Spaan Fellow Working Papers in Second or Foreign Language Assessment
Copyright © 2011
Volume 9: 17–46

University of Michigan

A Cognitive Diagnostic Analysis of the MELAB Reading Test

Hongli Li
Georgia State University

ABSTRACT With cognitive diagnostic analysis, each examinee receives a multidimensional skill profile expressing whether he/she is a master or nonmaster of each skill measured by the test. Fine-grained diagnostic feedback that facilitates teaching and learning can thus be provided to teachers and students. This study investigated cognitive diagnostic analysis as applied to the Michigan English Language Assessment Battery (MELAB) reading test. The Fusion Model (Hartz, 2002) was used to estimate examinee profiles on each reading subskill underlying the MELAB reading test. With data collected from multiple sources, such as the think-aloud protocol and expert rating, a tentative Q-matrix was initially developed to indicate the subskills required by each item. This Q-matrix was then validated via an application of the Fusion Model using data from the MELAB reading test. Four subskills were found to underlie the test, e.g., vocabulary, syntax, extracting explicit information, and understanding implicit information. Examinee skill mastery profiles were produced as the result of the cognitive diagnostic analysis. Finally, issues involved in the cognitive diagnostic analysis of reading tests were discussed, and areas for future research were also suggested.

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 reported score. 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). A typical procedure of the cognitive diagnostic analysis using CDMs is as follows: (i) identifying a set of skills involved in a test; (ii) demonstrating which skills are required for correctly answering each item in the test; (iii) estimating the profiles of skill mastery for individual examinees based on actual test performance data using the CDM; and (iv) providing score reporting and/or diagnostic feedback to examinees and other stakeholders (Lee & Sawaki, 2009b). Fine-grained diagnostic feedback can thus be provided to facilitate teaching and learning.

Despite their relatively new status, CDMs have been actively applied to large-scale language tests. For instance, the Rule Space Model (Tatsuoka, 1983) has been applied to the TOEFL reading (e.g, Kasai, 1997; Scott, 1998), the TOEIC reading (Buck, Tatsuoka, & Kostin, 1997), the TOEIC listening (Buck & Tatsuoka, 1998), and the SAT Verbal (Buck et al., 1998). Recently, the Fusion Model (Hartz, 2002) has been used for diagnostic analysis of the TOEFL iBT reading and listening (Jang, 2005; Lee & Sawaki, 2009a; von Davier, 2005). Wang and Gierl (2007) have also applied the Attribute Hierarchy Method (AHM) (Leighton, Gierl, & Hunka, 2004) to SAT Verbal. Despite the challenge in understanding the content domains and the complexity of the psychometric modeling procedure, these studies have shown the potential of using CDMs with existing language tests.

The Michigan English Language Assessment Battery (MELAB) is developed by the English Language Institute at the University of Michigan (ELI-UM) to evaluate advanced-level English language competence of adult nonnative speakers of English who will use English for academic purposes in a university setting. It consists of three parts: Part 1 composition, Part 2 a listening test comprising 60 multiple-choice items, and Part 3 a grammar/cloze/vocabulary/reading test with a total of 110 multiple-choice items. There is also an optional speaking test. A score for each part is reported, and the final MELAB score is also reported, which is the average of the scores of Part 1, Part 2, and Part 3. The speaking test score is not averaged into the Final MELAB score (ELI-UM, 2010).

The reading section of the MELAB is designed to assess examinees' understanding of college-level reading texts. It consists of four passages, each of which is followed by five multiple-choice items. According to the item-writing guidelines provided by the ELI-UM, 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). At present, since reading is only part of the grammar/cloze/vocabulary/reading section, no score is provided specifically to indicate an examinee's reading competence. Thus the diagnostic information on reading is rather limited. Gao (2006) developed a model of the cognitive processes used by examinees taking the MELAB reading test and validated the model with the tree-based regression (TBR) (Sheehan, 1997). This investigation has set a foundation for studying the diagnostic potential of the MELAB reading test.

In order to maximize the instructional and washback values of the MELAB, it is useful to explore how the CDMs can be used with the MELAB reading test. The purpose of this study is thus to investigate the use of cognitive diagnostic analysis with the MELAB reading test so as to provide rich diagnostic information for examinees.

Literature Review

Overview of Cognitive Diagnostic Models

With a CDM, examinees are assigned multidimensional skill profiles that classify them as masters or nonmasters of each skill involved in the test. Despite 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).

