuitively appealing to use the skill profile pattern as the matching variable, to use the total score as the internal matching variable without purification is a more reasonable option for the purpose of this study.

5.1.4 DSF Hypotheses

As shown in Chapter 4, the present study's hypotheses on the reading subskill differences between the East Asian and the Romance groups were generated based on substantive evidence such as the think-aloud verbal reports with a grounded theory approach. It was suggested that given the same English reading ability, East Asian ESL learners would show lower performance in linguistic skills, such as vocabulary and syntax, but higher performance in comprehension skills, such as extracting explicit information and connecting and synthesizing than their Romance counterparts.

The think-aloud verbal reports showed that the observed language group differences were consistent for all reading ability levels. In other words, there was no observable indication of the existence of non-uniform DSF for the four skills investigated. It was, therefore, decided that there was no basis for hypothesizing the existence of non-uniform DSF in this study. Therefore, I only hypothesized the existence of uniform DSF regarding the four reading skills. Hereafter in this study, DSF only refers to uniform DSF.

In light of these considerations, the hypotheses generated in Chapter 4 were tested via the application of the logistic regression approach as a DSF tool. The internal matching variable used was the total score on the MELAB, and no purification process was employed. The specific hypotheses related to DSF are as follows:

- Hypothesis 1: There is DSF for skill 1 (vocabulary) favoring the Romance group.
- Hypothesis 2: There is DSF for skill 2 (syntax) favoring the Romance group.
- Hypothesis 3: There is DSF for skill 3 (extracting explicit information) favoring the East Asian group.
- Hypothesis 4: There is DSF for skill 4 (connecting and synthesizing) favoring the East Asian group.

In addition, the literature has shown that female students are generally better readers than are male students (Logan & Johnston, 2009), though how gender differences relate to the specific subskills of reading is not clear. In order to be certain that the observed DSF was not attributable to gender, a two-stage DSF procedure was conducted in this study. In the first stage, only the total score was entered as the internal matching variable. Then, in the second stage, in addition to the total score, gender was entered as the external matching variable.

5.2 Methods

5.2.1 Data Sources

The data used for the DSF analysis were the outcomes of the cognitive diagnostic calibration as described in Chapter 3. Specifically, of the 2,019 examinees in the MELAB dataset, a total of 669 had a native language background in either East Asian languages or Romance languages. Table 5.1 shows the distribution of language backgrounds among the 669

examinees. As shown, 522 examinees had an East Asian language background, while 147 examinees had a Romance language background.

Table 5.1

Sample Size across Language Groups

Group	Native language	Sample size
East Asian (N = 522)	Chinese	410
	Korean	84
	Japanese	28
Romance (N = 147)	Spanish	75
	Romanian	37
	Portuguese	21
	French	12
	Italian	2

Table 5.2 shows the sample size of gender by language group. It can be seen that there are 239 East Asian males, 283 East Asian females, 57 Romance-language males, and 90 Romance-language females. A chi-square association test shows that there was no significant association between gender and native language ($\chi^2 = 2.285$, df = 1, $p = 0.131$).

Table 5.2

Sample Size (Gender by Language Group)

		Language Group		
		East Asian	Romance	Total
Gender	Male	239	57	296
	Female	283	90	373
Total		522	147	669

The continuous outcomes of the Fusion Model calibration for each of these 669 examinees were a posterior probability of mastery (PPM). In accord with Hartz (2002) and Roussos, DiBello, et al. (2007), a cut-off PPM criterion of 0.5 was used to reach a dichotomous mastery status for each examinee on each skill, i.e., non-master if PPM < 0.5, master if PPM >

0.5. Table 5.3 shows the descriptive statistics of the PPM and mastery status of each skill for the two groups. The average PPM for the Romance group was generally higher than that for the East Asian group across all skills. The percentage of masters for each skill was also reported with 0.5 as the cut-off point. Similarly, more examinees in the Romance group in the sample were masters of each of the skills.

Table 5.3

Descriptive Statistics of the PPM

Group	Skills	Mean of PPM	SD of PPM	Percentage of masters
East Asian (N = 522)	Skill 1 (Vocabulary)	0.264	0.320	20.1%
	Skill 2 (Syntax)	0.310	0.347	27.4%
	Skill 3 (Extracting explicit information)	0.363	0.399	35.2%
	Skill 4 (Connecting and synthesizing)	0.347	0.369	32.2%
Romance (N = 147)	Skill 1 (Vocabulary)	0.377	0.356	35.9%
	Skill 2 (Syntax)	0.373	0.359	33.1%
	Skill 3 (Extracting explicit information)	0.447	0.405	45.5%
	Skill 4 (Connecting and synthesizing)	0.396	0.401	37.9%

For more details, please see Appendix J, which provides the number of masters of each skill across gender by language group, descriptive statistics of the PPMs and total scores across gender by language group, and scatter plots of the distribution of PPMs on total scores across language groups.

5.2.2 DSF Procedure

In this study, the logistic regression procedure was adopted for the DSF analysis. The purpose was to determine if the two native language groups differed in terms of the probability of success with the subskills of reading when conditioned on their overall English reading ability.

Also, as discussed previously, the total score was used as the internal matching variable without purification. The following describes the variables and procedures used for the DSF detection.

As shown in Table 5.4, the dependent variable was a dichotomous mastery status for each skill and was coded 1 if the examinee was estimated to have attained mastery of that skill and coded 0 otherwise. The observed total score of the MELAB reading test was chosen as the internal matching variable, so that overall English reading ability was controlled for. The language group variable was 0 for East Asian languages and 1 for Romance languages. Gender was coded 1 for females, 0 for males.

Table 5.4

Variable Names and Coding

Variable	Variable meaning	Coding
$MASTERY_1$	Mastery status for skill 1 (Vocabulary)	1 if master, 0 if non-master
$MASTERY_2$	Mastery status for skill 2 (Syntax)	1 if master, 0 if non-master
$MASTERY_3$	Mastery status for skill 3 (Explicit information)	1 if master, 0 if non-master
$MASTERY_4$	Mastery status for skill 4 (Connecting and synthesizing)	1 if master, 0 if non-master
TOT	Observed total score of MELAB reading test	Continuous, range from 0 to 18
LAN	Language group	1 if Romance, 0 if Eastern Asian
GENDER	Gender of the examinee	1 if Female, 0 if Male

Stage 1: Total Score as the Matching Variable. The total score was used as the internal matching variable in order to determine if the two native language groups had different performance on the subskills of reading given the same overall English reading ability. As shown in the following equations with skill *i* (i from 1 to 4) as the example, only the total score was entered as a predictor in Model 1. Then the language group variable was added as an additional predictor to Model 2. If the -2 log-likelihood difference between Model 1 and Model 2 is larger than a χ^2 value with 1 degree of freedom, DSF exists.

Model 1 Mastery i = TOT

Model 2 Mastery $_i$ = TOT + LAN

Stage 2: Total Score and Gender as Matching Variables. In order to be certain that the observed DSF was not attributable to gender, gender was entered as an external conditioning variable in addition to the internal conditioning variable of the total score. As shown in the following equations with skill *i* (i from 1 to 4) as the example, the total score and gender were entered as predictors in Model 1. Then the language group variable was added as an additional predictor to Model 2. If the -2 log-likelihood difference between Model 1 and Model 2 is larger than a χ^2 value with 1 degree of freedom, DSF exists.

Model 1 Mastery i = TOT + GENDER

Model 2 Mastery i = TOT + GENDER + LAN

Finally, when multiple logistic regression analyses are conducted for DIF/DSF detection, an issue arises in regard to whether the alpha level of 0.05 should be adjusted to control the overall type I error rate. Different approaches are recorded in the literature, and Table 5.5 lists some typical examples.

Table 5.5

Alpha Levels Used in Some DIF/DSF Studies

Study	Overall sample size	Number of items/skills (i.e., number of logistic regression analyses)	Alpha level
Clauser et al. (1996)	2,000	440	0.01
Crane et al. (2007)	495/380	28	0.05
Kim (2001)	1,038	3	0.05
Monahan et al. (2007)	12,945	23	0.00217
Qi & Marley (2009)	440	46/44	0.05
Whitmore & Schumacker (1999)	200/400/600	20/40/60	0.01
Current study	669	4	0.05

Some studies used Bonferroni adjustment. For instance, Monahan, McHorney, Stump, and Perkins (2007) had 23 items in their DIF study and used an alpha level of 0.00217 (i.e., 0.05/23). However, Bonferroni adjustments may be too stringent. According to a simulation study conducted by Scott et al. (2009), in order to maintain acceptable power, the sample size of both the focal and reference groups should be at least 500 if Bonferroni adjustments are to be made. Some researchers used a significance level of 0.01 regardless of the number of items in the test and the number of logistic regression analyses. For instance, Clauser, Nungester, and Swaminathan (1996) used a significance level of 0.01 to detect DIF in a test with 440 items. Other researchers (e.g., Crane, Cetin, Cook, Johnson, Deyo, & Amtmann, 2007; Kim, 2001; Qi & Marley, 2009) did not adjust the alpha level. All these practices have been accepted in the field. As only four skills were under investigation in the present study, four logistic regression analyses were conducted at each stage. In addition, the overall sample size of 669 in the present study was moderate. Thus, due to the stated concern about statistical power, an alpha level of 0.05 was used in this study without adjustment.

In addition to statistical significance level, effect size has also been used to measure the magnitude of DIF. The increased portion of R^2 after the group variable is entered into the logistic regression could be used as an effect size measure. For instance, in a series of DIF studies involving the TOEFL CBT essay prompts, Breland, Lee, and Muraki (2005), Breland and Lee (2007), and Lee, Breland, and Muraki (2005) combined the R^2 change and p-values for the χ^2 test to judge whether the prompts showed statistically significant and also practically meaningful DIF. However, researchers have expressed different opinions about how to interpret the magnitude of R^2 change. Originally Cohen (1988, 1992) regarded R^2 values of 0.02, 0.13, and 0.26 as “small,” “medium,” and “large,” respectively, which was corresponding to Cohen’s d of

0.2, 0.5, and 0.8. According to Zumbo (1999), in order to classify an item as displaying DIF, the R^2 change after the group variable and the interaction term have been entered into the logistic regression should be at least 0.13. Jodoin and Gierl (2001), however, suggested R^2 differences of 0.035 for negligible DIF, 0.035 to 0.070 for moderate DIF, and greater than 0.070 for large DIF. So far, no study has investigated the use of R^2 change for the magnitude of DSF in the context of cognitive diagnostic assessment.

Therefore, in this study, R^2 change was not used as the criterion for the judgment of the existence of the DSF; however, R^2 values of Models 1 and 2 and R^2 change between Models 1 and 2 were reported in Appendix K.

5.3 Results

5.3.1 Existence of the DSF

Stage 1: Total Score as the Matching Variable. In Stage 1, the total score was used as the internal matching variable. Table 5.6 shows the -2 log-likelihood difference between the two models for each of the four skills being examined. The last column shows the -2 log-likelihood difference between Models 1 and2. A difference larger than the critical value of chi-square with one degree of freedom (i.e., $\chi^2_{(1, 0.05)} = 3.84$) indicates evidence of DSF. For skill 1 (vocabulary), the -2 log-likelihood difference was found to be 7.742, which is larger than the critical value of 3.84. For skill 2 (syntax), the -2 log-likelihood difference was found to be 2.274, which is smaller than 3.84. For skill 3 (extracting explicit information), the -2 log-likelihood difference was found to be 0.207, which is again smaller than 3.84. For skill 4 (connecting and synthesizing), the -2 log-likelihood difference was 3.775, which is a little bit smaller than 3.84. To summarize, DSF existed for skill 1.

Table 5.6

Summary of -2 Log-Likelihood Differences of Stage 1 Analysis

Skills	-2 log-likelihood of Model 1	-2 log-likelihood of Model 2	-2 log-likelihood difference between Models 1 and 2
Skill 1 (Vocabulary)	218.362	210.620	7.742*
Skill 2 (Syntax)	345.650	343.376	2.274
Skill 3 (Extracting explicit information)	341.480	341.273	0.207
Skill 4 (Connecting and synthesizing)	340.512	336.737	3.775

Note. * Larger than the critical value of $\chi^2_{(1, 0.05)} = 3.84$

Stage 2: Total Score and Gender as Matching Variables. In Stage 2, gender was controlled for as an external matching variable, in addition to the internal matching variable of the total score. For skills 1, 2, and 3, results were similar to those yielded in Stage 1. As shown in the last column of Table 5.7, for skill 1 (vocabulary), the -2 log-likelihood difference was 7.751, larger than 3.84. This indicated that DSF still existed for skill 1 when gender was controlled for. For skills 2 and 3, the -2 log-likelihood differences were smaller than 3.84 in both cases. However, for skill 4, the -2 log-likelihood difference was now 4.202, larger than 3.84. Therefore, DSF existed for skill 4 when gender was controlled for.

Table 5.7

Summary of -2 Log-Likelihood Differences of Stage 2 Analysis

Skills	-2 log-likelihood of Model 1	-2 log-likelihood of Model 2	-2 log-likelihood difference between Models 1 and 2
Skill 1 (Vocabulary)	217.416	209.665	7.751*
Skill 2 (Syntax)	338.462	335.838	2.624
Skill 3 (Extracting explicit information)	340.944	340.690	0.254
Skill 4 (Connecting and synthesizing)	335.724	331.522	4.202*

Note. * Larger than the critical value of $\chi^2_{(1, 0.05)} = 3.84$

To summarize, when only the total score was the internal matching variable, skill 1 exhibited DSF. However, when gender was also controlled for as an external matching variable in addition to the total score, both skill 1 and skill 4 exhibited DSF.

As shown in Appendix K, the R^2 change between Model 1 and Model 2, however, was rather small, being less than 0.01 in all cases. The smallness of the R^2 change was probably because each of the four skills was highly correlated to the total score, and thus after the total score had been entered into Model 1, much of the variance of the dependent variable had already been accounted for. Therefore, when language group was further entered into Model 2, there was not much variance left.

5.3.2 Interpretation of Logistic Regression Coefficients

The directions and values of the logistic regression coefficients for the four skills are interpreted as follows:

Skill 1 (Vocabulary). As shown in Table 5.8, when only the total score was used as the matching variable, language group was a statistically significant predictor, with a *p*-value of 0.007 and an odds ratio (i.e., Exp (β)) of 2.971. When gender was controlled for in addition to total score, Table 5.9 shows that language group was still a statistically significant predictor, with a *p*-value of 0.006 and an odds ratio of 2.986. However, Table 5.9 shows that gender itself was not a statistically significant predictor, with a *p*-value of 0.329.

Table 5.8

Regression Coefficients for Skill 1 when Matched on Total Scores

	β	S.E.	Wald	df	Sig.	Exp(β)
Language	1.089	0.401	7.381	1	0.007*	2.971
Total score	1.297	0.130	99.726	1	< 0.001*	3.657
Constant	-18.204	1.811	101.095	1	< 0.001*	0.000

*Note. * p < 0.05.*

Table 5.9

Regression Coefficients for Skill 1 when Matched on Total Scores and Gender

	β	S.E.	Wald	df	Sig.	Exp (β)
Language	1.094	0.401	7.447	1	0.006*	2.986
Total score	1.293	0.130	99.615	1	< 0.001*	3.642
Gender	-0.350	0.358	0.954	1	0.329	0.705
Constant	-17.946	1.815	97.774	1	< 0.001*	0.000

Note. * $p < 0.05$.

To summarize, given the same overall English reading ability, the odds that the Romance group would have mastery of vocabulary skill was about 3 times as large as the odds for the East Asian group regardless of gender.

Skill 2 (Syntax). As shown in Table 5.10 and Table 5.11, language group was not a significant factor whether gender was controlled for or not. Contrary to the hypothesized direction, the negative β coefficients for language group in both tables indicated a potential trend in the sample that given the same overall English reading ability, the East Asian group would be more likely to have mastery of syntax than was the Romance group. However, there is insufficient evidence to conclude the existence of language group difference regarding the skill of syntax.

In addition, as shown in Table 5.11, gender was a statistically significant predictor, with a *p*-value of 0.007. In order to further examine the effects of gender, as shown in Table 5.12, the language group variable was removed from the logistic regression. The results showed that gender remained a statistically significant predictor, with a *p*-value of 0.009. To summarize, given the same overall English reading ability, the odds for female ESL learners to have mastery of the syntax skill was about twice as large as the odds for male ESL learner regardless of native language.

Table 5.10

Regression Coefficients for Skill 2 when Matched on Total Scores

	β	S.E.	Wald	df	Sig.	Exp (β)
Language	-0.482	0.323	2.235	1	0.135	0.617
Total score	0.840	0.069	149.763	1	< 0.001*	2.317
Constant	-10.895	0.875	154.947	1	< 0.001*	0.000

Note. * $p < 0.05$.

Table 5.11

Regression Coefficients for Skill 2 when Matched on Total Scores and Gender

	β	S.E.	Wald	df	Sig.	Exp (β)
Language	-0.522	0.326	2.574	1	0.109	0.593
Total score	0.868	0.071	147.808	1	< 0.001*	2.382
Gender	0.788	0.293	7.227	1	0.007*	2.199
Constant	-11.707	0.969	145.955	1	< 0.001	0.000

Note. * $p < 0.05$.

Table 5.12

Regression Coefficients for Skill 2 when Language Group was Removed

	β	S.E.	Wald	df	Sig.	Exp(β)
Total score	0.853	0.070	149.122	1	< 0.001*	2.346
Gender	0.767	0.292	6.898	1	0.009*	2.153
Constant	-11.635	0.962	146.265	1	< 0.001*	0.000

Note. * $p < 0.05$.

Skill 3 (Extracting Explicit Information). As shown in Table 5.13 and Table 5.14, language group was not a statistically significant predictor whether gender was controlled for or not. The negative β coefficients for language group in both tables indicated a potential trend that given the same overall English reading ability, the East Asian group was more likely to have mastery of extracting explicit information than was the Romance group. However, there is insufficient evidence to conclude any significant language group difference. Further, as shown in Table 5.14, gender itself was not a statistically significant predictor, with a *p*-value of 0.446.