A large number of CDMs have been proposed (62 models as listed by Fu & Li, 2007). 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) is an updated version of the Rule Space Model. It specifies the hierarchical relations 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 (see Roussos, Templin, & Henson, 2007 for a full review). 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 the observed response by breaking the overall ability down into different components. The number of dimensions depends on the number of skill components involved in the assessment.

Second, CDMs are inherently confirmatory. The loading structure of a CDM is the Q-matrix, i.e., a particular hypothesis about which skills are required for successfully answering each item. We will let *k* stand for the number of skills being measured, *i* for the number of items, and *j* for the number of examinees. $Q = \{q_{ik}\}$, where $q_{ik} = 1$ when skill *k* is required by item *i*, and $q_{ik} = 0$ when skill *k* is not required by item *i.* As shown in Table 1, skill A is required by item 1, whereas skill B and skill C are required by item 2.

Table 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 noncompensatory 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 noncompensatory models are the Rule Space Model, the Attribute Hierarchy Method, the DINA (deterministic input noisy and) model of Haertel (1984, 1989, 1990), the NIDA model of Junker and Sijtsma (2001), the HYBRID Latent Class Model of Gitomer and Yamamoto (1991), and 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) are compensatory.

Finally, unlike traditional IRT models which generally model continuous latent variables, the latent variables modeled in CDMs are discrete. Currently, most CDMs and the associated estimation procedures only allow for dichotomous latent variables (e.g., mastery vs. nonmastery), 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 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.

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:

$$PXij=1\alpha j,\theta j=\pi i* \ k=1krik*(1-\alpha jk) \ qikpci \ (\theta j) \qquad (1)$$

Where,

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

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

There are two ability parameters, α_j and θ_j:.

α_j refers to a vector of cognitive skill mastery for examinee *j* for the skill *k* specified by the Q-matrix (α_{jk} = 1 if examinee *j* has mastered skill *k*; α_{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 skills required for item *i*, will correctly apply all the skills to solving the 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 the item 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 lower the c_i is, the more the item response function depends on residual ability θ_j. Therefore, c_i is regarded as the Q-matrix completeness index.

The biggest advantage of the Fusion Model over other CDMs is that it acknowledges the incompleteness of the Q-matrix and compensates for this by including the residual parameter c_i, which represents all the other skills that have been used by the examinees but have not been specified in the Q-matrix (Roussos, DiBello, et al., 2007). As we do not have a full understanding of the cognitive processes underlying reading comprehension, 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 (Bolt et al., 2008) helps to modify the Q-matrix by removing nonsignificant item parameters, thereby facilitating the process of building a valid Q-matrix. As demonstrated in Hartz (2002), the Fusion Model uses a stepwise reduction algorithm to increase the estimation accuracy of the item parameters by eliminating noninformative 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 2.0 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 the item 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).

The Fusion Model has been intensively studied in 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. Henson and Templin (2004) developed a procedure for analyzing National Assessment of Educational Progress (NAEP) data with the framework of the Fusion Model.

Due to its relatively new status, the Fusion Model 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 TOEFL iBT. 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 TOEFL iBT 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. Overall, 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 TOEFL iBT reading and listening were used. Different from Jang's study, only four skills were identified as underlying the TOEFL reading test. 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).

Overall, given the complexity of reading comprehension, the Fusion Model has great potential for conducting cognitive diagnostic analysis with reading tests. Therefore, in the current study, the Fusion Model was applied to the MELAB reading test to investigate its diagnostic potential.

Q-Matrix Construction and Validation

The Q-matrix is an essential input for using the Fusion Model and any other CDMs. However, because the construct being tested and the underlying cognitive processes associated with it are usually not fully understood, establishing a Q-matrix, especially for an existing test, is a challenging task. In the present study, the following procedures were followed to construct and validate the Q-matrix for the MELAB reading test.

Initial Q-Matrix Construction

Based on information gathered from students' verbal reports and from content experts, Gao (2006) developed a model of cognitive process underlying the MELAB reading test. Both the MELAB and TOEFL are English-language proficiency tests used by North American universities to determine whether to admit international students. Therefore, Gao's cognitive model was compared to the TOEFL reading model constructed by Jang (2005).

As shown in Table 2, the reading subskills of the MELAB and TOEFL fall into five categories: vocabulary, syntax, extracting explicit information, connecting and synthesizing, and making inferences. In the present study, this initial framework was further revised and validated with evidence from students' verbal reports, expert ratings, and the extant literature.

Table 2. Summarizing Cognitive Models of Reading Built by Gao and Jang

Category	Gao (2006)	Jang (2005)
Vocabulary	• Recognize and determine the meaning 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; scan the text for specific details, which includes (i) matching key vocabulary items in the question to key vocabulary items in the relevant part of the text, and (ii) identifying or formulating a synonym or a paraphrase of the literal meaning of a word, phrase, or sentence in the relevant part of the text	• 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; skim the text for gist	• Summarizing • Mapping contrasting ideas into mental framework
Making inferences	• Speculate beyond the text, e.g., use background/topical knowledge • Draw inferences and conclusions based on information implicitly stated in the text	• Inferencing • Textually implicit information

Think-Aloud Protocol

To supplement the initial framework shown in Table 2, think-aloud protocols (Ericsson & Simon, 1993; Pressley & Afflerbach, 1995) were conducted in order to gather information about possible cognitive processes involved in responding to the MELAB items. In total, 13 ESL learners participated in the study, and their background information is shown in Table 3.

Table 3. Background Characteristics of Think-Aloud Participants

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

A brief training session was provided prior to the formal think-aloud activity. During the concurrent think-aloud session, the participant talked out what he/she was thinking while reading the passage and responding to the 20 reading items of the MELAB Form E. It was important not to distract the participant; therefore, only when a silence of 10 seconds or so had occurred, would I prompt the participant with questions such as "What are you thinking now?" Then after finishing all five questions following one passage, the participant would recount the processes he/she had used. At this retrospective think-aloud session, I asked some questions mainly for clarification and further inquiry. The whole process was recorded using a digital voice recorder.

I read through the transcribed verbal reports line-by-line in order to understand the reading skills involved. The initial framework was mostly confirmed by the data. First, it was difficult to distinguish whether students determined the meaning of specific words by using context clues or by using phonological/orthographic/vocabulary knowledge. 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 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 appeared to involve different levels of elements. In some cases, students only needed to read and connect information from adjacent sentences in a single paragraph. However, in other cases, students had to read across different paragraphs or the whole passage in order to identify the main idea of the passage. Only two items were found to test main ideas, and thus it was not practical to have a separate skill for main ideas. Therefore, a final decision was made to use the general skill designated as connecting and synthesizing. The fifth category pertained to making inferences, in which students went beyond the text in order to draw conclusions based on implicit information in the text.

In addition to the above five skills, skills relating to metacognition, test-taking, and guessing were noticeable in the think-aloud verbal reports. For instance, some students read the questions before reading the passage or skipped the question when they were not able to answer a question upon first encountering it. Also, some students consistently guessed at or eliminated alternative choices. A residual skill category seemed to exist, 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.

With reference to Gao (2006) and Jang (2005), a coding scheme, as shown in Table 4, was built based on the cognitive framework and the think-aloud data. The think-aloud data helped to build the coding scheme, which was later used to guide the coding of the data.

Table 4. 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.

Expert Rating

Four experts were invited to identify the skills required by each of the 20 items in the MELAB Form E. All were advanced doctoral students in education or applied linguistics and had experience in teaching English reading to ESL students. Their qualifications and experience are summarized in Table 5. Each expert was provided with the four MELAB reading passages, a one-page introduction to the MELAB, a coding scheme, and a coding form. In order to acquaint the experts with the rating task, a training session was held prior to the formal rating.

Table 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

After training, each expert read the passages and performed the rating task independently. They identified the skills for each item and also made annotations about the evidence on which they based their assessments. When they had finished rating each passage, the experts convened and compared their ratings. Spearman rho was calculated to indicate the agreement between the ratings given by each expert. As shown in Table 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 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).

Initial Q-Matrix

With reference to the coding scheme, an initial Q-matrix was constructed based on evidence from the think-aloud verbal report and the expert rating. However, a frequently encountered problem here is that students' verbal reports may not agree with expert rating (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 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 7. The number 1 indicates that the skill is required by the item, whereas 0 indicates that the skill is not required by the item. The residual skills were thought to be affiliated with each item; their coding is thus not listed in the table.

Table 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

Empirical Validation of the Q-Matrix

Response data from 2,019 examinees to each of the 20 reading items of the MELAB Form E were used for the empirical validation. 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 or may have run out of time and thus were not instigating the processes required by item solution (Plough, personal communication, March 25, 2010). The data set were analyzed with Arpeggio, and the following procedures were used.