Table 5.13

Regression Coefficients for Skill 3 when Matched on Total Scores

	β	S.E.	Wald	df	Sig.	Exp (β)
Language	-0.147	0.332	0.195	1	0.659	0.864
Total score	0.919	0.074	155.903	1	< 0.001*	2.508
Constant	-10.819	0.859	158.780	1	< 0.001*	0.000

Note. * $p < 0.05$.

Table 5.14

Regression Coefficients for Skill 3 when Matched on Total Scores and Gender

	β	S.E.	Wald	df	Sig.	Exp (β)
Language	-0.167	0.332	0.253	1	0.615	0.846
Total score	0.924	0.074	156.133	1	< 0.001*	2.519
Gender	0.217	0.284	0.581	1	0.446	1.242
Constant	-10.996	0.891	152.339	1	< 0.001*	0.000

Note. * $p < 0.05$.

Skill 4 (Connecting and Synthesizing). As shown in Table 5.15, language group was not a statistically significant predictor, with a *p*-value of 0.055, though the negative β coefficient (i.e., -0.643) indicated a potential trend that given the same overall English reading ability, the East Asian group was more likely to have mastery of connecting and synthesizing information than was the Romance group. However, when gender was controlled for, as shown in Table 5.16, language group became a statistically significant predictor, with a *p*-value of 0.043 and an odds ratio of 0.507.This indicated that given the same overall English reading ability and gender, the odds that the Romance group would have mastery of the skill of connecting and synthesizing information was only half as large as the odds for the East Asian group.

In addition, as shown in Table 5.16, gender itself was a statistically significant predictor, with a *p*-value of 0.024. In order to further examine the effects of gender, language group was removed from the logistic regression. As shown in Table 5.17, gender was still a significant predictor, with a *p*-value of 0.031. To summarize, given the same overall English reading ability,

the odds that female ESL learners would have mastery of the skill of connecting and synthesizing information was about 1.9 times as large as the odds that male ESL learners would have mastery of this skill regardless of native language group.

Table 5.15

Regression Coefficients for Skill 4 when Matched on Total Scores

	β	S.E.	Wald	df	Sig.	Exp (β)
Language	-0.643	0.335	3.681	1	0.055	0.526
Total score	0.908	0.074	152.440	1	< 0.001*	2.480
Constant	-11.092	0.888	156.029	1	< 0.001*	0.000

Note. * $p < 0.05$.

Table 5.16

Regression Coefficients for Skill 4 when Matched on Total Scores and Gender

	β	S.E.	Wald	df	Sig.	Exp(β)
Language	-0.679	0.336	4.090	1	0.043*	0.507
Total score	0.928	0.076	150.882	1	< 0.001*	2.529
Gender	0.659	0.293	5.065	1	0.024*	1.932
Constant	-11.713	0.962	148.169	1	< 0.001*	0.000

Note. * $p < 0.05$.

Table 5.17

Regression Coefficients for Skill 4 when Language Group was Removed

	β	S.E.	Wald	df	Sig.	Exp (β)
Total score	0.904	0.073	154.351	1	< 0.001*	2.469
Gender	0.629	0.291	4.657	1	0.031*	1.875
Constant	-11.568	0.944	150.107	1	< 0.001*	0.000

Note. * $p < 0.05$.

5.3.3 Summary of the Results

Based on qualitative data from a group of ESL learners and based on the literature, the specific reading subskill differences between East Asian and Romance ESL learners were hypothesized in Chapter 4. They were tested via a series of DSF analyses through logistic regression techniques. The following provides a summary of the results:

- Hypothesis 1: There is DSF for skill 1 (vocabulary) favoring the Romance group.
- Hypothesis 2: There is DSF for skill 2 (syntax) favoring the Romance group.
- Hypothesis 3: There is DSF for skill 3 (extracting explicit information) favoring the East Asian group.
- Hypothesis 4: There is DSF for skill 4 (connecting and synthesizing) favoring the East Asian group.

Hypothesis 1 was supported. It seems that given the same overall English reading ability, it is more likely for the Romance group to have mastery of vocabulary skill when compared to the East Asian group.

Neither hypothesis 2 nor hypothesis 3 was supported. Given the same overall English reading ability, there was insufficient evidence to conclude that the East Asian group and the Romance group would have different mastery of the skill of syntax or different mastery of the skill of extracting explicit information.

Hypothesis 4 was supported only when gender was controlled for. Given the same overall English reading ability and gender, it is more likely that the East Asian group would have mastery of the skill of connecting and synthesizing information compared to the Romance group. However, when gender was not controlled for, language group was no longer a statistically significant predictor, with a p-value of 0.055.

In addition, though gender was not included in the hypotheses, it was found that given the same overall English reading ability, female ESL learners were more likely to have mastery of the skill of syntax and the skill of connecting and synthesizing information regardless of native language.

CHAPTER 6

DISCUSSION, IMPLICATIONS, LIMITATIONS AND FUTURE RESEARCH

With the purpose of providing diagnostic information for second-language reading instruction, this dissertation used a sequential mixed-methods design to examine the differences between two groups—those with an East Asian language background and those with a Romance language background—in regard to the reading subskills represented in the MELAB reading test. The study established specific hypotheses regarding the subskill differences between the two groups by using a grounded theory approach that draws on think-aloud reports from a sample of ESL students. Via a series of DSF analyses through logistic regression techniques, the hypotheses were tested by comparing the subskill profiles of the two groups. The subskill profile of each examinee was identified by applying the item-skill Q-matrix to a Fusion Model of cognitive diagnostic modeling.

This chapter discusses the overall findings and its implications for the second-language reading instruction and cognitive diagnostic assessment of reading. Finally, limitations of the study and important areas for future research are also addressed.

6.1 Discussion of the Overall Findings

The following briefly discusses the overall findings of the DSF analysis. Hypothesis 1, which postulated that there is DSF for vocabulary favoring the Romance group, was supported. The DSF analysis shows that given the same overall English reading ability, it is more likely for Romance ESL learners to have mastery of vocabulary compared to East Asian ESL learners.

This result is also well supported in the literature and the theory about how linguistic skills are transferred from L1 to L2: the closer one's L1 and L2 are, the easier it is for the ESL learner's L1 skills to be transferred to that learner's L2 skills. Due to the influence of Latin, English and Romance languages share many linguistic features. One such distinctive commonality is the use of Roman alphabets, in which each letter represents a phoneme. However, Chinese characters and Japanese Kanji belong to logographic systems, in which each symbol maps into a morpheme. Although Korean Hangul is alphabetic, it does not use Roman alphabets and requires assembling individual symbols into syllable blocks (Taylor & Taylor, 1995). The different writing systems of the respective languages may lead to different word recognition processes (Koda, 2005). Due to the differences between how words are recognized in East Asian languages as compared to English, East Asian ESL learners are likely to experience more difficulty in recognizing English words than are Romance ESL learners.

Hypothesis 2, which postulated that there is DSF for the skill of syntax favoring the Romance group, was not supported. In fact, contrary to the hypothesized direction, the negative β coefficient for language group predictor indicates a potential trend that given the same overall English reading ability, the East Asian group may be more likely to have mastery of syntax than the Romance group is, even though this result was not statistically significant. It has to be admitted that the syntax difference between the two groups' native languages is not as clear-cut as the vocabulary difference. For instance, both Chinese and English primarily have Subject–Verb–Object as the word order, though Chinese relies less on word order than English does. Even though Chinese ESL learners are challenged by a large number of English syntax features which do not exist in Chinese, the common S–V–O might reduce their disadvantage in the mastery of syntax. If only Korean and Japanese ESL learners, whose native languages have Subject–Object–

Verb as the word order, had been compared to Romance ESL learners in the DSF analysis, a clearer difference regarding syntax may have been detected. Another possible reason is the intensive training in grammar that East Asian learners receive during their English learning and test preparation processes. The grammar-translation method has greatly influenced English instruction in East Asian countries, and the learning of grammar has received considerably more attention than have other communicative skills, such as listening and speaking. Therefore, focused training may have more than compensated for East Asian learners' initial disadvantages in mastering English syntax.

Hypothesis 3, which postulated that there is DSF for the skill of extracting explicit information favoring the East Asian group, was not supported. The think-aloud verbal reports indicated that East Asian learners were more skilled at extracting explicit information at the local level than were Romance learners, probably due to East Asian learners' training in and the experience with multiple-choice reading tests. The nonsignificant result of the DSF analysis might be due to a possible interaction between vocabulary knowledge and extracting explicit information. When extracting explicit information at the local level, readers match lexical and/or syntactic information in the question to those in the relevant part of the text, and they may also identify or formulate a synonym or a paraphrase of the literal meaning of a word, phrase, or sentence in the relevant part of the text. East Asian learners may identify and locate information more efficiently compared to their Romance counterparts; however, their relative disadvantage in vocabulary may reduce this efficiency. Still, it should be noted that the negative β coefficient associated with language group does trend in the correct direction in the sample data as stated in the hypothesis. Specifically, the odds for the East Asian group to have mastery of the skill of extracting explicit information was about 1.25 times as large as the odds for the Romance group.

Hypothesis 4, which postulated that there is DSF for the skill of connecting and synthesizing information favoring the East Asian group, was supported when gender was controlled for in the DSF analysis. The think-aloud activity showed that East Asian learners were very efficient at connecting and synthesizing information, probably due to their intensive training in taking multiple-choice reading tests. However, when only total scores were controlled for, the language group was not a statistically significant predictor with a *p*-value of 0.055, even though the negative β coefficient did trend in the correct direction in the sample data. When gender was controlled for in addition to the total score, however, the language group did become a statistically significant predictor with a *p*-value of 0.043. It seems that given the same overall English reading ability and gender, it is more likely for the East Asian group to have mastery of the skill of connecting and synthesizing information than for the Romance group. The skill of connecting and synthesizing is a very broad category. As defined in Chapter 3, when connecting and synthesizing, sometimes it is only necessary for readers to connect ideas from two adjacent sentences in the same paragraph, whereas at other times readers have to synthesize information from the overall passage. Therefore, skill 4 involves different reading components at different levels. The heterogeneity of these components may make any group differences more difficult to detect. If a more fine-grained category of subskill had been defined, such as "identifying the main idea," any group differences may have been clearer.

In addition to language group differences, gender differences also emerged. As shown in the DSF analysis, given the same overall English reading ability, female ESL learners were more likely to have mastery of syntax than were male ESL learners. Also, for those with the same overall English reading ability, female ESL learners were more likely to have mastery of connecting and synthesizing information than were males. When gender was controlled for,

given the same overall English reading ability, East Asian ESL learners were more likely to have mastery of connecting and synthesizing information than the Romance ESL learners. Overall, it seems that female ESL learners with an East Asian language background were more skilled than other learners at connecting and synthesizing information.

The DSF results agree with the literature on gender differences in reading. It has been found that female students generally perform better in reading than do male students (Klinger, Shulha, & Wade-Woolley, 2009; Logan & Johnston, 2009). Researchers have investigated the factors causing gender differences from different perspectives. Using cognitive process taxonomy, Halpern (2000, 2004) found that female students more rapidly access phonological, semantic, and episodic information from long-term memory, whereas male students perform better on tests of verbal analogies, which involve mapping verbal relationships in working memory, as well as tasks involving transformations in visuo-spatial working memory. Halpern (2006) also argued that girls tend to receive better grades in school, especially when the teacher's test material closely resembles what was taught. Given the fact that English reading instruction is intensively test-oriented in East Asian countries, female East Asian students thus may gain an advantage in reading tests compared to other groups. Therefore, the observed DSF favoring female students may be partially attributable to the English instruction they have received.

6.2 Implications for Second-Language Reading Instruction

With traditional unidimensional IRT, examinee reading performance is usually expressed as a single score. This score helps to rank examinees along a single continuum, but it provides little diagnostic information. Two examinees may earn the same total score but have different strengths and weaknesses. Understanding their different characteristics helps to facilitate

learning and instruction. As a result of cognitive diagnostic analyses, examinees are assigned multidimensional skill profiles by being classified as masters versus non-masters of each skill involved in the test, so that fine-grained diagnostic feedback can be provided.

Using a cognitive diagnostic approach and a DSF procedure, this dissertation confirmed the hypothesis that given the same overall English reading ability it is more likely for Romance ESL learners to have mastery of the skill of vocabulary than for East Asian ESL learners. Also, given the same overall English reading ability and gender, it is more likely for East Asian ESL learners to have mastery of the skill of connecting and synthesizing information than for Romance ESL learners. In addition, given the same overall English reading ability, female ESL learners are more likely to have mastery of the skill of syntax and the skill of connecting and synthesizing information. The following sections suggest some instructional strategies for addressing specific weaknesses in ESL learners' reading skills that have been observed in this dissertation.

6.2.1 Vocabulary

Numerous studies have shown that word recognition is a major predictor of later reading abilities (e.g., Adams, 1990, 1999; Juel, 1988; Perfetti, 1999, 2007; Perfetti, Landi & Oakhill, 2005). Lack of vocabulary has been identified as the principal obstacle in reading comprehension. A rule of thumb is that readers must know 95% of the words in a text if they are to read it successfully (Grabe, 2009); however, this is rarely true for ESL learners. East Asian ESL learners are especially challenged by English vocabulary due to the vast difference between the writing system of their native languages and that of English.

Vocabulary learning has been the focus for East Asian ESL learners; however, as discussed in Chapter 4, most East Asian students learn English in order to take tests for

admission or employment purposes. Some testing organizations, such as the College English Test (CET) Committee in China, publish manuals with all the words that might appear in their tests. It by no means stretches the imagination to expect that teachers explicitly teach those words in the classroom, students concentrate on learning the words in such manuals, and textbooks are also designed to include exercises on those words. In fact, East Asian participants in the think-aloud activity reported that they "recited" these words every day for test preparation. In addition to explicitly learning words from the classroom and mechanically reciting words outside class, many East Asian students rarely acquire words incidentally, such as reading for entertainment, watching English movies, or having conversations with native English speakers. As a result of lack of experience in such activities, many East Asian ESL learners' vocabulary knowledge tends to be isolated and mechanical; that is, they tend not to fully understand either the words' usage or connotations.

A large number of instructional strategies are available to ESL teachers (Lems, Miller, & Soro, 2010). In light of the findings of this dissertation, extensive reading and increasing phonological awareness are especially recommended to help improve East Asian ESL learners' vocabulary skill.

Extensive Reading. ESL reading instruction in East Asian countries tends to focus on intensive reading (Powell, 2005). With intensive reading, readers take a text, study it line by line, and refer frequently to a dictionary in order to understand the grammar and vocabulary of the text (Palmer, 1917). Intensive reading is usually conducted in the classroom, and the reading materials are short and formal. It is also followed by various drills and exercises in order for students to practice what has been emphasized in the instruction. Intensive reading is necessary and important; however, overly or solely relying on intensive reading is restrictive. Students are

only exposed to a small amount of text, and even if they have "learned" a word, they have very limited opportunities to encounter or use the word in a variety of contexts. Also due to the limited exposure to the language, they lack the ability to acquire vocabulary on their own.

Extensive reading, however, is different in that students read a large amount of longer, easy to understand materials relatively fast, mostly out of the classroom and according to their own pace and schedule. The purpose is overall understanding rather than word-by-word decoding or grammar analysis. Considerably greater exposure to authentic reading in English will help students "overcoming the many L1–L2 differences that exist for L2 reading development" (Grabe, 2009, p. 150). This is especially helpful for East Asian ESL learners.

Extensive reading is beneficial for reading proficiency, especially in vocabulary learning (Hitosugi & Day, 2004; Horst, 2005; Kweon & Kim, 2008). Stanovich (1986) makes a strong argument for a reciprocal causal relationship between reading and vocabulary; i.e., vocabulary growth leads to improved reading comprehension, and amount of reading leads to vocabulary growth. Those who have large vocabularies can read more material, and more reading can help them acquire a larger vocabulary. This is the "rich get richer through reading" idea. Consistent exposure to English texts through extensive reading can help East Asian ESL learners gain more vocabulary in context. This, in turn, can help them become more efficient readers.

Extensive reading also has the potential to train ESL students to become proficient at acquiring vocabulary on their own. Beginning ESL learners largely rely on explicit instruction to learn words, so as to build their basic vocabulary. However, many East Asian ESL learners, some even after over 10 years of English learning in the classroom, still lack the ability to learn English vocabulary on their own. Krashen (1981) argued that students can acquire language on their own, if (a) they receive enough exposure to comprehensible language and (b) it is done in a

relaxed, stress-free atmosphere. This gives theoretical support to using extensive reading to strengthen students' self-directed learning ability.

To summarize, extensive reading can be an important complement to the test-oriented intensive reading instruction that currently prevails in East Asian countries. In this way, ESL learners may develop implicit understanding regarding when and how words are used, and they may also become independent learners of English vocabulary.

Increasing Phonological Awareness. One reason for East Asian ESL students' difficulty in English word recognition is their lack of phonological awareness. East Asian ESL learners have been found to be less sensitive to phonological information in English word recognition, compared to those with a Roman alphabetic L1 background (Biederman & Tsao, 1979; Brown & Haynes, 1985; Koda, 1990; Tzeng & Wang, 1983). Some of the East Asian students in the think-aloud activity in this dissertation showed a lack of phonological awareness, which negatively affected their reading performance.

Phonological awareness refers to the reader's awareness of the phonological structure, or sound structure, of the spoken word (Gillon, 2004; Stahl & Murray, 1994). It is regarded as an important and reliable predictor of later L1 reading ability (e.g., Ball, 1997; Ehri et al., 2001). As indicated by Baddeley (2006), storage, rehearsal, and reinforced memory of new words in phonological form in the working memory is the foundation of all vocabulary learning. Although the effects of phonological awareness are not fully investigated in L2 reading, it plays an important role in L2 reading development (Bernhardt, 2011; Grabe, 2009). Therefore, it is very important to aim at increasing East Asian ESL students' phonological awareness in order to help them achieve more effective word recognition.

Different strategies are available for increasing students' phonological awareness. One option is through explicit classroom instruction. ESL instructors may use some tasks to help students improve their ability in this regard. For example, oddity tasks involve the detection of similar or dissimilar sounds, deletion and substitution tasks require the manipulation of sounds, and segmentation activities teach how to segment at multiple phonological levels (Anthony, Lonigan, Driscoll, Phillips, & Burgess, 2003). Explicit instruction can help students learn the phonological rules of English words in a structured and effective way.

In addition, oral reading, or reading aloud after class, helps students build phonological awareness on their own. East Asian ESL classrooms are usually quiet, with the instructor talking while the students listen. Also because oral reading fluency is usually not tested on large-scale assessments, many students rarely read aloud on their own. During reading aloud, in addition to visual processing, readers are actively using the phonological cues by hearing the words in context. This helps to improve not only vocabulary learning and reading ability, but also listening and speaking ability.

To summarize, East Asian ESL learners are at a disadvantage in English word recognition due to their lack of phonological awareness. Explicit training in the classroom and also oral reading outside class should help them increase their phonological awareness and improve their English reading ability overall.

6.2.2 Syntax, Connecting and Synthesizing

The DSF study shows that given the same overall reading ability, female ESL learners were more likely to have mastery of syntax and connecting and synthesizing information than were male ESL learners. Males with a Romance language background were especially challenged regarding the skill of connecting and synthesizing information. The following

sections summarize some instructional strategies for addressing gender differences in reading. It also suggests teaching text structure to ESL learners (especially males with a Romance language background) to help them improve their ability to connect and synthesize information during reading.

Instructional Strategies to Address Gender Differences in Reading. Many studies drawing on a range of perspectives have investigated gender differences in reading. Neuroimaging studies suggest that male and female students have different patterns of functioning activation during reading (Pugh et al., 1996; Shaywitz et al., 1995). Male and female students also have been found to use different reading strategies (Thompson, 1987) and even to benefit from different types of reading instruction (Johnston, Watson, & Logan, 2009). Another finding is that female students have a more positive attitude toward reading than do male students (McKenna, Kear, & Ellsworth, 1995; Sainsbury & Schagen, 2004). Female students have also been found to read more frequently than males do (Hall & Coles, 1999; Mullis, Martin, Kennedy, & Foy, 2007). Furthermore, it is generally reported that male students have poorer attention during literacy lessons than female students (Logan, Medford, & Hughes, as cited in Logan & Johnston, 2010). A general trend among all these studies is that female students perform slightly better in reading than do males (Chiu & McBride-Chang, 2006; Mullis, Campell, & Farastrup, 1993).

While most of the above-cited studies are about L1 reading in K-12 settings, only a small number of studies have been conducted on gender differences in L2 reading. A consistent finding is that in learning a foreign language female students use strategies more actively than do males (Dreyer & Oxford, 1996; Goh & Foong, 1997). Oxford (1993) suggests that females tend to be higher L2 achievers mainly because of their higher level of strategy use, which is also supported

by Sheorey and Mokhtari (2001). Researchers (e.g., Bügel & Buunk, 1996; Brantmeier, 2003; Pae, 2004) have also investigated whether observed gender differences may be confounded with the content of reading materials. It thus has been suggested that gender-free passages should be selected to reduce gender differences in L2 reading tests.

As discussed, studies on gender differences in L2 reading mainly focus on reading strategies and reading topics. ESL instructors do not seem to be as well-informed about gender differences as L1 reading instructors are. Given the similarities between L1 and L2 reading processes (Grabe, 2009), the following section provides ESL instructors with some general strategies for addressing male students' disadvantages in reading, which are mainly drawn from the literature on L1 reading.

ESL instructors may choose classroom activities that are especially effective for male students. As suggested by Connell and Gunzelmann (2004), instructors can provide activities that require the use of visual–spatial strengths, given that males have been reported to excel in visual–spatial tasks. In addition, instructors could integrate physical activity and allow time for movement, because males are more prone to movement and physical activity than females. Other suggestions include providing opportunities for male students to demonstrate learning through the use of hands-on materials and maximizing student use of technology in the instructional process. In general, these suggestions focus on providing a supportive learning environment for male students to develop their reading skills.

Other strategies have also been found useful for addressing affective factors regarding reading (Younger & Warrington, 2005). For example, individual goal-setting and mentoring could be provided to male students. Some sociocultural strategies could be used to help increase students' self-confidence, to reduce their non-conformist behaviors, and to integrate them within

school life. Furthermore, in order to boost male students' interest in reading and motivation to read, instructors could choose reading materials that are more appealing to males, such as historical nonfiction, adventure tales, and stories about sports and war (Bauerlein & Stotsky, 2005). Setting up book clubs has also been recommended as an effective way to engage male students.

These suggestions are intended to help improve male ESL learners' reading skills, but it is important to note that the reading achievement of female students should not be negatively impacted. ESL instructors need to select appropriate strategies depending on the age group and the specific class setting.

Explicit Instruction of Text Structure. As defined in Chapter 3, the skill of connecting and synthesizing includes the following components: (a) understand the relationship between sentences and the organization of the text using cohesion and rhetorical organization knowledge; (b) synthesize information presented in different sentences or parts of the text; (c) identify the main idea, theme, or concept; and (d) skim the text for the gist. The DSF study showed that male ESL learners, Romance ESL learners, and especially male Romance ESL learners were challenged in connecting and synthesizing information during reading. Though many instructional strategies are available, explicit instruction of text structure is particularly recommended. Text structure focuses on helping readers understand how the information in a text is organized (Taylor, 1992), which may lead to improvements in connecting and synthesizing information as well as in overall reading comprehension.

Text structure theory addresses how the overall structure of a text may affect reading comprehension (e.g., Meyer, 1975; Meyer & Rice, 1982; Meyer et al., 2010). For example, five basic types of expository rhetorical organization have been identified: comparison, problem–

and–solution, cause–and–effect, sequence, and description (e.g., Meyer, 1985). Subsets of these structures (e.g., comparison with description) build on each other to make the logical structure of the text, which can be shown graphically. Ideas at the top levels of the hierarchical, logical structure are the main ideas (Meyer, Young, & Bartlett, 1989).With a clear understanding of the text structure, students can more effectively understand the main idea of the text and more systematically retrieve details from memory later (Meyer et al., 2010). Using the text structure strategy can improve reading comprehension for readers with adequate vocabulary skills, but poorer reading comprehension skills (Meyer, Brandt, & Bluth, 1980) or adults reading unfamiliar, expository texts (Meyer & Poon, 2001). In general, teaching text structure to children or adults across the life span can lead to improvements in reading comprehension (e.g., Meyer et al., 2010; Williams et al., 2005).

Despite consistent findings on the positive effects of teaching text structure, text structure strategy has not received much attention in the domain of L2 reading (Chang, 2002). An early study conducted by Carrell (1985) reported increased recall of information from the text after text structure strategy was taught to intermediate-level ESL learners. However, subsequently, only a few studies (e.g., Carrell, 1992; Chen, 1990; Chu, 1999) investigated this issue. Still, a general trend found in these studies is that structure-aware readers consistently outperformed structure-unaware readers in reading comprehension and recall (Chang, 2002). It is, thus, of crucial importance to teach ESL learners text structure strategy.

Many approaches are available to teach students text structure strategy. Meyer and her colleagues have taught students to use signaling words to help them recognize the different structures of expository texts (e.g., Meyer, 1985; Meyer & Rice, 1982; Meyer & Poon, 2001). For example, words and phrases such as *because* and *in order to* indicate causation; words such

as *problem, solution, answer, solve* indicate problem–solution; words and phrases such as *in contrast, instead,* and *however*, indicate contrast and comparison. In particular, a web-based system called the Intelligent Tutoring of the Structure System (ITSS) has been designed to provide text structure training to students at different grade levels (e.g., Meyer & Wijekumar, 2007). Other researchers have proposed teaching students text structure strategy by using a graphic organizer (e.g., Berkowitz, 1986; Gallini & Spires, 1995) and writing hierarchical outlines (e.g., Slater, 1985; Taylor & Beach, 1984). Moreover, students can be taught to use headings, subheadings, and topic sentences in order to understand the structure of a text (Seidenberg, 1989). Overall, a large number of methods and resources are available for text structure instruction.

Most of the studies referenced are about expository text structure, but another important text structure is narrative structure. For example, a story generally consists of characters, setting, plot, attempt, reaction, outcomes, and ending (Fitzgerald, 1989; Mandler & Johnson, 1977; Taylor, 1992). Numerous studies have shown that teaching students the narrative structure of storytelling improves their comprehension (e.g., Singer & Donlan, 1982; Pearson & Fielding, 1991). In particular, it has been suggested that story grammar instruction be used to teach narrative structure, through which students are given a framework to use to help them understand different elements of the story (Gurney, Gersten, Dimino, & Carnine, 1990). This approach has been found especially effective for improving students' understanding of the main ideas of the story (Idol & Croll, 1987).

Most of the above-cited studies investigated L1 reading. However, given the similarities between L1 and L2 reading and findings on the positive effects of using text structure strategy in L2 reading (e.g., Carrell, 1985; Chang, 2002), explicit instruction of text structure appears to be

very promising in strengthening ESL learners' ability to connect and synthesize information and their overall reading comprehension.

6.3 Implications for Cognitive Diagnostic Assessment of Reading

6.3.1 Developing Diagnostic Assessment

As summarized by DiBello, Roussos, and Stout (2007), a systematic cognitive diagnostic assessment involves six steps: (i) describing assessment purpose; (ii) describing skill space; (iii) developing assessment tasks; (iv) specifying psychometric model; (v) performing model calibration and evaluation; and (vi) score reporting. However, currently, very few large-scale tests are designed with a cognitive diagnostic purpose; therefore, in most application studies, a preexisting test is analyzed with a complex cognitive diagnostic model (CDM). Some successful retrofitting studies, such as Jang (2005) and Klein et al. (1981), have demonstrated that it is possible to extract richer diagnostic information than the test was designed to elicit. Another benefit of retrofitting is that this practice can deepen our understanding of the construct being tested. However, a major challenge involved in retrofitting is the time-consuming process of constructing the post-hoc Q-matrix. In addition, sometimes calibrating a unidimensional preexisting test with a multidimensional CDM may not be psychometrically efficient (Haberman & von Davier, 2007). In order for a test to generate detailed diagnostic feedback, it is essential that it be built for a skills-based diagnostic purpose (DiBello, Roussos, & Stout, 2007).

The evidence-centered design (ECD) developed by Mislevy and colleagues (e.g., Mislevy, 1994; Mislevy, Steinberg, & Almond, 2003) offers a good framework for developing diagnostic assessment. The general form of the ECD is as follows:

> A construct-centered approach [to assessment design] would begin by asking what complex of knowledge, skills, or other attribute should be assessed.... Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (Messick, 1994, p. 17)

As stated by Mislevy, Almond, and Lukas (2003), good assessment tasks are not built in isolation. Instead, the starting point of designing a complex assessment is to determine the inferences we wish to make. Then, we need to know the observations that are necessary in order to make the inferences, the situations that evoke those observations, and the chain of reasoning that connects them (Missick, 1994). As a result, successful test development needs coordination among different specialists, such as statisticians, item developers, and interface designers.

Despite the well-structured ECD framework, understanding the skill space of the construct being tested is essential for developing skills-based cognitive assessment. However, due to the complexity of reading comprehension, its cognitive processes are not fully understood (Lee & Sawaki, 2009b). What's more, as reported by Li (1992), even when test developers anticipate that certain skills will be required by an item, test takers may actually use different skills in their reading process. Similarly, Anderson, Bachman, Perkins, and Cohen (1991) found no statistically significant relationship between learner-reported skills and the intended purposes of the test questions.

Think-aloud protocols (or cognitive interviews) have been found valuable to help test developers gain more in-depth understanding of the cognitive processes underlying the construct being tested and also how test items function with the test takers. For instance, Kaliski, Huff, and

Thurber (2011) used think-aloud protocols to inform the development of a large-scale World History placement test using the ECD framework. Fifteen high school students who took the exam in the past year participated in the think-aloud activity. Analysis of their verbal reports revealed whether the tasks were eliciting evidence of the intended targets of measurement and identified what item features contributed to the sources of complexity within the items. In addition to being used for understanding the skill space of the construct, think-aloud protocols are helpful in various stages of test development, involving test design, test development, and test validation (Almond et al., 2009). For instance, Johnstone, Bottsford-Miller, and Thompson (2006) found that think-aloud protocols could successfully detect design issues in a large-scale mathematics test, such as "unclearly defined constructs, inaccessibility of items, unclear instructions, incomprehensible language, and illegible text and graphics." Readers can refer to Almond et al. (2009) for a thorough discussion on the use of think-aloud protocols for different phases of test development.

To summarize, test developers play critical roles in the process of cognitive diagnostic assessment. The ECD offers a general framework that test developers could follow in designing tests for diagnostic purposes. In particular, think-aloud protocols can be beneficial at different stages of test development.

6.3.2 Selecting Diagnostic Models

With a large number of CDMs available (62 models as listed by Fu & Li, 2007), the question is "which one should I use" for reading tests (Jiao, 2009, p. 55). Lee and Sawaki (2009b) and Rupp and Templin (2008) presented good reviews on the available CDMs and software. In addition to a full understanding of the conditions and assumptions of the CDMs, one major decision is to make the choice between compensatory and non-compensatory models for

diagnostic analysis of reading tests.

CDMs allow for both compensatory and non-compensatory relationships among subskills. Some of the widely known non-compensatory models are the Rule Space Model, the Attribute Hierarchy Method, the DINA, the NIDA, the HYBRID model, and the Reparameterized Unified Model (RUM) also known as the Fusion Model. The DINO model and the NIDO model are compensatory. Non-compensatory models have been more popular in cognitive diagnostic analysis, probably because they can generate more fine-grained diagnostic information than compensatory models can. Early applications of CDMs were mostly with mathematics, where the solution is usually broken down into a series of steps. All the steps must be successfully performed in order to solve the mathematics problem. Therefore, it is generally agreed that non-compensatory models are appropriate for mathematics tests (Roussos, Templin, & Henson, 2007).

However, the question of whether we should use non-compensatory or compensatory models with reading tests does not have a clear-cut answer. Lee and Sawaki (2009a) applied three different CDMs to IBT TOEFL reading and listening data, and found that non-compensatory and compensatory models yielded similar results. Jang (2005) also found that reading skills involved in the IBT TOEFL appeared to be a mixture of non-compensatory and compensatory interactions. In the literature on reading, Stanovich (1980) proposed a compensatory-interactive model of reading. A major claim of the model is that "a deficit in any particular process will result in a greater reliance on other knowledge source, regardless of their level in the processing hierarchy" (p. 32). However, according to the simple view of reading (Gough & Tunmer, 1986), reading comprehension (RC) is the product of comprehension (C) and decoding (D), i.e., RC = C x D. The multiplication indicates a non-compensatory relationship. In

fact, except for extreme cases when examinee ability in one subskill is zero, the additive property of compensatory models is theoretically equivalent to the multiplicative property of non-compensatory models. No matter which model is used, the more skills the examinee acquires, the more likely it is that the examinee can correctly answer the item requiring those skills. Therefore, at a macro-level, whether a compensatory or non-compensatory model is used for reading tests is probably inconsequential.

However, if interested, we could empirically test the relationships between reading subskills using a log-linear approach (Henson, Templin, & Willse, 2008). Henson et al. reparameterized the cognitive diagnostic modeling family with a log-linear approach. In this way, estimation could be conducted with more commonly used software such as Mplus (Muthén & Muthén, 2010). An interaction term in the log-linear model indicates the relationship between the subskills. With this approach, it is not necessary to choose between a compensatory or non-compensatory model, and the relationship between subskills can vary across items. With more evidence for the robustness of the log-linear approach for cognitive diagnostic analysis, it may prove to be an effective estimation method.

6.3.3 Potential Use of Scale Scores

Cognitive diagnostic analysis via the Fusion Model or most other CDMs is usually technically challenging. It involves a principal dilemma: On the one hand, the use of the CDMs is especially helpful for classroom instructors. On the other hand, currently, only a small number of psychometricians are trained to use multidimensional CDMs. Therefore, an important task is to make the CDMs "absolutely opaque to classroom teachers, to coordinators of language education programs, and to other in-the-trenches educators" (Davidson, 2010, p. 106).

In order to reduce the sophistication involved in model calibration and thus maximize the advantages of the CDMs, one available option for classroom teachers and non-technical researchers is to use scale scores for the subskills (Henson, Templin, & Douglas, 2007). Given that the cognitive structure of a test is well validated, a scale score could be calculated by averaging the scores of the items associated with a given skill. Item scores may also be weighted while contributing to the sum score. With a simulation study, Henson et al. (2007) concluded that scale scores could be used to estimate the continuous posterior probability of mastery (PPM) with only a moderate reduction in the accuracy of the classification rates. The weighted sum score approach, which takes into consideration unequal contributions of the item scores, may be more appropriate for complex associations between skills and items.

As a post-hoc analysis, the scale scores for each skill in the present study were obtained by averaging the scores of the items requiring the skill. Spearman's rho between the average scale score and the average PPM extracted from the Fusion Model calibration for skill 1 (vocabulary), skill 2 (syntax), skill 3 (extracting explicit information), and skill 4 (connecting and synthesizing) was respectively 0.967, 0.90, 915, and 0.884 (Please see Appendix L for the scatter plots of scale scores and the PPMs). Jang (2005) also found high correlations between the scale scores and PPMs in her study, which was regarded as evidence for the validity of the Fusion Model calibration.

The Fusion Model as well as other IRT-based CDMs has the advantage of IRT models, such as being sample-independent and item- (or skill-) independent. The PPMs are probabilities of latent subskill mastery, whereas the scale scores are the observed skill scores. The relationship between the PPMs and scale scores is thus similar to the relationship between the IRT ability scores and the classical raw scores (Suen, personal communication, December 27, 2010). The

PPMs have more desirable psychometric features than do the scale scores; however, the scale scores can be an easy and quick way for less technically competent users to derive diagnostic information from a test with a clear cognitive structure.

6.4 Limitations and Future Research

6.4.1 Cognitive Diagnostic Analysis

As discussed, a primary limitation of this study is that of retrofitting the cognitive diagnostic modeling with the MELAB, which is originally an English proficiency test. Because the cognitive structure of the MELAB reading test was not clearly specified *a priori*, the retroactive Q-matrix construction process proved to be challenging and time-consuming.