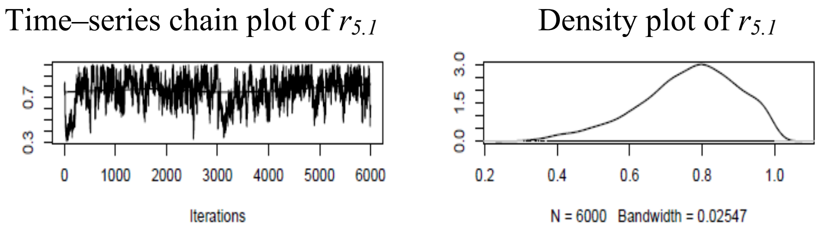
MCMC Convergence Checking

The software Arpeggio uses 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 parameters, which may provide a better understanding of the true standard errors involved (Patz & Junker, 1999). However, 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. A time–series chain plot provides a graphical check of the stability of the generated parameter values, whereas a density plot checks graphically if the mean of a parameter has stabilized. Other criteria, such as the Heidelberg–Welch diagnostic and the Geweke *Z,* were also examined. 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 is used, and small p-values (such as < 0.05) indicate non-convergence. 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 (Ntzoufras, 2009).

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 will 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 (Sinharay, 2004). After several trials, a chain length of 60,000 and burn-in steps of 30,000 was found to be appropriate.

Visual examination of the plots showed that 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), $r_{5.1}$ (the 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 1, 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.


Time–series chain plot of $r_{5.1}$

Density plot of $r_{5.1}$

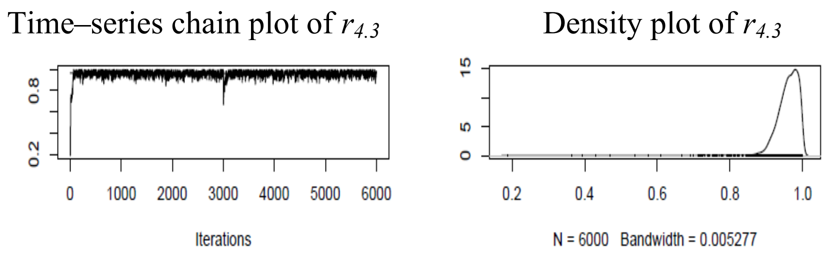

Time–series chain plot of $r_{4.3}$ Density plot of $r_{4.3}$

Figure 1. Sample Time–Series Chain Plots and Density Plots.

Some numerical criteria were also used to help judge MCMC convergence. As shown in Table 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 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)

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 7 for details). Thus, the information pertaining to skill 5 was probably insufficient for estimation. Therefore, I decided to combine skill 4 (connecting and synthesizing) and skill 5 (making inferences) to create a new skill 4: "understanding implicit information through connecting ideas and making inferences." Hereafter, this new skill 4 is referred to as "understanding implicit information."

High values of *r* and *c* parameters indicate possibility for model simplification (Hartz, 2002; Roussos, DiBello, et al., 2007). It has been recommended that an *r* parameter larger than 0.9 should be removed from the Q-matrix, as this indicates that the affiliated skill is not significantly important for the item. Also, when *c* is larger than 2, the skills required to successfully answer the item are completely specified by the Q-matrix, and thus *c* can be dropped (Hartz, 2002). A more parsimonious model with few parameters is usually preferred. However, whether to drop a certain Q-matrix entry depends on both statistical criteria and substantive knowledge. First, the seven large *c* parameters were dropped from the Q-matrix one at a time, as they did not greatly change the Q-matrix structure. Then five of the large *r* parameters 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 retained because the skill affiliated with each of these was the only one 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 fluctuations. All the parameters met the Heidelberg–Welch diagnostic and Geweke Z convergence criteria.

Model Fit

Just as with any other statistical models, only when the Fusion Model fits the data, the interpretation of the estimated parameters is meaningful. 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.

The first index is the residual between the observed and model-predicted p-values 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. The chart at the top of Figure 2 shows the observed p-value versus the predicted 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 predicted p-value when the refined Q-matrix was used. The two lines were very close or overlapped for most of the items. Table 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.