A noticeable indeterminacy in the cognitive diagnostic analysis is the grain size of the subskills (Lee & Sawaki, 2009b). The more skills identified, the richer the diagnostic information that can be provided; however, including a high number of skills places a stress on the capacity of statistical modeling, given the fixed length of a test. Two major factors considered were the modeling capacity and the meaningfulness of the skill mastery profile. Hartz (2002) suggested that one skill should be assigned to at least three items to obtain sufficient information to estimate the skill with the Fusion Model. Gao (2006) suggested that 10 reading skill components underlying the MELAB reading test. However, given the fact that the MELAB consists of only 20 items, this study only involved five subskills, such as vocabulary, syntax, extracting explicit information, synthesizing and connecting, and making inferences. Skill 5 (making inferences) is an important higher-order reading skill. However, only 2 items were finally identified as requiring this skill. Even though this skill is of great interest, it was removed from the final analysis due to insufficient statistical information to estimate parameters related to

this skill. Another option might be collapsing skill 5 (making inferences) and skill 4 (connecting and synthesizing) into a single higher-order reading skill. Still, the best approach is to design a reading test with a clear cognitive structure in order to estimate the subskills of interest.

Jang (2009) and Sawaki, Kim, and Gentile (2009) commented on the skill granularity issue. For the same TOEFL reading test, Jang identified nine skills, whereas Sawaki et al. identified only four skills. In particular, Jang identified two vocabulary skills, one with and the other without the use of context clues, but Sawaki et al. included only one vocabulary skill. Sawaki et al. acknowledged that they had considered the two different approaches but decided not to include the context clues for two reasons. First, only when a reader is not sufficiently familiar with a word in question, using context clues is required as part of the process of responding to a vocabulary item. Also, though two vocabulary skills may help to extract more fine-grained diagnostic information, using two may not be feasible if a test includes only a small number of items requiring vocabulary as an essential skill. To summarize, as Jang (2009) suggested, decisions about the grain sizes of the subskills should be made by considering theoretical (construct representativeness), technical (availability of test items), and practical (purposes and context of using diagnostic feedback) factors.

Given this indeterminacy of the grain sizes, there are always alternative Q-matrices as a function of the definitions and categories of subskills (Lee & Sawaki, 2009b). The final Q-matrix used in this dissertation is thus not the only Q-matrix possible for the MELAB reading test used in this study. It is expected that a deeper understanding of the construct of reading and also of the statistical modeling approach will provide more evidence for the appropriate granularity of the subskills of reading in cognitive diagnostic research.

6.4.2 Grounded Theory Study

One limitation of the grounded theory study is the limited diversity of participants in the think-aloud activity. Only one Korean student with low-level English reading ability and one Japanese student with high-level English reading ability participated. It would be desirable to include at least one high-level Korean student and one low-level Japanese student. Additionally, no Romanian native speaker participated in the think-aloud activity; however, data from the Romanian examinees were included in the dataset for the DSF analysis (see Table 5.1). Qualitative studies do not attempt to have a representative sample, because the results of such studies are not intended to be generalized to a wider population. However, it is still important to include participants from more language groups in the think-aloud activity, given the fact that the hypotheses generated from this study were later tested with a larger dataset.

Another limitation is the use of a pre-existing framework in data analysis. This grounded theory study was conducted after the Q-matrix construction study and before the DSF analysis. During the Q-matrix construction part, major subskills (vocabulary, syntax, extracting explicit information, connecting and synthesizing) were already identified based on the literature and the think-aloud verbal reports. Therefore, the differences between the two groups were examined with the think-aloud protocols within these existing categories. However, it is possible that there were other differences between the two native language groups other than the subskill differences.

6.4.3 DSF Analysis

In this study, DSF analysis was conducted to investigate whether the Romance group and the East Asian group perform differently at the subskill level when their overall English reading

ability is controlled for. The following sections discuss some alternative methods for DSF analysis and also address some potential confounding variables that have not been controlled for.

Alternative DSF Procedures. In this study, logistic regression was used for the DSF analysis. The dependent variable, a dichotomous skill mastery status, was derived from the continuous posterior probability of mastery (PPM). An arbitrary cut-off criterion of 0.5 was used to reach a dichotomous mastery status for each examinee for each subskill, i.e., non-master if PPM < 0.5, master if PPM > 0.5.

However, it is important to note that alternative cut-off points exist. For instance, cut-off points of 0.4 and 0.6 could be used (Jang, 2005), i.e., non-master if PPM < 0.4, master if PPM > 0.6, and unclassified if 0.4 < PPM < 0.6. Those within the range of 0.4 to 0.6 are sometimes referred to as near masters or partial masters (Karelitz, 2008). If this alternative classification method is used, as shown in Figure 6.1, about 7.1% of examinees fall within the intermediate level for skill 1. This number was 7.3% for skill 2, 6.7% for skill 3, and 6% for skill 4. The result accords with Romàn (2009)'s observation that approximately 7% of the examinees could not be classified as either masters or non-masters in the diagnostic studies using the Fusion Model.

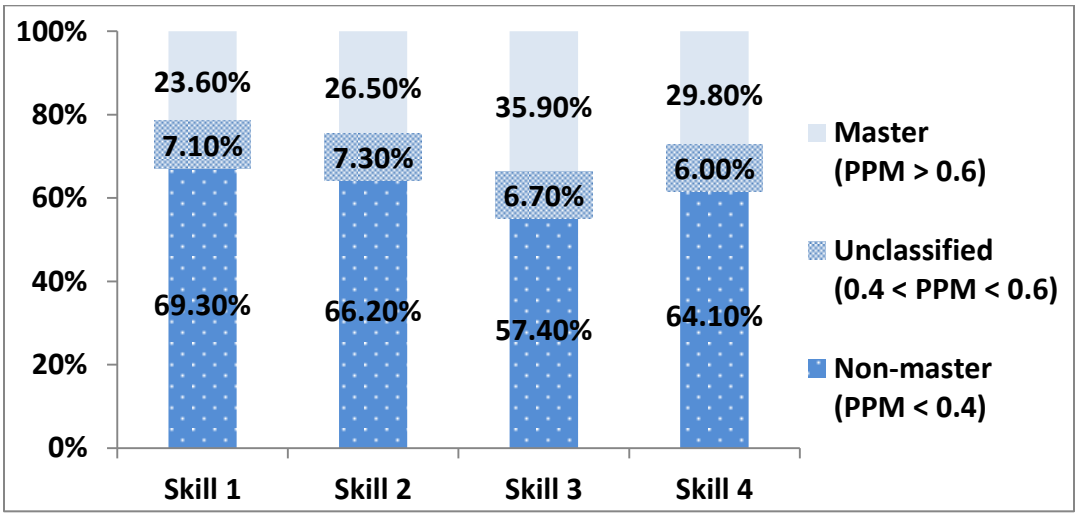

Figure 6.1. Alternative skill mastery classification.

In addition to the 0.4 and 0.6 cut-off points, we could also use percentiles for classification, such as upper 1/3, middle 1/3, and bottom 1/3. Also, the top 27%, the middle 46%, and the bottom 27% could make up three groups with different skill mastery status. If a three-category skill mastery status is used as the dependent variable, a polytomous logistic regression could be used for the DSF analysis. Or the middle category could be dropped, and a dichotomous logistic regression is still used for the DSF analysis.

Furthermore, in this study, four separate logistic regression analyses were conducted to examine the DSF for each subskill. A significance level of 0.5 was used to judge the statistical significance of the DSF (Scott et al., 2009). However, it should be noted that the four subskills are correlated with each other, because they are all the underlying components of reading comprehension. An alternative approach is to conduct a multivariate logistic regression analysis involving the four subskills within one single analysis as a way to control the overall error rate.

Finally, in this study only the statistical significance level was used to judge whether DSF existed between the two language groups. Further studies are needed to examine how the effect size measure of R^2 change could be incorporated into the decision making of the DSF within a cognitive diagnostic framework.

Potential Confounding Factors. Reading is a complex process, and any statistical modeling is only an approximation of the actual reading process. Using an existing dataset for the DSF analysis also restricted the possibility of controlling for some potential confounding variables.

One salient confounding factor is the testlet effect. In the MELAB reading test, the 20 multiple-choice items were associated with four reading passages, and thus the conditional independence assumption may be violated and the testlet effects may be present. Although

traditional IRT modeling can integrate testlet effects (Wainer, Bradlow, & Wang, 2007), as of 2010, cognitive diagnostic modeling has not caught up with accommodating testlet effects in its estimation approach (de la Torre, personal communication, May 1, 2010). It is possible that the estimation of the person and item parameters can be biased due to the presence of this testlet effect. A brief post-hoc analysis was conducted to study whether the presence of testlet effect could have confounded with the native language group. As shown in Figure 6.2, the average scores of the East Asian and Romance groups were graphed across each passage. The nearly parallel lines indicate that the overall performance of the two native language groups was influenced by the testlet effect in similar ways. This was also confirmed by a repeated measure ANOVA analysis. The F ratio associated with the interaction between language group and passage was 1.862, with a degree of freedom being 3 and a *p*-value of 0.134. Therefore, for the sake of the hypothesis testing between the two groups, the presence of testlet effect may not have had substantive influence.

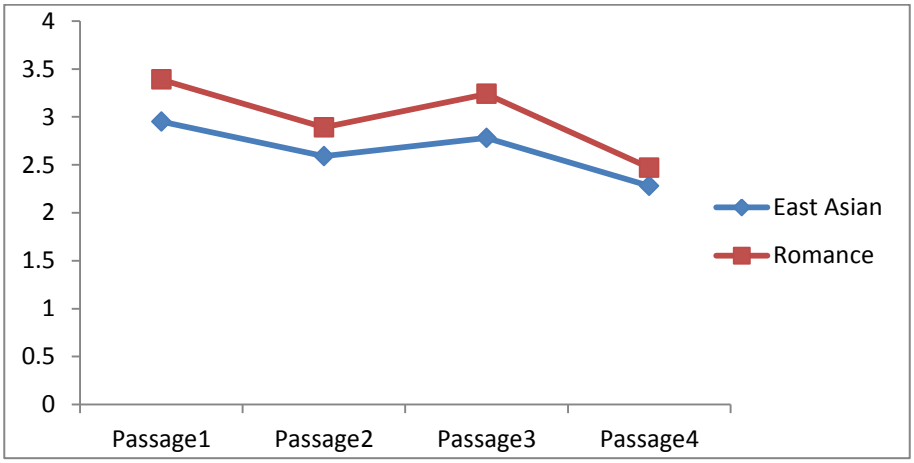

Figure 6.2. Average scores of East Asian and Romance groups across passages.

Even though testlet effects do not appear to be a problem for this study, still the accommodation of testlet effects in cognitive diagnostic modeling is an important area for future study. A potential solution is to treat each individual passage as a subskill, and thus code the item

as 1 for the passage it belongs to (Lei, personal communication, August 27, 2010). However, there is a statistical trade-off between the complexity of the Q-matrix and the accuracy of the parameter estimation for a test with a fixed length (Hartz, 2002). Given that the MELAB reading test only has 20 items, I did not attempt to add four more passage subkills to the Q-matrix. An important area for future research is to investigate how the passage subskills may influence the power of parameter estimation with the Fusion Model.

In addition to the testlet effect, there are other potential confounding variables. The overall purpose of this dissertation was to examine reading skill differences between two native language groups: East Asian versus Romance. However, although skills and strategies are conceptually distinct, they are not mutually exclusive. Skills refer to techniques that are automatic, whereas strategies are deliberate actions taken to achieve goals. The line between them is somewhat blurred. Therefore, although this dissertation primarily used the term *skills*, I am aware that skills and strategies are closely associated and even overlap in some circumstances.

Due to the complexity of the Q-matrix construction and concerns about the limited capacity of statistical modeling, only skills that are of substantial importance in correctly answering the items were coded in the Q-matrix. It was expected that the residual ability parameter in the Fusion Model might capture all those not specified in the Q-matrix, whether they are skills or strategies. However, it is not unlikely that those uncounted factors or variables may confound the native language group differences, such as metacognition, guessing, comparing options, eliminating options, and reading test items before reading the passage. At the same time, the examinees' other individual differences, such as gender, age, background knowledge, interest, motivation, and engagement, may have influenced the reading process as

well (Bernhardt, 2011). Among these, only gender was involved in the present DSF study, because many studies have shown salient gender differences in reading (Klinger et al., 2009) and because the MELAB dataset provided a complete record of examinee gender. Nevertheless, a better approach is to conduct an experimental study to examine the potential influence of the specific factors involved in the reading process.

REFERENCES

Abbott, M. (2006). ESL reading strategies: Differences in Arabic and Mandarin speaker test performance. *Language Learning, 56,* 633–670.

Adams, M.J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Adams, M.J. (1999). Afterword: The science and politics of beginning reading practices. In J. Oakhill & R. Beard (Eds.), *Reading development and the teaching of reading* (pp. 213–227). Oxford: Blackwell.

Akamatsu, N. (1999). The effects of first language orthographic features on word recognition processing in English as a second language. *Reading and Writing, 11* (4), 381–403.

Alderson, J.C. (1984). Reading in a foreign language: A reading problem or a language problem? In J.C. Alderson & A.H. Urquhart (Eds.), *Reading in a foreign language*. London: Longman.

Alderson, J.C. (1990a). Testing reading comprehension skills (Part One). *Reading in a Foreign Language, 6* (2), 425–438.

Alderson, J.C. (1990b). Testing reading comprehension skills (Part Two) Getting students to talk about talking a reading test (a pilot study). *Reading in a Foreign Language, 5* (2), 253–270.

Alderson, J.C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.

Alderson, J. C. (2010). "Cognitive diagnostic and Q-matrices in language assessment": A commentary. *Language Assessment Quarterly, 7,* 96–103.

Alderson, J.C., & Lukmani, Y. (1989). Cognition and reading: cognitive levels as embodied in test questions. *Reading in a Foreign Language, 5* (2), 1989.

Almond, P. J., Cameto, R., Johnstone, C. J., Laitusis, C., Lazarus, S., Nagle, K., Parker, C. E., Roach, A. T., & Sato, E. (2009). *White paper: Cognitive interview methods in reading test design and development for alternate assessments based on modified academic achievement standards (AA-MAS).* Dover, NH: Measured Progress and Menlo Park, CA: SRI International.

Alptekin, C. (2006). Cultural familiarity in inferential and literal comprehension in L2 reading. *System, 34*. 494–508.

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*, 41–66.

Angoff, W. H. (1972). *A technique for the investigation of cultural differences*. Paper presented at the annual meetings of the American Psychological Association, Honolulu.

Anthony, J. L., Lonigan, C. J., Driscoll, K., Phillips, B. M., & Burgess, S. R. (2003). Phonological sensitivity: A quasi-parallel progression of word structure units and cognitive operations. *Reading Research Quarterly, 38*, 470–487.

Baddeley, A.D. (2006) Working memory: an overview. In Pickering,S. (Ed.), *Working memory and education* (pp. 1–31). New York: Academic Press.

Baker, L. & Brown, A.L. (1984). Metacognitive skills and reading. In P. D. Pearson (Ed.), *Handbook of reading research*. New York: Longman.

Ball, E. W. (1997). Phonological awareness: Implications for whole language and emergent literacy programs. *Topics in Language Disorders, 17*, 14–26.

Bates, E., Devescovi, A., & D'Amico (1999).Processing complex sentences: A cross-linguistic study. *Language and Cognitive Processes, 14*(1), 69–123.

Bauerlein, M. & Stotsky, S. (2005, January). Why Johnny won't read. [Electronic Version] Washington Post. Retrieved from http://www.washingtonpost.com/wp-dyn/articles/A33956-2005Jan24.html

Berkowitz, S. (1986). Effects of instruction in text organization on sixth-grade students' memory for expository reading. *Reading Research Quarterly, 21*, 161–178.

Bernhardt, E. B. (2005). Process and procrastination in second language reading. *Annual Review of Applied Linguistics*, 25, 133–150.

Bernhardt, E. B. (2011). *Understanding advanced second-language reading*. New York: Routledge.

Bernhardt, E. B. & Kamil, M. L. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics, 16*(1), 15–34.

Birenbaum, M., Kelly, & Tatsuoka, K (1993). Diagnosing knowledge states in algebra using the rule-space model. *Journal for Research in Mathematics Education 24*, 442–59.

Biederman, I., & Tsao, Y.-C. (1979). On processing Chinese ideographs and English words: Some implications from Stroop-test results. *Cognitive Psychology, 11*, 125–132.

Bloom, B. S. (Ed.). (1956). *A taxonomy of educational objectives: Handbook I cognitive domain.* New York: Longmans, Green.

Bolt, D., Chen, H., DiBello, L., Hartz, S., Henson, R., Roussos, L., Stout, W., & Templin, J. (2008). The Arpeggio Suite: software for cognitive skills diagnostic assessment [Computer software and Manual]. St. Paul, MN: Assessment Systems.

Bolt, D., Li, Y. & Stout, W. (2003). *A low-dimensional IRT approach to linking calibrations based on the Fusion Model*. Unpublished Manuscript, University of Wisconsin-Madison.

Bossers, B. (1991). On thresholds, ceilings and short-circuits: The relation between L1 reading, L2 reading and L2 knowledge. In J. H. Hulstijn & J. F. Matter (Eds.), *Reading in two languages. AILA review, 8*, 45–60.

Brantmeier, C. (2003). Beyond linguistic knowledge: Individual differences in second language reading. *Foreign Language Annals, 36* (1), 33–43.

Breland, H., Lee, Y-W, Muraki, E. (2005). Comparability of TOEFL CBT essay prompts: response-mode analyses. *Educational and Psychological Measurement, 65* (4), 577– 599.

Breland, H., & Lee, Y-W (2007). Investigating uniform and non-uniform gender DIF in computer-based ESL writing assessment. *Applied Measurement in Education, 20* (4), 377–403.

Brown, T. L., & Haynes, M. (1985). Literacy background and reading development in a second language. In T. H. Carr (Ed.), *The development of reading skills* (pp. 19–34). San Francisco, CA: Jossey-Bass.

Bryant, A. & Charmaz, K., (Eds.). (2007), *The Sage handbook of grounded theory.* London: Sage Publications.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157.

Buck, G., VanEssen, T., Tatsuoka, K., Kostin, I., Lutz, D., & Phelps, M. (1998). *Development, selection and validation of a set of cognitive and linguistic attributes for the SAT I Verbal: Analogy section* (Research Report, RR-98-19). Princeton, NJ: Educational Testing Service.

Bügel, K., & Buunk, B.P. (1996). Sex differences in foreign language text comprehension: The role of interests and prior knowledge. *Modern Language Journal, 80*, 15–31.