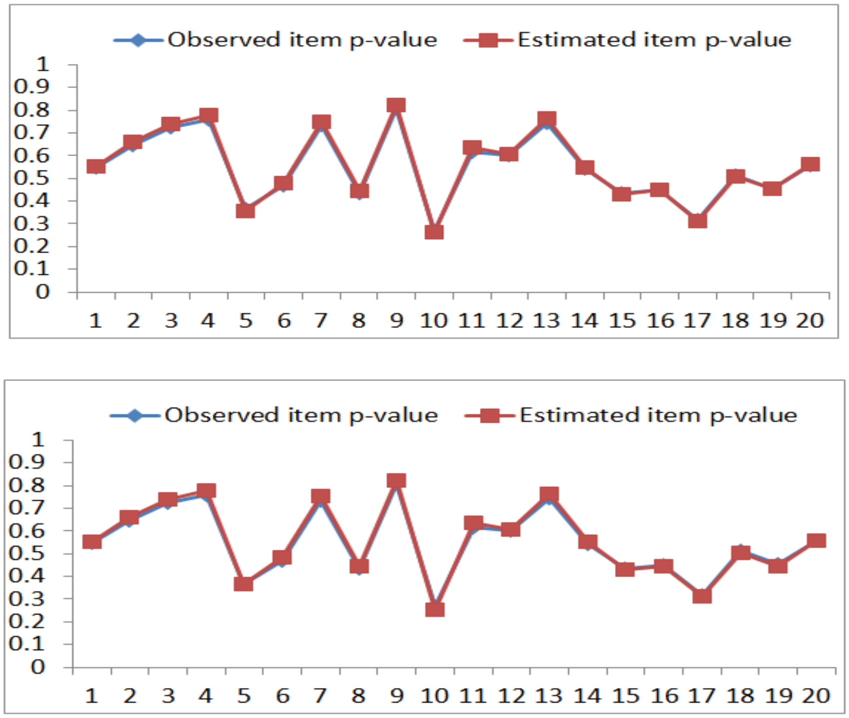


Figure 2. Observed Versus Predicted P-Values Across Items.

Table 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	-0.006
Mean square error	0.000	0.000

Second, 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. As shown in Table 10, when both the observed and predicted total scores for 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 model had more parameters. This has been a common phenomenon because adding more parameters to the model always improves overall model fit; however, the question is whether the improved model fit is worthy given that the model is more complex than before (Kline, 2005). The purpose is to find a parsimonious model that still fits the data reasonably well.

Table 10. Comparison of Observed and Predicted Total Score across Examinees

Difference between observed and predicted total score	Initial Q-matrix	Refined Q-matrix
Mean	0.000	0.000
Mean square error	0.070	0.080

The scatter plots of the observed and predicted total scores for all 2,019 examinees are shown in Figure 3. 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 charts. The correlation between the observed and predicted total scores was 0.960, regardless of whether the initial Q-matrix or 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 (Jang, 2005; Roman, 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 result (Roussos, DiBello, et al., 2007).

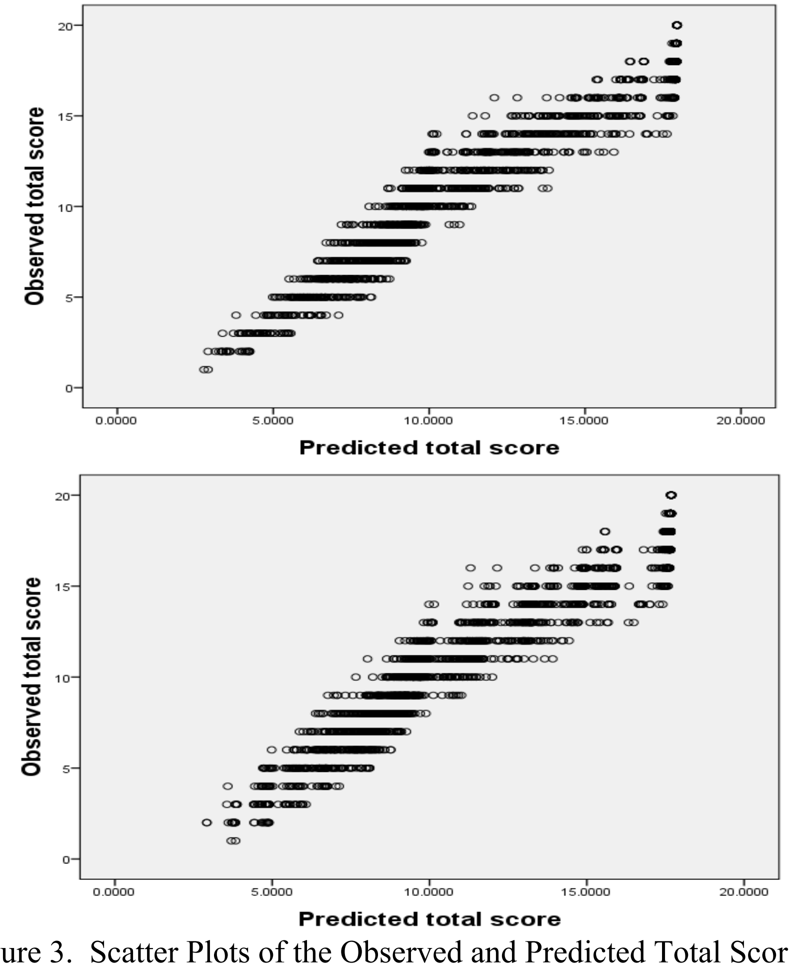


Figure 3. Scatter Plots of the Observed and Predicted Total Scores.