Camilli, G. & Shepard, L.A. (1994). *Methods for identifying biased test items*. Hollywood, CA: Sage Publications.

Carrell, P.L. (1985). Facilitating ESL reading by teaching text structure. *TESOL Quarterly, 19* (4), 727–752.

Carrell, P. L. (1991). Second language reading: Reading ability or language proficiency. *Applied Linguistics, 12*, 159–79.

Carrrell, P.L. (1992). Awareness of text structure: Effects on recall. *Language Learning, 42* (1), 1–20.

Carver, R.P. (1992). What do standardized tests of reading comprehension measure in terms of efficiency, accuracy, and rate? *Reading Research Quarterly, 27*(4), 347–359.

Chang, C. (2002). *The reader effect (instruction/awareness of text structure) and text effect (well-structured vs. bad-structured texts) on first and second/foreign language reading comprehension and recall--what does research teach us?* Retrieved from http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED465180&ERICExtSearch_SearchType_0=no&accno=ED465180

Charmaz, K. (2006). *Constructing grounded theory. A practical guide through qualitative analysis*. London: Sage.

Chen, D. (1990). *Effects of formal schema and metacognition on reading comprehension of Chinese and English expository texts by Chinese readers* (Unpublished doctoral dissertation). University of Minnesota, Twin Cities, MN.

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

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25* (1), 15–37.

Cheng, T-Y. (1993). *The syntactical problems Chinese college students meet in reading English technical textbooks*. Indiana: ERIC Document Reproduction Service No. ED 364094.

Chiu, M. M., & McBride-Chang, C. (2006). Gender, context, and reading: a comparison of students in 41 countries. *Scientific Studies of Reading, 10*, 331–362.

Cho, J-R., & McBride-Chang, C. (2005a). Correlates of Korean hangul acquisition among kindergartners and second graders. *Scientific Studies of Reading, 9*, 3–16.

Cho, J-R., & McBride-Chang, C. (2005b).Levels of phonological awareness in Korean and English: A 1-year longitudinal study. *Journal of Educational Psychology, 97*,564–571.

Choi, I-C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing, 25* (1), 39–62.

Chu, H-C. (1999). *The effects of culture-specific rhetorical conventions on the L2 reading recall of Chinese students* (Unpublished doctoral dissertation). University of Texas-Austin, Austin, TX.

Clarke, M.A. (1980). The short circuit hypothesis of ESL reading—or when language competence interferes with reading performance. *Modern Language Journal, 64,* 203–209.

Clauser, B. E., Mazor, K. & Hambleton, R. K. (1993). The effect of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education, 6* (4), 269–279.

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test scores and an educational background variable. *Journal of Educational Measurement, 33*, 453–464.

Coady, J. (1979). A psycholinguistic model of the ESL reader. In R. Mackay, B. Barkman, & R. R. Jordan (Eds.), *Reading in a second language* (pp. 5–12). Rowley, MA: Newbury House.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Connell, D. & Gunzelmann, B. (2004). The next gender gap: Why are so many boys floundering while so many girls are soaring? *Instructor*, *113* (6), 14–17.

Corder, S. P. (1983). A role for the mother tongue. In S. Gass & L. Selinker (Eds.), *Language transfer in language learning* (pp. 85–97). Rowley, MA: Newbury House.

Cortazzi, M., & Jin, L. (1996). Cultures of learning: language classrooms in China. In Coleman, H. (Ed.), *Society and the language classroom* (pp. 169–206). Cambridge University Press, Cambridge.

Crane, P.K., Cetin, K., Cook, K.F., Johnson, K., Deyo, R., & Amtmann, D. (2007) Differential item functioning impact in a modified version of the Roland-Morris Disability Questionnaire, *Quality of Life Research, 16,* 981–990.

Creswell, J.W. (2002). *Educational research: Planning, conducting, and evaluating quantitative and qualitative approaches to research*. Upper Saddle River, NJ: Merrill/Pearson Education.

Creswell, J.W., Plano Clark, V.L., Gutmann, M.L., & Hanson, W.E. (2003). Advanced mixed methods research designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 209–240). Thousand Oaks, CA: Sage.

Crystal, D. (2004). *The stories of English*. London: Allan Lane.

Cummins, J. (1984). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera (Ed.), *Language proficiency and academic achievement* (pp. 2–19). Clevedon: Multilingual Matters.

Cutts, R. L. (1997). *An empire of schools: Japan's universities and the molding of a national power elite*. Armonk, N.Y., M.E. Sharpe.

Cziko, G.A. (1980). Language competence and reading strategies: A comparison of first and second-language oral reading errors. *Language Learning, 30,* 101–114.

Davidson, F. (2010). Why is cognitive diagnosis necessary? A reaction. *Language Assessment Quarterly, 7* (1), 104–107.

Davis, F. (1944). Fundamental factors of comprehension in reading. *Psychometrika, 9,* 185–197.

DiBello, L. (2002). Skill-based scoring models for PSAT/NMSAT. In Kristen L. Huff (Organizer), *Reporting more than scores: Skills-based scoring of a national test*. Symposium conducted at the meeting of the National Council of Measurement in Education, New Orleans, LA.

DiBello, L. & Crone, C. (2001). *Enhanced score reporting on a national standardized test*. Paper presented at the International Meeting of the Psychometric Society, Osaka, Japan.

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

DiBello, L. V., & Stout, W. F. (2007). Guest editor's introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44*(4), 285–291.

Dogan, E., Guerrero, A., Tatsuoka, K. (2005). *Using DIF to investigate strengths and weaknesses in mathematics achievement profiles of 10 different countries.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Montreal, Canada.

Dreyer, C., & Oxford, R. L. (1996). Learning strategies and other predictors of ESL proficiency among Afrikaans-speakers in South Africa. In Oxford, R. L. (Ed.), *Language learning strategies around the world: Cross-cultural perspectives* (pp. 61–74). Manoa: University of Hawaii Press.

Ehri, L., Nunes, S., Willows, D., Schuster, B., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly, 36,* 250–287.

Ehrlich, M-F. (1996). Metacognitive monitoring in the processing of anaphoric devices in skilled and less skilled comprehenders. In C. Cornoldi & Oakhill (Eds.), *Reading comprehension difficulties* (pp. 221–49). Mahwah, NJ: L. Erlbaum.

ELI-UM (2003). *The MELAB technical manual*. Retrieved from http://www.lsa.umich.edu/UMICH/eli/Home/Test%20Programs/MELAB/Officers%20&%20Professionals/Revised02TechManual.pdf

Ericsson, K. A., & Simon, H. A. (1984, 1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Esling, J., & Downing, J. (1986). What do ESL students need to learn about reading? *TESL Canada Journal, Special Issue, 1*, 55–68.

Fischer, G.H. (1973). The linear logistic model as an instrument in educational research. *Acta Psychologica, 37*, 359–374.

Fischer, G.H. (1983). Logistic latent trait models with linear constraints. *Psychometrika*, *48*, 3–26.

Fitzgerald, J. (1989). Research on stories: Implications for teachers. In K.D. Muth (Ed.), *Children's comprehension of text: Research into practice*. (pp. 2–36). Newark, DE: International Reading Association.

Flynn, S., & Espinal, I. (1985). The head-initial/head-final parameter in adult Chinese L2 acquisition of English. *Second Language Research, 1*, 93–117.

Fortson, B. W. (2004). *Indo-European Language and Culture: An Introduction*. Malden, Massachusetts: Blackwell.

Fries, C. (1945). *Teaching and learning English as a foreign language*. Michigan: University of Michigan Press, Ann Arbor.

Fu, J. (2005).*The polytomous extension of the fusion model and its Bayesian parameter estimation* (Unpublished doctoral dissertation). University of Wisconsin-Madison, Madison, WI.

Fu, J., & Li, Y. (2007). *An integrated review of cognitively diagnostic psychometric models*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Gallini, J.K., & Spires, H. (1995). Macro-based, micro-based, and combined strategies in text processing. *Reading Psychology, 16* (1), 21–41.

Gao, L. (2006). Toward a cognitive processing model of MELAB reading test item performance. *Spann Fellow Working Papers in Second or Foreign Language Assessment, 4*, 1–39. English Language Institute, University of Michigan, MI.

Gierl, M. J. (1997). Comparing the cognitive representations of test developers and students on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research, 91*, 26–32.

Gierl, M. J. Alves, C., Roberts, M., & Gotzmann, A. (2009). *Using judgments from content specialists to develop cognitive models for diagnostic assessments.* Paper presented at the 2009 annual meeting of the National Council on Measurement in Education, San Diego, CA.

Gierl, M. J., & Cui, Y. (2008) Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: interdisciplinary Research & Perspective, 6*(4), 263–268.

Gierl, M.J., Zheng, Y., & Cui, Y. (2008). Using the Attribute Hierarchy Method to identify and interpret cognitive skills that produce group differences. *Journal of Educational Measurement, 45* (1). 65–89.

Gillon, G. (2004). *Phonological awareness: from research to practice*. New York: Guilford Press.

Gitomer, D. H., & Yamamoto, K. (1991). Performance modeling that integrates latent trait and class theory. *Journal of Educational Measurement, 28*, 173–189.

Glaser, B.G. (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory*. Mill Valley, CA: Sociology Press.

Glaser, B. G. (1992). *Basics of grounded theory analysis*. Mill Valley, CA: Sociology Press.

Glaser, B. G. (1998). *Doing grounded theory: Issues and discussions*. Mill Valley, CA: Sociology Press.

Glaser, B. G. & Strauss, A. L (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine Publishing Company.

Goh, C. C. M. & Foong, K. P. (1997). Chinese ESL students' learning strategies a look at frequency, proficiency, and gender. *Hong Kong Journal of Applied Linguistics, 2*, 39–53.

Goodman, K.S. (1971). Psycholinguistic universals in the reading process. In P. Pimsleur & T. Quinn (Eds.), *The psycholinguistics of second language learning* (pp.135–142). Cambridge: Cambridge University Press.

Goodman, K.S. (1976). Reading: A psycholinguistic guessing game. In H. Singer & R.B. Ruddell (Eds.), *Theoretical models and processes of reading* (2nd ed., pp. 497–508). Newark, DE: International Reading Association.

Gough, P. B. and W. E. Tunmer (1986). Decoding, reading, and reading disability. *Remedial and special Education 7*(1), 6–10.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge, UK: Cambridge University Press.

Gray, W.S. (1919). *The relation between study and reading (Proceedings of the annual meeting of the National Education Association).* Washington, DC: National Education Association.

Gray, W. S. (1960). The major aspects of reading. In H. Robinson (Ed.), *Sequential development of reading abilities* (Vol. 90, pp. 8–24). Chicago: Chicago University Press.

Green, D.W., & Meara, P. (1987). The effects of script on visual search. *Second Language Research, 3*, 102–117.

Gurney, D., Gersten, R., Dimino, J.A., & Carnine, D. (1990). Story Grammar: Effective literature instruction for learning disabled high school students. *Journal of Learning Disabilities, 23*, 335–342.

Haberman, S. J. & von Davier, M. (2007). Some notes on models for cognitively based skills diagnosis. In C. V. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol.26, Psychometrics) (pp. 979–1027). Amsterdam: Elsevier.

Haertel, E. H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement, 8,* 333–346.

Haertel, E. H. (1989). Using restricted latent class models to map skill structure of achievement items. *Journal of Educational Measurement, 26,* 301–321.

Haertel, E. H. (1990). Continuous and discrete latent structure models of item response data. *Psychometrika, 55,* 477–494.

Hall, C., & Coles, M. (1999). *Children's reading choices*. London/New York: Routledge.

Halpern, D. (2000). *Sex differences in cognitive abilities* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Halpern, D. (2004). A cognitive-process taxonomy for sex differences in cognitive abilities. *Current Directions in Psychological Science,13*(4), 135–139.

Halpern. D. F. (2006). Assessing gender gaps in learning and academic achievement. In *Handbook of educational psychology* (2nd ed.), P.A. Alexander and P.H.Winne (Eds.), Lawrence Erlbaum Associates, Mahwah, NJ.

Hamada, M., & Koda, K. (2010). The role of phonological decoding in second language word-meaning inference. *Applied Linguistics, 31,* 513–531.

Hancin-Bhatt, B., & Nagy, W. (1994). Lexical transfer and second language morphological development. *Modern Language Journal, 75,* 27–38.

Harrington, M. (1987). Processing transfer: Language-specific processing strategies as a source of interlanguage variation. *Applied Psycholinguistics, 8,* 351–377.

Hartz, S.M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Henning, G., Hudson, T., &Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing,2*, 141–54.

Henson, R. A., Roussos, L., & Templin, J. L. (2004). *Cognitive diagnostic "fit" indices*. Unpublished ETS project report, Princeton, NJ.

Henson, R. A., Roussos, L., & Templin, J. L. (2005). *Fusion model "fit" indices*. Unpublished ETS project report, Princeton, NJ.

Henson, R. A., & Templin, J. L. (2004). *Modifications of the Arpeggio algorithm to permit analysis of NAEP*. Unpublished ETS project report, Princeton, NJ.

Henson, R. A., & Templin, J. L. (2007). *Large-scale language assessment using cognitive diagnosis models*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Henson, R., Templin, J., & Douglas, J. (2007). Use of subscores for estimation of skill masteries. *Journal of Educational Measurement, 44*, 361–376.

Henson, R., Templin, J., & Willse, J. (2008). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74* (2), 191–210.

Hitosugi, C. I., & Day, R. R. (2004). Extensive reading in Japanese. *Reading in a Foreign Language, 16*(1). Retrieved from http://nflrc.hawaii.edu/rfl/April2004/hitosugi/hitosugi.html.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel—Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *Canadian Modern Languages Review, 61*, 355–382.

Huang, M. (2009). *Brief analysis of Chinese students' errors in English reading*. Retrieved from www.nj0516.cn/Article/Class22/200908/3579.html

Idol, L. & Croll, V. (1987). Story-mapping training as a means of improving reading comprehension. *Learning Disability Quarterly, 10*, 214–229.

Inoue, A & Fodor, J.D. (1995). Information-paced parsing of Japanese. In R. Mazuka & N.Nagai (Eds.), *Japanese sentence processing* (pp. 9–64). Hillsdale: Lawrence Erlbaum Associates.

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

Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly, 6*(3), 210–238.

Jiao, H. (2009). Diagnostic classification models: Which one should I use? *Measurement: Interdisciplinary Research & Perspective, 7* (1), 65–67.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329–349.

Johnson, D.M., & Reynolds, F. (1941). A factor analysis of verbal ability. *Psychological Record, 4*, 183–195.

Johnston, R.S, Watson, J. E., & Logan, S (2009). Enhancing word reading, spelling and reading comprehension skills with synthetic phonics teaching: studies in Scotland and England. In C. Wood, & V. Connelly, (Eds.), *Contemporary perspectives on reading and spelling*. London: Routledge.

Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Available at http://education.umn.edu/NCEO/OnlinePubs/Tech44/

Jones, J. (1999). From silence to talk: cross-cultural ideas on students' participation in academic group discussion. *English for Specific Purposes, 18*(3), 243–259.

Juel, C. (1988). Learning to read and write: A longitudinal study of fifty-four children from first through fourth grades. *Journal of Educational Psychology, 80*, 437–447.

Juffs, A. (1998). Main verb versus reduced relative clauses ambiguity resolution in L2 sentence processing. *Language Learning, 48*, 107–147.

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Kaliski, P.K., Huff, K. L., & Thurber, A. (2011). *Using think-aloud interviews in evidence-centered assessment design for the Advanced Placement World History Exam.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Karelitz, T. M. (2008). How binary skills obscure the transition from non-mastery to mastery. *Measurement: Interdisciplinary research & Perspective, 6*(4), 268–272.

Kasai, M. (1997). *Application of the rule space model to the reading comprehension section of the test of English as a foreign language (TOEFL)* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18* (1), 89–114.

Klein, M., Birenbaum, M. Standiford, S., & Tatsuoka, K. K. (1981). *Logical error analysis and construction of tests to diagnose student 'bugs' in addition and subtraction of fractions* (Tech. Rep. 81–6–NIE), University of Illinois, CERL, Urbana-Champaign, IL.

Klinger, D. A., Shulha, A. A., & Wade-Woolley, L. (2009). *Towards an understanding of gender differences in literacy achievement: Literature review*. Retrieved from http://www.eqao.com/Research/pdf/E/ENGLISH_Literature_Review_May_11_2010.pdf

Koda, K. (1990). The use of L1 strategies in L2 reading: Effects of L1 orthographic structures on L2 phonological recoding strategies. *Studies in Second Language Acquisition, 12,* 393–410.

Koda, K. (1993). Transferred L1 strategies and L2 syntactic structure during L2 sentence comprehension. *Modern Language Journal, 77,* 490–500.

Koda, K. (2000a). Cross-linguistic interaction in the development of L2 intraword awareness: Effects of logographic processing experience. *Psychologia, 43,* 26–46.

Koda, K. (2000b). Cross-linguistic variations in L2 morphological awareness. *Applied Psycholinguistics, 21,* 297–320.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. NY: Cambridge University Press.

Krashen, S. (1981). *Second language acquisition and second language learning*. Oxford: Pergamon Press.

Kubiak, A., O'Neill, K., & Payton, C. (1992, April). *The effects of using educational background variables in DIF analysis*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Kweon, S.-O., & Kim, H.-R. (2008). Beyond raw frequency: Incidental vocabulary acquisition in extensive reading. *Reading in a Foreign Language, 20* (2), 191–215.

Larsen-Freeman, D., & Long, M. (1991). *An introduction to second language acquisition research*. New York, NY: Longman.

Lee, Y-W., Breland, H., Muraki, E. (2005). Comparability of TOEFL CBT writing prompts for different native language groups. *International Journal of Testing, 5 (2),* 131–158.

Lee, Y-W., & Sawaki, Y. (2009a). Application of three cognitive diagnosis models to ESL reading and listening. *Assessments: Language Assessment Quarterly, 6*(3), 239–263.

Lee, Y-W., & Sawaki, Y. (2009b). Cognitive diagnosis approaches to language assessment: An overview. *Assessments: Language Assessment Quarterly, 6*(3), 172–189.

Leighton, J.P. (2004). Avoiding misconceptions, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*, 1–10.