ImStats computes the observed proportion-correct score 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 skill classification. Therefore, *ImStats* is also used as internal validity evidence, because it uses the test data itself to help verify the authenticity of the model (DiBello, Roussos, & Stout, 2007).

In Figure 4, 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 11, the average proportion-correct score of the item masters was around 0.9 in both cases, whereas the average proportion-correct score of the item non-masters was less than 0.45. To summarize, the differences, as shown in Figure 4 and Table 11, provide important evidence for good model fit.

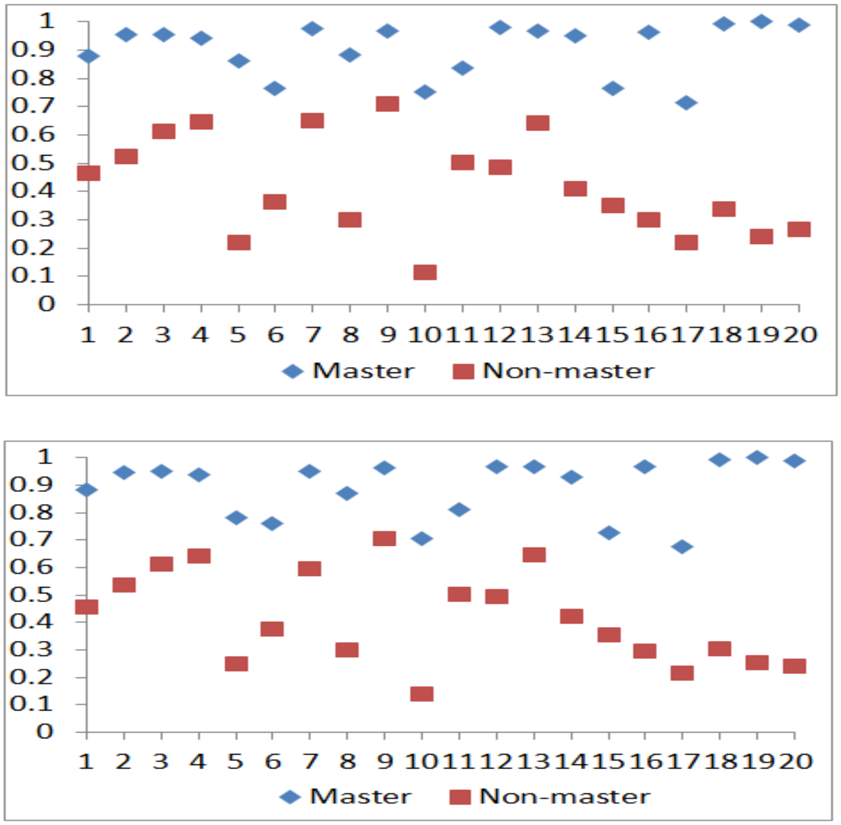


Figure 4. Proportion-Correct Scores of Item Masters and Non-Masters.

Table 11. Comparison of Average Proportion-Correct Scores of Masters and Nonmasters

Statistic	Initial Q-matrix	Refined Q-matrix
Mean proportion-correct score of item masters	0.905	0.888
Mean proportion-correct score of item nonmasters	0.417	0.416

Based on the above model-fit evidence, the model fits 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 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.

Diagnostic Results

With the recommended Q-matrix and the item response data of the 2,019 examinees, the Fusion Model calibration was conducted using Arpeggio software. The estimated item parameters and examinee skill mastery status are reported as follows.

Item Parameters

The item parameters are shown in Table 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. For comparison, the last column shows the proportion-correct score of the item.

Table 12. Item Parameters of the Final Calibration

Item	π	r*1	r*2	r*3	r*4	C	Proportion-correct score
1	0.860	0.710	■		0.849	1.649	.548
2	0.984	0.853		■		0.868	.648
3	0.972				0.852	1.296	.724
4	0.993			0.958		1.113	.761
5	0.752	0.495	■		0.729*	1.701	.365
6	0.765	0.880		0.653		1.333	.471
7	0.983		■	0.852		1.266	.736
8	0.885	0.579			0.815	1.020	.436
9	0.985			0.918		1.519	.807
10	0.643	0.237			0.821*	■	.265
11	0.872			0.868		1.025	.618
12	0.971	0.757	0.828	■		1.337	.600
13	0.995				0.914	1.154	.746
14	0.986	0.766			0.852	0.762	.541
15	0.723	0.804	0.697		0.801*	■	.436

16	0.936	0.854	0.635	0.441		[redacted]	.448
17	0.619		0.414		0.753	[redacted]	.318
18	0.967		0.438	0.712		[redacted]	.512
19	0.976	0.861			0.254	[redacted]	.454
20	0.954			0.306		[redacted]	.558

Note. Items 5, 10, and 11 were originally affiliated with skill 5 (making inferences).

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 the table was 0.891, indicating that the identified skills for the items were generally adequate and reasonable. However, the π parameter for item 17 was as low as 0.619. This indicates that the probability that examinees would correctly answer item 17 was only 0.619, given that they had acquired the required skills of syntax and understanding implicit information. Item 17 was a rather difficult item. As shown in Table 12, the proportion-correct score for items 17 was only 0.318, whereas the average proportion-correct score for all the items was 0.550. This is probably one of the reasons that 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*. For example, $r_{10.1}$ was 0.237. This indicates that the probability of an examinee correctly answering item 10 when he/she has not mastered skill 1 (vocabulary) is 0.237 times of the probability of correctly answering item 10 when skill 1 has been mastered. This shows that vocabulary is a very important skill for item 10. However, some *r* parameters were rather large. For instance, $r_{7.3}$ was 0.852. This indicates that the probability of correctly answering item 7 when skill 3 (extracting explicit information) has not been mastered is 0.852 times of 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 Table 12, 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 *r* parameters of the MELAB items were on the higher end, indicating that the diagnostic capacity of the MELAB reading test is low. This is probably because the MELAB reading test is not built for diagnostic purposes.

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 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 $pci(\theta j)$ 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. The outcome of the Fusion Model was examined, and this 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 eight *r* parameters were larger than 0.9. 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 also 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 seven large *c* parameters were dropped for model parsimony, while the rest of the *c* parameters were maintained in the Q-matrix.

Examinee Skill Mastery Status

Continuous posterior probability of mastery (PPM) indicates the probability that an examinee is a master of the skill being studied. As shown in Figure 5, 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.