Leighton, J.P., & Gierl, M.J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26,* 3–16.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41,* 205–237.

Lems, K., Miller, L.D., Soro, T.M. (2010). *Teaching reading to English language learners*. New York: The Guilford Press.

Lennon, R.T. (1962). What can be measured? *The Reading Teacher, 15*, 326–337.

Levy, R., Mislevy, R.J., & Sinharay, S. (2006). *Posterior predictive model checking formultidimensionality in item response theory*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco, CA.

Levy, R., Mislevy, R.J., & Sinharay, S. (2007). *Posterior predictive model checking for conjunctive multidimensionality in item response theory*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

Li, P., Bates, E., & MacWhinney, B. (1993). Processing a language without inflections: A reaction time study of sentence interpretation in Chinese. *Journal of Memory and Language, 32*, 169–192.

Li, W. (1992). *What is a test testing? An investigation of the agreement between students' test-taking processes and test constructors' presumption* (Unpublished MA thesis). Lancaster University.

Liao, Y. (2007). Investigating the construct validity of the grammar and vocabulary section and the listening section of the ECCE: Lexico-grammatical ability as a predictor of L2 listening ability. *Spaan Fellow Working Papers in Second or Foreign Language Assessment,5*, 37–78. English Language Institute, University of Michigan, MI.

Liu, N. F., & Littlewood, W. (1997). Why do many students appear reluctant to participate in classroom learning discourse? *System, 25* (3), 371–384.

Logan, S., & Johnston (2009). Gender differences in reading ability and attitudes: examining where these differences lie. *Journal of Research in Reading, 32* (2), 199–214.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing, 10*(3), 211–234.

Lunzer, E., Waite, M. & Dolan, T. (1979). Comprehension and comprehension tests. In E., Lunzer, & K., Gardner (Eds.), *The effective use of reading (37–71)*. London: Heinemann Educational Books.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology, 9*, 111–151.

Matthews, M. (1990). Skill taxonomies and problems for the testing of reading. *Reading in a Foreign Language, 7* (1), 511–517.

Mazuka, R., & Itoh, K. (1995). Can Japanese speakers be led down the garden path? In R. Mazuka & N. Nagai. (Eds.), *Japanese sentence processing* (pp. 295–329). Hillsdale, NJ: Erlbaum.

McKenna, M.C., Kear, D.J. & Ellsworth, R.A. (1995). Children's attitudes toward reading: A national survey. *Reading Research Quarterly, 30*(4), 934–956.

Meyer, B. J. F. (1975). *The organization of prose and its effects on memory*. Amsterdam: North Holland.

Meyer, B. J. F. (1985). Prose analysis: Purposes, procedures, and problems. In B. K. Britton & J. Black (Eds.), *Analyzing and understanding expository text* (pp. 11–64, 269–304). Hillsdale, NJ: Erlbaum.

Meyer, B.J.F., Brandt, D.M., & Bluth, G.J. (1980). Use of the top level structure in text: Key for reading comprehension of ninth grade students. *Reading Research Quarterly*, *16*(1), 72–103.

Meyer, B. J. F., & Poon, L. W. (2001). Effects of structure strategy training and signaling on recall of text. *Journal of Educational Psychology, 93*, 141–159.

Meyer, B. J. F., & Rice, G. E. (1982). The interaction of reader strategies and the organization of text. *Text, Interdisciplinary Journal for the Study of Discourse, 2*, 155–192.

Meyer, B. J. F., & Wijekumar, K. (2007). A web-based tutoring system for the structure strategy: Theoretical background, design, and findings. In D.S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 347–375). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Meyer, B.F., Wijekumar, K., Middlemiss, W., Higley, K., Lei, P., Meier, C., & Spielvogel, J. (2010). Web-based tutoring of the structure strategy with or without elaborated feedback or choice for fifth-and seventh-grade readers. *Reading Research Quarterly*, *45* (1), 62–92.

Meyer, B. J. F., Young, C. J., & Bartlett, B. J. (1989). *Memory improved: Reading and memory enhancement across the life span through strategic text structures*. Hillsdale, NJ: Lawrence Erlbaum.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105–107.

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

Milewski, G.B., & Baron, P.A. (2002). *Extending DIF methods to inform aggregate reports on cognitive skills*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439–483.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, Delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics. 32* (1), 92 –110.

Montero, D. H., Monfils, L., Wang, J., Yen, W. M., & Julian, M. W. (2003). *Investigation of the application of cognitive diagnostic testing to an end-of-course high school examination.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Mullis, I., Campbell, J. & Farstrup, A. (1993). *NAEP 1992: Reading report card for the nation and the states*. Washington, DC: U.S. Department of Education.

Mullis, I., Martin, M.O., Kennedy, A.M., & Foy, P. (2007). *PIRIL 2006 international report: IEA's progress in international reading literacy study in primary schools in 40 countries*. Chestnut Hill, MA: Boston College.

Munby, J. (1978). *Communicative syllabus design*. Cambridge, UK: Cambridge University Press.

Muthén, B. O., & Muthén, L. K. (2010). *Mplus* 6 [Computer software]. CA: Los Angeles.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: John Wiley & Sons.

Oxford, R. L. (1993). Instructional implications of gender differences in second/foreign language (L2) learning styles and strategies. *Applied Language Learning, 4* (1–2), 65–94.

Pae, T. (2004). DIF for examinees with different academic backgrounds. *Language Testing, 21* (1), 53–73.

Palmer, H. E. (1917). *The scientific study and teaching of languages*. London: Harrap. (Reissued in 1968 by Oxford University Press).

Paris, S.G., Lipson, M.Y., & Wixson, K.K. (1983). Becoming a strategic reader, *Contemporary Educational Psychology, 8*, 293–316.

Patz, R. J., & Junker, B.W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*(4), 342–366.

Pearson, P.D., & Fielding, L. (1991). Comprehension instruction. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol.2, pp. 815–860). White Plains, NY: Longman.

Perfetti, C. A. (1999). Comprehending written language: A blueprint of the reader. In C. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 167–208). Oxford University Press.

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading,11*(4), 357–383.

Perfetti, C. A., & Dunlap, S. (2008). Learning to read: General principles and writing system variations. In K. Koda & A. Zehler (Eds.), *Learning to read across languages* (pp. 13–38). Mahwah, NJ: Erlbaum.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Oxford: Blackwell.

Perfetti, C. A., Marron, M. A., & Foltz, P. W. (1996). Sources of comprehension failure: Theoretical perspectives and case studies. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention*. Mahwah, NJ: Lawrence Erlbaum.

Perfetti, C. A., & Zhang, S. (1995). Very early phonological activation in Chinese reading, *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(1), 24–33.

Pettit, N.T. & Cockriel, I.W. (1974). A factor study on the literal reading comprehension test and the inferential reading comprehension test. *Journal of Reading Behavior, 6*, 63–75.

Poole, G. (2003). Assessing Japan's institutional entrance requirements. *Asian EFL Journal*, 5 (1). Retrieved from http://www.asian-efl-journal.com/march03.sub5a.php.

Powell, S. (2005). Extensive reading and its role in Japanese high school. *The Reading Matrix, 5* (2). Retrieved from http://www.readingmatrix.com/articles/powell/article.pdf.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Erlbaum.

Pugh, K. R., Shaywitz, B. A., Shaywitz, S. E., Constable, R. T., Skudlarski, P., Fulbright, R. K., et al. (1996). Cerebral organization of component processes in reading. *Brain, 119*, 1221–1238.

Qi, C. H., & Marley, S. C. (2009). Differential item functioning analysis of the preschool language scale-4 between English-speaking Hispanic and European American children from low-income families. *Topics in Early Childhood Special Education, 29*(3), 171–180.

Ramstedt, G. J. (1928) *A Korean Grammar*. Memoires de la societe Finno-Ougrienne LXXXII. Helsinki.

Rao, Z. (2001). Matching teaching styles with learning styles in East Asian contexts. *The Internet TESL Journal*, VII (7). Retrieved from http://iteslj.org/Techniques/Zhenhui-TeachingStyles.html

Raugh, H.E. (2008). The origins of the transformation of the defense language program. *Applied Language Learning, 16*(2), 1–12. Retrieved from http://www.dliflc.edu/academics/academic_materials/all/ALLissues/all16two.pdf

Ringbom, H. (1987). *The role of the first language in foreign language learning*. Clevedon: Multilingual Matters.

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel–Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*, 105–116.

Romàn, A. I. S. (2009). *Fitting cognitive diagnostic assessment to the cognitive assessment tool for statistics* (Unpublished doctoral dissertation). Purdue University, Lafayette, OH.

Ross, S. J. (2008). Language testing in Asia: Evolution, innovation, and policy challenges. *Language Testing, 25* (1), 5–13.

Rost, D. (1993). Assessing the different components of reading comprehension: Fact or fiction? *Language Testing, 10*(1), 79–82.

Roussos, L. A., DiBello, L.V., Stout, W. F., Hartz, S. M., Henson, R. A., & Templin, J. H. (2007). The fusion model skills diagnostic system. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. New York: Cambridge University Press.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44* (4), 293–311.

Roussos, L., Xu, X., & Stout, W. (2003). *Skills diagnosis data simulation program, version 1.1*. Unpublished ETS project report: Princeton, NJ.

Rupp, A. A. & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*(4), 219–262.

Rupp, A. A., Templin, J., & Henson, R. J. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

Ryan, K., & Bachman, L.F. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9*(1), 12–29.

Sainsbury, M. & Schagen, I. (2004). Attitudes to reading at ages nine and eleven. *Journal of Research in Reading, 27*(4), 373–386.

Sasaki, M. (2008). The 150-year history of English language assessment in Japanese education. *Language Testing, 25(1)*, 63–83.

Savignon, S., & Berns, M. S. (Eds.). (1984). *Initiatives in communicative language teaching*. Reading, MA: Addison-Wesley.

Sawaki, Y., Kim, H. J., & Gentile, C. (2009). Q-Matrix construction: defining the kink between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly, 6*(3), 190–209.

Sawaki, Y., Stricker, L., & Oranje, A. (2009). Factor structure of the TOEFL internet-based test. *Language Testing, 26*, 5–30.

Schrader, S. v. (2006). *On the feasibility of applying skills assessment models to achievement test data* (Unpublished doctoral dissertation). University of Iowa, Iowa city, IA.

Scott, H. S. (1998). *Cognitive diagnostic perspectives of a second language reading test* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Petersen, M.A., Sprangers, M. A.G. (2009). A simulation study provided sample size guidance for differential item (DIF) studies using short scales. *Journal of Clinical Epidemiology, 62*, 288–295.

Seidenberg, P. L. (1989). Relating text-processing research to reading and writing instruction for learning disabled students. *Learning Disabilities Focus, 5* (1), 4–12.

Shaywitz, B. A., Shaywitz, S. E., Pugh, K. R., Constable, R. T., Skudlarski, P., Fulbright, R. K., et al. (1995). Sex differences in the functional organization of the brain for language [see comment]. *Nature, 373*, 607–609.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159–194.

Sheorey, R., & Mokhtari, K. (2001).Differences in the metacognitive awareness of reading strategies among native and non-native readers. *System, 29*, 431–449.

Shibatani, M. (1990). *The languages of Japan*. Cambridge University Press, Cambridge.

Singer, H., & Donlan, D. (1982). Active comprehension: Problem-solving schema with question generation for comprehension of short stories. *Reading Research Quarterly, 17,* 166–185.

Sinharay, S. (2004). Experiences with Markov Chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics, 29*, 461–488.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*, 375–394.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics, 31*, 1–33.

Sinharay, S., Johnson, M.S., & Stern, H.S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30,* 298–321.

Slater, W. H. (1985). Teaching expository text structure with structural organizer. *Journal of Reading, 28* (8), 712–718.

Spearritt, D. (1972). Identification of subskills of reading comprehension by maximum likelihood factor analysis. *Reading Research Quarterly, 8,* 92–111.

Stahl, S. A., & Murray, B. A. (1994). Defining phonological awareness and its relationship to early reading. *Journal of Educational Psychology*, *86*(2), 221–234.

Stanovich, K.E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16*(1), 92–111.

Stanovich, K. E. (1986). Matthew effects in reading: some consequences of individual differences in acquisition of literacy. *Reading Research Quarterly, 21* (4), 360–407.

Steinberg, L., Thissesn, D., & Wainer, H. (1990). Validity. In H. Wainer, N. J. ns, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (pp. 187–231). Hillsdale, NJ: Lawrence Erlbaum.

Stevenson, M., Schoonen, R. & de Glopper, K. (2006). Revising in two languages: A multidimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing, 15* (3), 201–233.

Stoker, H. W., & Kropp, R. P. (1960). The perspective validates and factorial content of the Florida state-wide ninth grade testing program. *Florida Journal of Educational Research, 2*, 105–114.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, Calif: Sage Publications.

Strauss, A.L., Corbin, J.M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage.

Suen, H.K., & Yu, L. (2006). Chronic consequences of high-stakes testing? Lessons from the Chinese civil service exam. *Comparative Education Review, 58* (1), 46–65.

Swaminathan, H. (1994). Differential item functioning: A discussion. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and issues*. Ottawa, Canada: University of Ottawa.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–370.

Swan, M. (1985a). A critical look at the communicative approach (1). *ELT Journal, 39*(1), 2–12.

Swan, M. (1985b). A critical look at the communicative approach (2). *ELT Journal*, *39*(2), 76–87.

Tatsuoka, K.K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345–354.

Tatsuoka, K. K., Corter, J., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 11*(4), 901–926.

Tatsuoka, K.K., & Tatsuoka, M.M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7,* 215–231.

Taylor, B. M. (1992). Chapter 9: Text structure, comprehension, and recall. In S. J. Samuels, & A. E. Farstrup (Eds.), *What research has to say about reading instruction.* (2nd ed.). (pp. 220–235). Newark, DE: International Reading Association.

Taylor, I. (1998). Learning to read in Chinese, Korean and Japanese. In A.Y. Durgunoglu & L. Verhoeven (Eds), *Literacy development in a multilingual context: Cross-cultural perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Taylor, B. M., & Beach, R. W. (1984). The effects of text structure instruction on middle-grade students' comprehension and production of expository text. *Reading Research Quarterly, 19* (2), 134–146.

Taylor, I., & Taylor, M.M. (1995). *Writing and literacy in Chinese, Korean, and Japanese*. Philadelphia: John Benjamins.

Templin, J. L. (2005). *Generalized linear mixed proficiency models for cognitive diagnosis* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11,* 287–305.

Templin, J. L., Henson, R. A., & Douglas, J. (2006). *General theory and estimation of cognitive diagnosis models: Using Mplus to derive model estimates*. Unpublished manuscript.

Thompson, I. (1987). Memory in language learning. In A. Wenden & J. Rubin (Eds.), *Learner strategies in language learning* (pp. 43–56). Englewood Cliffs, NJ: Prentice Hall International Ltd.

Thorndike, E. L. (1917a). The psychology of thinking in the case of reading, *Psychological Review, 24,* 220–234.

Thorndike, E. L. (1917b). Reading as reasoning: a study of mistakes in paragraph reading. *Journal of Educational Psychology, 8*, 323–332.

Thorndike, E. L. (1917c). The understanding of sentences: a study of errors in reading. *Elementary School Journal, 8*, 98–114.

Thurgood, G., & LaPolla, R. J. (Eds.). (2003). *Sino-Tibetan languages*. London: Routledge.

Thurstone, L.L. (1946). Note on a reanalysis of Davis' reading tests. *Psychometrika, 11*, 185–188.

Tzeng, O.J., & Wang, W. S.-Y. (1983). The first two R's: The way different languages reduce speech to script affects how visual information is processed in the brain. *American Scientist, 71*, 238–243.

Vacca, R.T. (1980). A study of holistic and subskill instructional approaches to reading comprehension. *Journal of Reading, 23*, 512–518.

Vernon, P.E. (1962). The determinants of reading comprehension. *Educational and Psychological Measurement, 22*, 269–286.

von Davier, M. (2006). Multidimensional latent trait modeling (MDLTM) [Software program].Princeton, NJ: Educational Testing Service.

Vygotsky, L.S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wainer, H., Bradlow, E.T., & Wang, X. (2007) *Testlet response theory and its applications*. New York: Cambridge University Press.

Walczyk, J. (1995). Testing a compensatory-encoding model. *Reading Research Quarterly, 30*, 396–408.

Walczyk, J. (2000). The interplay between automatic and control processes in reading. *Reading Research Quarterly, 35*, 554–566.

Walker, C. M., Azen, R., & Schmitt, T. (2006). Statistical versus substantive dimensionality: The effect of distributional differences on dimensionality assessment using DIMTEST. *Educational and Psychological Measurement*, 66, 721–738.

Wang, C., & Gierl, M. J. (2007). *Investigating the cognitive attributes underlying student performance on the SAT® critical reading subtest: An application of the Attribute Hierarchy Method.* Paper presented at the 2007 annual meeting of the National Council on Measurement in Education.

Wang, M., Koda, K., & Perfetti, C.A. (2003). Alphabetic and non-alphabetic L1 effects in English semantic processing: A comparison of Korean and Chinese English L2 learners. *Cognition, 87*, 129–149.

Weir, C.J. & Porter, D. (1990). Reading skills: hierarchies, implicational relationships and identifiability. *Reading in Foreign Language, 7*(1), 505–510.

Weir, C.J. & Porter, D. (1994). The multi-divisible or unitary nature of reading: the language tester between Scylla and Charybdis. *Reading in a Foreign Language, 10* (2), 1–19.

Whitmore, M. L., & Schumacker, R. E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Educational and Psychological Measurement, 59*, 9, 10—927.

Williams, J. P., Hall, K. M., Lauer, K. D., Stafford, K. B., DeSisto, L. A., & deCani, J. S. (2005). Expository text comprehension in the primary grade classroom. *Journal of Educational Psychology, 97*, 538–550.

Wu, A. D., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of Differential Item Functioning. *International Journal of Testing, 6*(3), 287–300.

Xu, X., & von Davier, M. (2006). *A General Diagnostic Model applied to language testing data* (Research Report, RP-05-16), Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008). *Fitting the structured General Diagnostic Model to NAEP data*. (Research Report, RR-08-27), Princeton, NJ: Educational Testing Service.

Yan, D., Almond, R. G., & Mislevy, R. J. (2004). *Comparisons of cognitive diagnostic models* (Research Report, RR-04-02), Princeton, NJ: Educational Testing Service.

Yoneyama, S. (1999). *The Japanese high school: silence and resistance.* London, Routledge.

Younger, M. & Warrington, M. (2005). *Raising boys' achievement* (Research Report No.636). Cambridge, England: University of Cambridge Faculty of Education.

Zappe, S. (2007). *Response process validation of equivalent test forms: How qualitative data can support the construct validity of multiple test forms* (Unpublished doctoral dissertation). The Pennsylvania State University, State College, PA.

Zenisky, A.L., Hambleton, R.K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement, 63 (1)*, 51–64.

Zhang, W. (2006). *Detecting differential item functioning using the DINA model* (Unpublished doctoral dissertation). University of North Carolina at Greensboro, Greensboro, NC.

Zhong, L. (2006). *Culture root and academic writing: Factors that influence Chinese international students' academic writing at universities in North America.* Paper presented at Internationalizing Canada's universities: Practices, challenges, and opportunities. Retrieved from http://www.york.ca/yorkint/global/conference/canada/papers/Zhong-Lan.pdf

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.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26*, 55–66.

Appendix A Consent Form for the Think-Aloud Activity

Informed Consent Form for Social Science Research
The Pennsylvania State University

Title of Project: Cognitive diagnostic analysis of the MELAB reading test

Principal Investigator: Hongli Li, Graduate student
226 CEDAR Bldg
University Park, PA 16802
(814) 321-1584; HUL151@psu.edu

Advisor: Dr. Hoi K. Suen
103 CEDAR Bldg
University Park, PA. 16802
(814)-865-2235, HoiSuen@psu.edu

1. **Purpose of the Study:** The purpose of this research study is to explore the use of cognitive diagnostic analysis with Michigan English Language Assessment Battery (MELAB) reading test and also to evaluate whether the latent subskills underlying MELAB-readings are compensatory or conjunctive.

2. **Procedures to be followed:** You will be asked to participate in a think-aloud activity. Before this activity, you will fill in a personal background information sheet. Then you will be asked to read 2 short reading passages, each followed by 5 multiple-choice items. You will be asked to think-out-aloud what you are thinking in your head when reading the passages and answering the items during the task. You will also be asked to recall retrospectively what you thought after the task. You will be audiorecorded during this think-aloud process.

3. **Discomforts and Risks:** There are no risks in this research beyond those in everyday life.

4. **Benefits:** As a result of participating in the study, you may gain insights into your reading skills and processes. You might have a better understanding of how to monitor your reading processes when you try to respond to a reading passage.

5. **Duration:** It will take about 45-60 minutes to finish the activity.

6. **Statement of Confidentiality:** Your participation in this research is confidential. The recordings will be stored and secured in a locked cabin at 329B of IST building. Only the principal investigator will have access to them. The recordings will be destroyed within 3 years following the making of the recordings. The Pennsylvania State University's Office for Research Protections, the Institutional Review Board and the Office for Human Research Protections in the Department of Health and Human Services may review records related to this research study. In the event of a publication or presentation resulting from the research, no personally identifiable information will be shared.

7. **Right to Ask Questions:** Please contact Hongli Li at (814) 321-1584 with questions, complaints or concerns about this research. You can also call this number if you feel this study has harmed you. If you have any questions, concerns, problems about your rights as a research participant or would like to offer input, please contact The Pennsylvania State University's Office for Research Protections (ORP) at (814) 865-1775. The ORP cannot answer questions about research procedures. All questions about research procedures can only be answered by the research team.

8. **Payment for participation:** You will receive $20 in cash for your participation.

9. **Voluntary Participation:** Your decision to be in this research is voluntary. You are free to stop participating in the research at any time. Refusal to take part in or withdrawing from this study will involve no penalty or loss of benefits you would receive otherwise.

You must be 18 years of age or older to take part in this research study. If you agree to take part in this research study and the information outlined above, please sign your name and indicate the date below.

You will be given a copy of this form for your records.

_____ _____

Participant Signature Date

_____ _____

Person Obtaining Consent Date

Appendix B Email Invitation for the Think-Aloud Activity

Subject: Participants Wanted for an Education Study at Penn State

Dear_____,

I am Hongli Li, a Ph.D. candidate of Educational Psychology program at the Penn State University. I am currently conducting a cognitive diagnostic analysis of the Michigan English Language Assessment Battery (MELAB) reading test for my research.

I would like to invite you to participate in this study. The study will take approximately 45-60 minutes. If you agree to participate, you will receive $ 20 in cash after you complete this study. Thank you in advance and I appreciate your help.

What: You will be asked to participate in a think-aloud activity. In this activity, you will be asked to read 2 short reading passages from Michigan English Language Assessment Battery (MELAB), each followed by 5 multiple-choice items. You will be asked to think-out-aloud what you are thinking when reading the passage and answering the items during the task. You will also be asked to recall retrospectively what you thought after the task. This think-aloud process will be audio-recorded, and your background information will also be collected. Your personal information will be kept confidential, and your name will not appear in any written documents nor will be disclosed to any third party.

The purpose of this study is NOT to test your English ability, but to help me understand the reading skills that are generally needed to answer the MELAB reading test.

Who: You must be over 18 years old; English is your second/foreign language; You have an intermediate to advanced English proficiency.

When & Where: The session will be arranged at the study rooms of Pattee Library.

How: Please reply to this email or call me at (814)-321-1584 if you are interested. I will further contact you for confirmation and the arrangement of the session.

Your participation is very important and I appreciate your help.
Hongli

Hongli Li
Phd candidate
Educational Psychology
Penn State University
Email: HUL151@psu.edu
Tel: (814)-321-1584

Appendix C Verbal Script for the Think-Aloud Activity

I am Hongli Li, a Ph.D. candidate of Educational Psychology program at the Penn State University. I am currently conducting a cognitive diagnostic analysis of the Michigan English Language Assessment Battery (MELAB) reading test for my research.

In this study, I am interested in the cognitive processes you use to answer the reading items. You will be asked to verbally report your thinking processes while reading the passage and answering the items and your remembrances about your thoughts after completing the task. The whole session will take approximately 45-60 minutes. You will be completing a total of 10 items based on two passages during this session. Because I will be asking you to talk quite a bit during the session, I will be using a digital audio recorder to make sure that I capture everything that you tell me. This is completely voluntary and I want to be sure you are comfortable with being part of this study. Before we start the think-aloud, you will be asked to fill in a background information sheet. You do not need to respond to any question in the sheet if you do not feel comfortable doing so. Your personal information will be kept confidential, and your name will not appear in any written documents nor will be disclosed to any third party.

Do you agree to participate?
[If yes], Would you please fill out the consent form? Please feel free to ask me if you have any questions or need any explanations about the study and the consent form.
[The researcher provides the participant with the consent from to read and sign].

Great! Now would you please fill out the background information sheet? You may leave it blank if you do not feel comfortable providing response to any of the following question.
[The researcher provides the participant with the background information sheet to read and fill].

Great! Now let’s begin with the think-aloud activity!

Appendix D Think-Aloud Participant Background Information Sheet

This is a brief form about your background information. Please be assured that your personal information will be used only for research purposes and will remain *strictly confidential*. You may leave it blank if you do not feel comfortable providing response to any of the following question.

Email address: _____ Phone number:_____

Gender:_____

Country of origin: _____

First (native) language: _____

Department or program where you are studying_____

Degree which you are pursuing_____

How many years have you learned English? _____

How would you rate your reading ability in English?
(1: minimal, 2: basic, 3: good, 4: very good, 5: excellent) *circle the appropriate number*

How long have you been in the U.S. (or any other English-speaking country?)

Have you taken the TOEFL? (circle one: PBT/ CBT/ iBT)
What was your total score? _____ (Approximate date of exam: _____/_____/_____)
Reading _____ Listening _____ Writing _____ Speaking (if IBT) _____ Grammar (if CBT) _____

If you have taken IELTS instead,
What was your score? _____ (Approximate date of exam: _____/_____/_____)
Reading _____ Speaking _____ Listening _____ Writing _____

Appendix E Think-Aloud Training Materials

In bringing up children, every parent watches eagerly the child's acquisition of each new skill: the first spoken words, the first independent steps, or the beginning of literacy. It may be tempting to hurry the child beyond his natural learning rate, but this can set up dangerous feelings of failure and states of anxiety in the child.

Adapted from College English Test Band 4, July 1998

Which of the following word has the closest meaning to “acquisition”?

a) Improve

b) Forget

c) Remember

d) Learn

What a child may feel if his parents hurry him to learn more?

a) Excited

b) Worried

c) Intensified

d) Satisfied

Appendix F Consent Form for Expert Rating

Informed Consent Form for Social Science Resear
The Pennsylvania State University

Title of Project: Cognitive diagnostic analysis of the MELAB reading test

Principal Investigator: Hongli Li, Graduate student
226 CEDAR Bldg
University Park, PA 16802
(814) 321-1584; HUL151@psu.edu

Advisor: Dr. Hoi K. Suen
103 CEDAR Bldg
University Park, PA. 16802
(814)-865-2235, HoiSuen@psu.edu

10. **Purpose of the Study:** The purpose of this research study is to explore the use of cognitive diagnostic analysis with Michigan English Language Assessment Battery (MELAB) reading test and also to evaluate whether the latent subskills underlying MELAB-readings are compensatory or conjunctive.

11. **Procedures to be followed:** You will be asked to participate in an expert rating activity. Before this activity, you will fill in a background information sheet, and then you will receive a brief training of the rating task. For the rating task, you will be provided with 4 short reading passages, each followed by 5 multiple-choice items. You will be asked to identify what reading skills are required to answer each item correctly. Finally, you will discuss your rating with other experts in a group.

12. **Discomforts and Risks:** There are no risks in this research beyond those in everyday life.

13. **Benefits:** As a result of participating in the study, you may gain insights into reading skills and processes as required by the MELAB test. You might also have a better understanding of how to help English as Second Language (ESL) students monitor their reading processes and improve their reading ability.

14. **Duration:** It will take about 2 hours to finish the activity.

15. **Statement of Confidentiality:** Your participation in this research is confidential. The Pennsylvania State University's Office for Research Protections, the Institutional Review Board and the Office for Human Research Protections in the Department of Health and Human Services may review records related to this research study. In the event of a publication or presentation resulting from the research, no personally identifiable information will be shared.

16. **Right to Ask Questions:** Please contact Hongli Li at (814) 321-1584 with questions, complaints or concerns about this research. You can also call this number if you feel this study has harmed you. If you have any questions, concerns, problems about your rights as a research participant or would like to offer input, please contact The Pennsylvania State University's Office for Research Protections (ORP) at (814) 865-1775. The ORP cannot answer questions about research procedures. All questions about research procedures can only be answered by the research team.

17. **Payment for participation:** You will receive $20 in cash for your participation.

18. **Voluntary Participation:** Your decision to be in this research is voluntary. You are free to stop participating in the research at any time. Refusal to take part in or withdrawing from this study will involve no penalty or loss of benefits you would receive otherwise.

You must be 18 years of age or older to take part in this research study. If you agree to take part in this research study and the information outlined above, please sign your name and indicate the date below.

You will be given a copy of this form for your records.

_____ _____

Participant Signature Date

_____ _____

Person Obtaining Consent Date

Appendix G Reading Expert Background Information Sheet

This is a brief form about your background information. Please be assured that your personal information will be used only for research purposes and will remain *strictly confidential*. You may leave it blank if you do not feel comfortable providing response to any of the following question.

Demographic information

Country of origin: _____

First (native) language: _____

Other languages you speak _____

How long have you been in the U.S. (or any other English-speaking country?)_____

Your ESL teaching experience

1) How long have you taught English to ESL/EFL students?

2) What courses have you taught? in what programs? to what kinds of students?

Appendix H Sample Expert Rating Form

	Linguistic Skills		Comprehension Skills			Residual Skills
Item	**Vocabulary**	**Syntax**	**Extracting explicit information**	**Connecting and synthesizing**	**Making inferences**	
1						
2						

Appendix I Item Statistics for the MELAB Reading Dataset

	Mean (or proportion-correct score)	Standard deviation	N	Corrected item-total correlation	Cronbach's alpha if item deleted
Item1	.55	.498	2019	.308	.805
Item2	.65	.478	2019	.415	.799
Item3	.72	.447	2019	.356	.802
Item4	.76	.427	2019	.404	.799
Item5	.36	.481	2019	.354	.802
Item6	.47	.499	2019	.337	.803
Item7	.74	.441	2019	.383	.800
Item8	.44	.496	2019	.418	.798
Item9	.81	.395	2019	.335	.803
Item10	.27	.442	2019	.348	.802
Item11	.62	.486	2019	.344	.802
Item12	.60	.490	2019	.366	.801
Item13	.75	.435	2019	.412	.799
Item14	.54	.498	2019	.471	.795
Item15	.44	.496	2019	.241	.808
Item16	.45	.497	2019	.453	.796
Item17	.32	.466	2019	.251	.807
Item18	.51	.500	2019	.417	.798
Item19	.45	.498	2019	.447	.797
Item20	.56	.497	2019	.436	.797

Appendix J Descriptive Statistics of Examinee Performance

Table J.1

Number of Masters of Skill 1 (Gender by Language Group)

		Native Languages		
		East Asian	Romance	Total
Gender	Male	55	21	76
	Female	50	32	82
Total		105	53	158

Table J.2

Number of Masters of Skill 2 (Gender by Language Group)

		Native Languages		
		East Asian	Romance	Total
Gender	Male	60	17	77
	Female	83	32	115
Total		143	49	192

Table J.3

Number of Masters of Skill 3 (Gender by Language Group)

		Native Languages		
		East Asian	Romance	Total
Gender	Male	81	27	108
	Female	103	40	143
Total		184	67	251

Table J.4

Number of Masters of Skill 4 (Gender by Language Group)

		Native Languages		
		East Asian	Romance	Total
Gender	Male	70	22	92
	Female	98	34	132
Total		168	56	224

Table J.5.

Descriptive Statistics of the PPM and Total Score (East Asian Male, N = 239)

Descriptive statistics	PPM of skill 1	PPM of skill 2	PPM of skill 3	PPM of Skill 4	Total score
Mean	0.289	0.291	0.359	0.338	9.933
SD	0.343	0.353	0.403	0.372	4.338

Table J.6.

Descriptive Statistics of the PPM and Total Score (East Asian Female, N = 283)

Descriptive statistics	PPM of skill 1	PPM of skill 2	PPM of skill 3	PPM of Skill 4	Total score
Mean	0.243	0.327	0.367	0.355	10.056
SD	0.298	0.342	0.397	0.367	3.856

Table J.7

Descriptive Statistics of the PPM and Total Score (Romance Male, N = 57)

Descriptive statistics	PPM of skill 1	PPM of skill 2	PPM of skill 3	PPM of Skill 4	Total score
Mean	0.377	0.330	0.423	0.408	10.930
SD	0.378	0.343	0.414	0.400	4.088

Table J.8.

Descriptive Statistics of the PPM and Total Score (Romance Female, N = 90)

Descriptive statistics	PPM of skill 1	PPM of skill 2	PPM of skill 3	PPM of Skill 4	Total score
Mean	0.376	0.399	0.462	0.389	11.456
SD	0.344	0.368	0.402	0.403	3.757

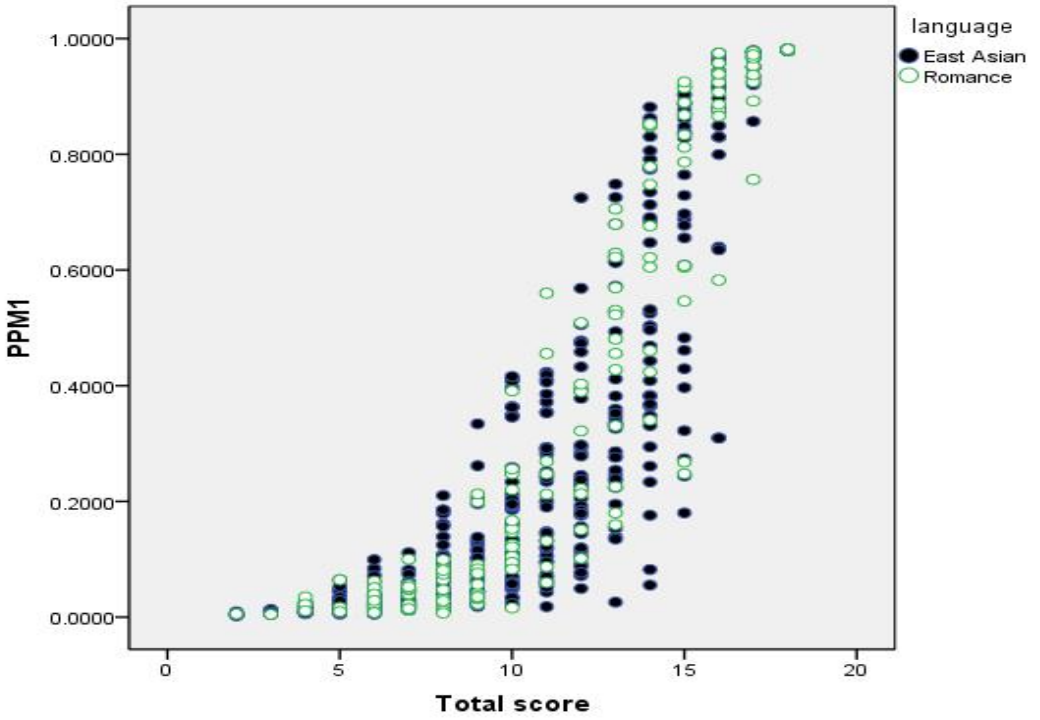

Figure J.1. Distribution of the PPM of skill 1 (vocabulary).

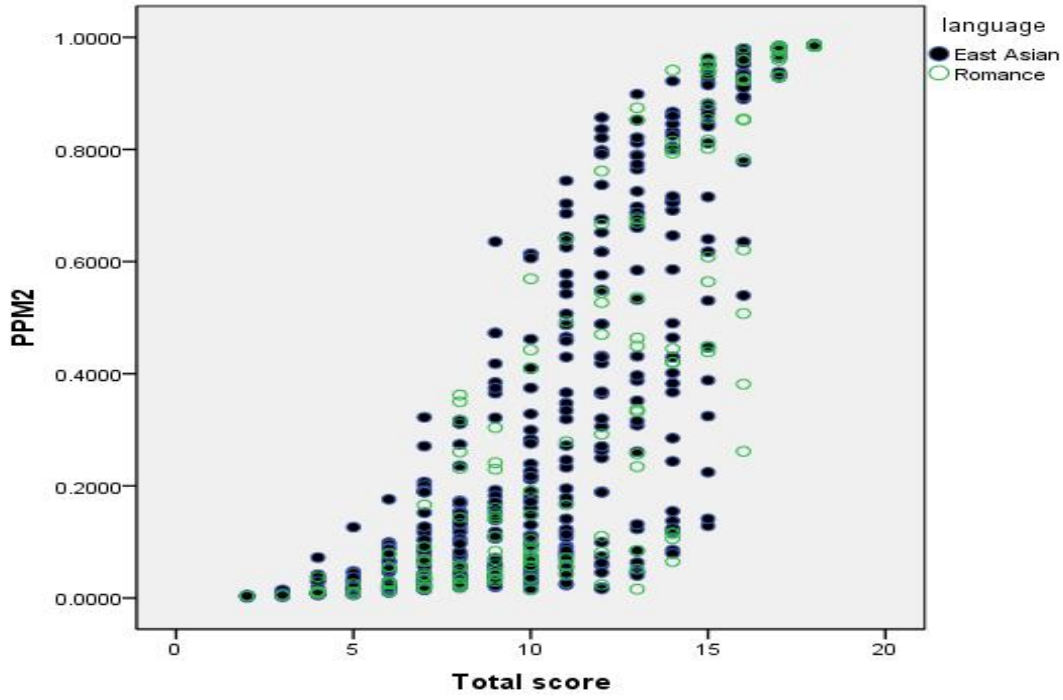

Figure J.2. Distribution of the PPM of skill 2 (syntax).

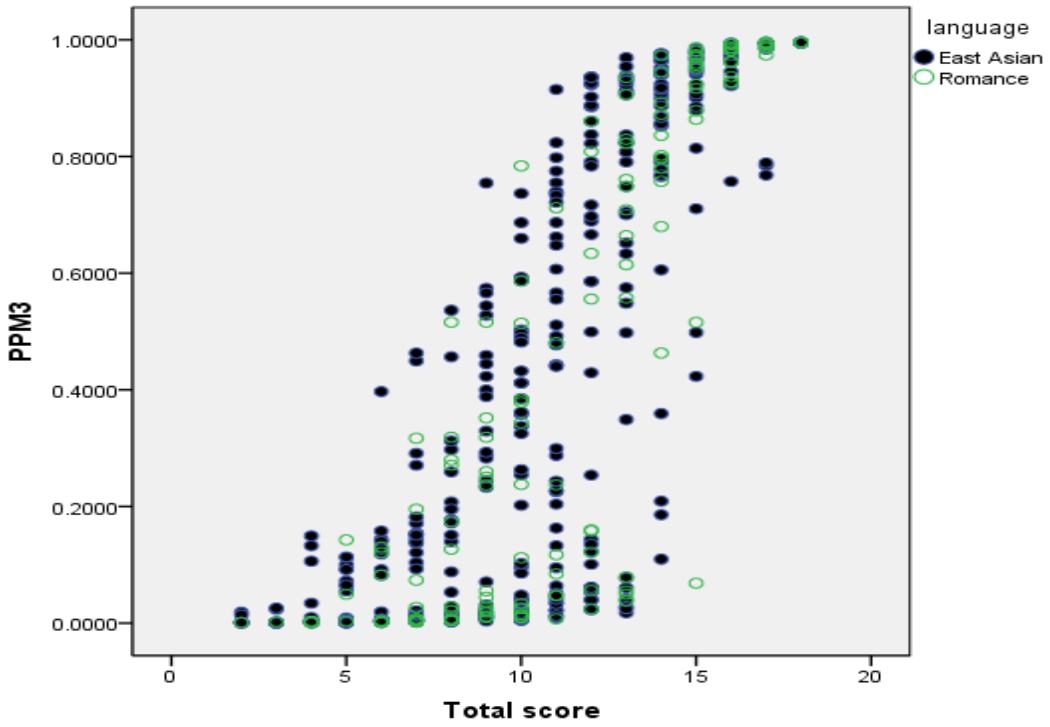

Figure J.3. Distribution of the PPM of skill 3 (extracting explicit information).

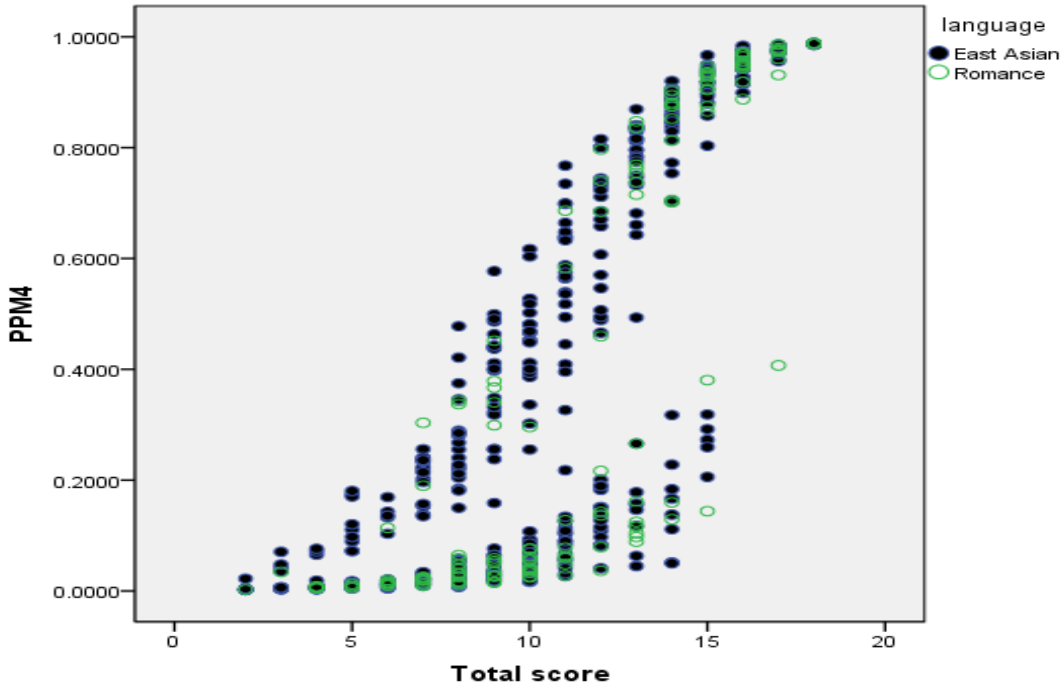

Figure J.4. Distribution of the PPM of skill 4 (connecting and synthesizing).

Appendix K R^2 change between Model 1 and Model 2

Table K.1

Summary of R^2 Change of Stage 1 Analysis

Skills	R^2 of Model 1	R^2 of Model 2	R^2 change between Models 1 and 2
Skill 1 (Vocabulary)	0.805	0.813	0.008
Skill 2 (Syntax)	0.708	0.710	0.002
Skill 3 (Extracting explicit information)	0.758	0.759	0.001
Skill 4 (Connecting and synthesizing)	0.743	0.746	0.003

Table K.2

Summary of R^2 Change of Stage 2 Analysis

Skills	R^2 of Model 1	R^2 of Model 2	R^2 change between Models 1 and 2
Skill 1 (Vocabulary)	0.806	0.814	0.008
Skill 2 (Syntax)	0.716	0.719	0.003
Skill 3 (Extracting explicit information)	0.759	0.759	0.000
Skill 4 (Connecting and synthesizing)	0.747	0.751	0.004

Appendix L Scatter Plots of Scale Scores versus PPMs

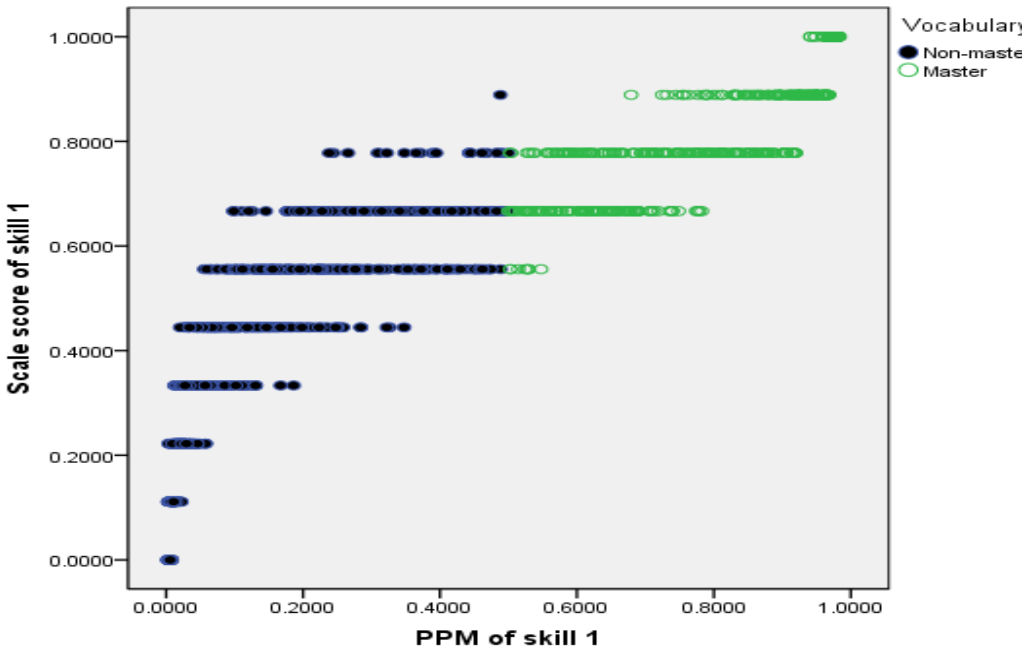

Figure L.1. Scatter plot of scale score versus PPM of skill 1.

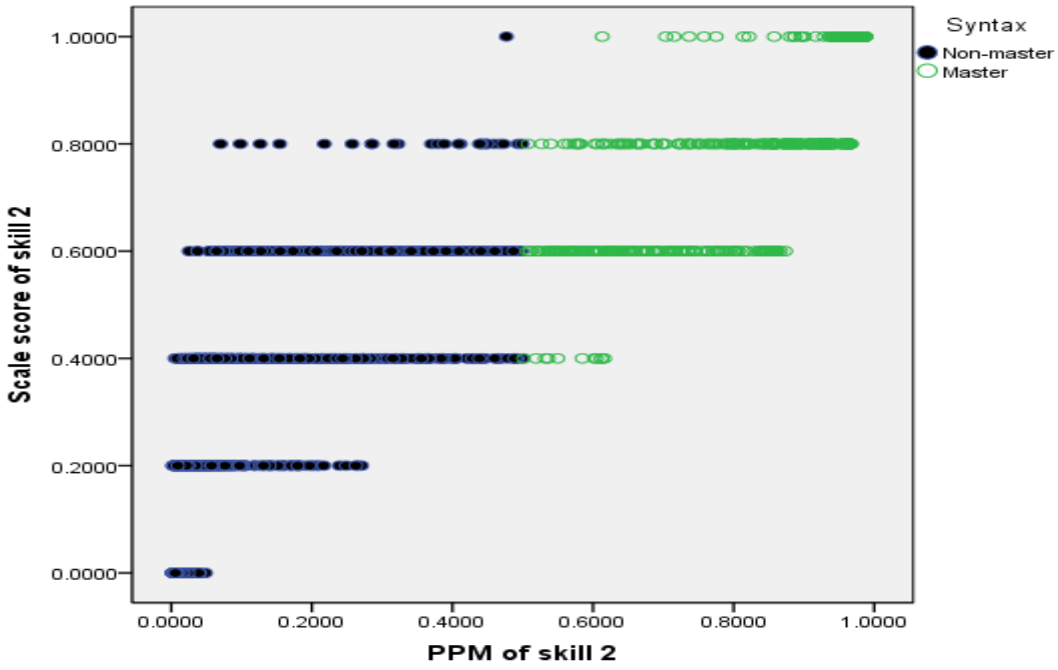

Figure L.2. Scatter plot of scale score versus PPM of skill 2.

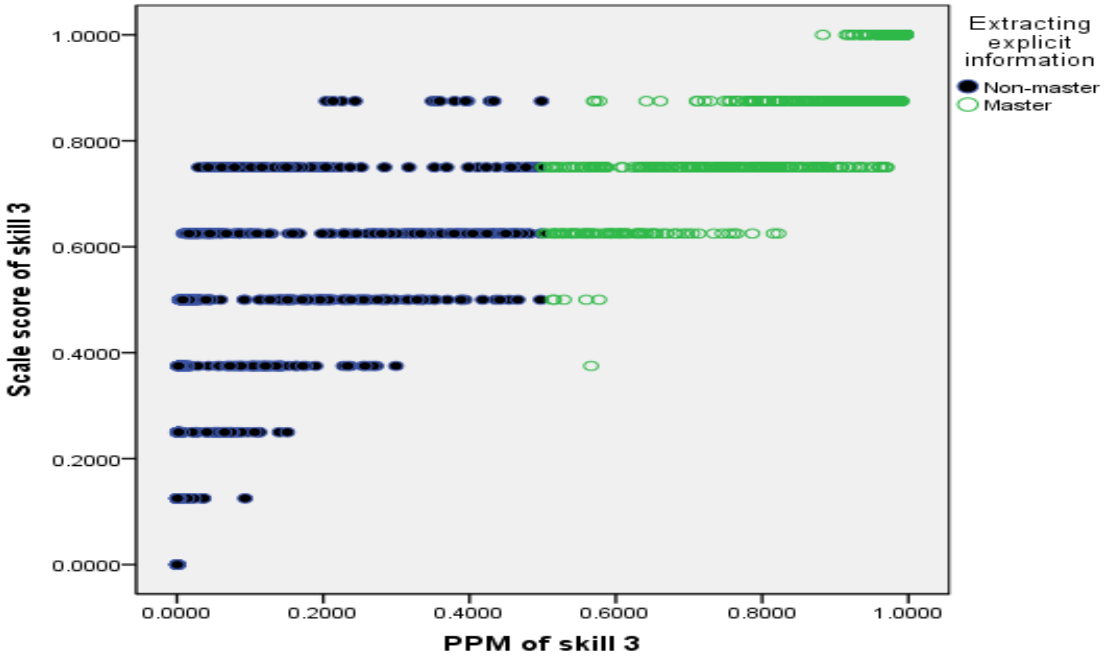

Figure L.3. Scatter plot of scale score versus PPM of skill 3.

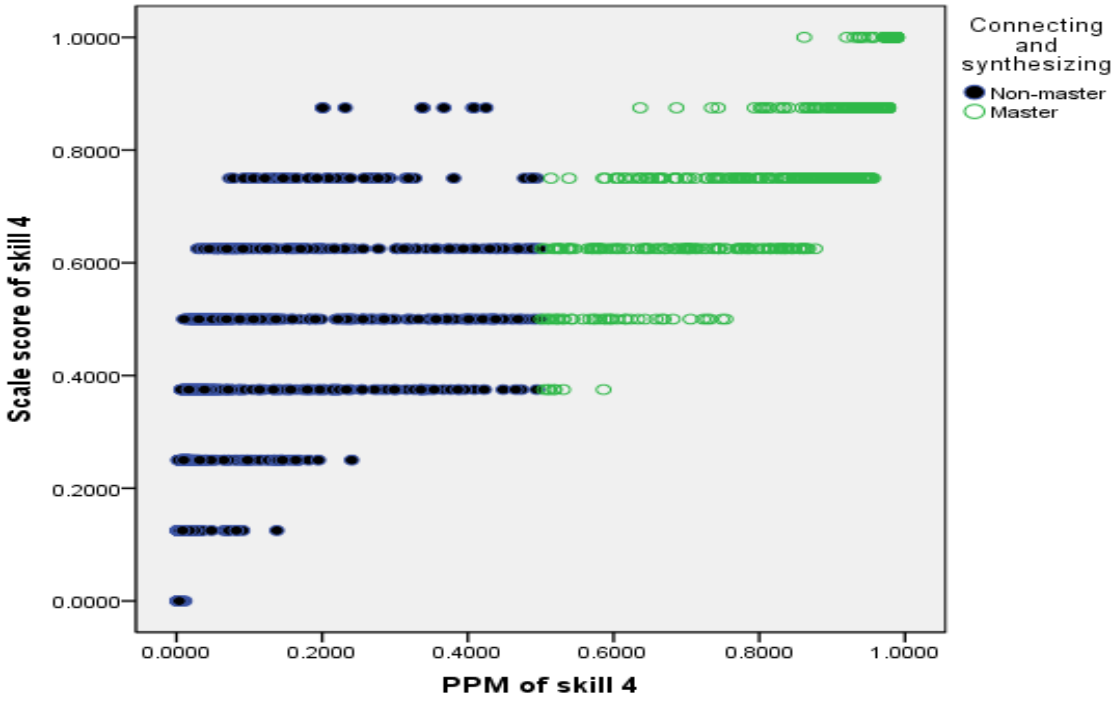

Figure L.4. Scatter plot of scale score versus PPM of skill 4.

Vita

Hongli Li was born in Fufeng, Shaanxi, China on March 28, 1977. She graduated from Xi'an Jiaotong University in 1998 with a Bachelor's degree in English for Science and Technology. She also earned a Master's degree in English for Business Purposes at Dongbei University of Finance and Economics in 2001. From July 2001 to July 2005, she taught at the Foreign Languages Department of Central University of Finance and Economics in Beijing.

From August 2005 to August 2007, she studied in the Division of English as an International Language at the University of Illinois at Champaign-Urbana and received a Master's degree in Teaching English as a Second Language (TESL). During that two-year period, she taught ESL courses to international graduate students and developed an interest in language assessment and statistics.

Since August 2007, she has been a doctoral student in the Educational Psychology Program (measurement track) at the Pennsylvania State University, and she anticipates receiving a doctoral degree in May 2011. As of June 2011, she will be an assistant professor of research, measurement, and statistics in the Department of Educational Policy Studies at Georgia State University.