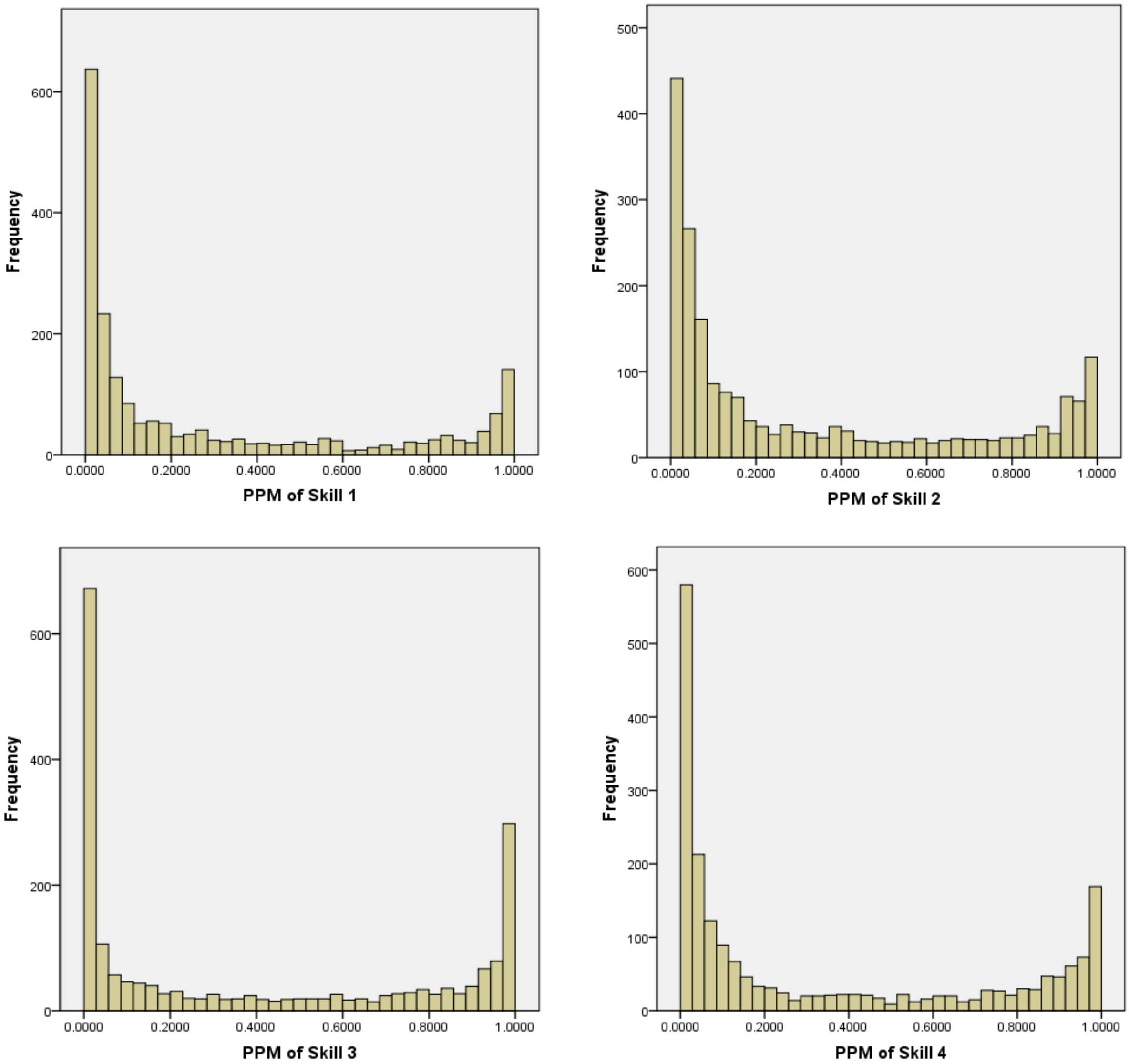


Figure 5. Continuous Posterior Probability of Mastery *(PPM)*.

In addition, a dichotomous mastery/non-mastery status can be accomplished by using a cut-off point of 0.5 (Lee & Sawaki, 2009a). 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 6, about 25.6% of examinees were masters of skill 1 (vocabulary), 28.7% were masters of skill 2 (syntax), 40.1% were masters of skill 3 (extracting explicit information), and 32.3% were masters of skill 4 (understanding implicit information). 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, only less than 7% of examinees had PPMs between 0.4 and 0.6 for all the four skills; therefore, a more refined polytomous status would not have changed the classification results much.

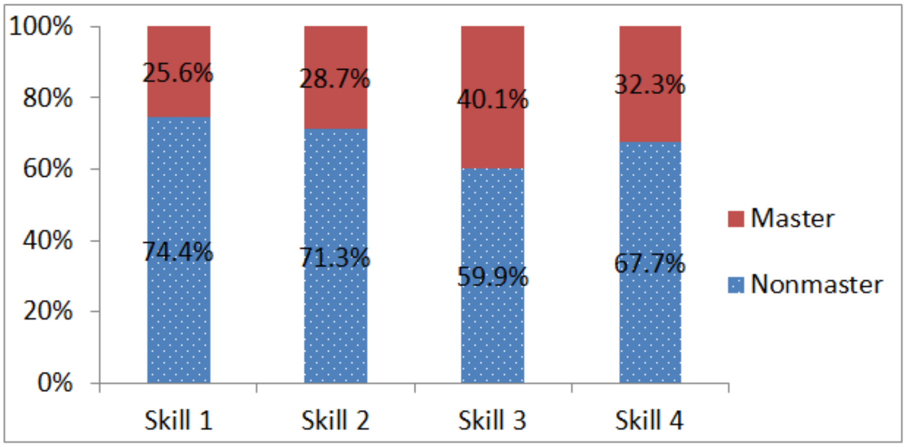


Figure 6. Categorical Mastery Status.

These results are relatively congruent with my expectation. Skill 1 (vocabulary) seemed to be the most difficult skill for MELAB examinees, which agreed with the consistent finding that lack of vocabulary is the major obstacle in reading comprehension (Garc*ia*, 1991). A rule of thumb is that readers need to know 95% of the words in a text to read it successfully (Grabe, 2009); however, this is rarely the case for ESL learners. In addition, 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 understanding implicit information through connecting ideas and making inferences. Thus, skill 4 requires more cognitive processes and is 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 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 (understanding implicit information). As illustrated in Table 13, about half of the examinees were non-masters of all four skills, i.e., their profiles showed 0000, and about 16% were masters of all four skills, i.e., their profiles showed 1111. The third largest number of examinees had the 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 13. Skill Mastery Patterns

Skill mastery patterns	Frequency	Percentage
0000	1042	51.61%
1111	324	16.05%
0010	136	6.74%
0111	96	4.75%
0011	70	3.47%
0110	60	2.97%
1011	58	2.87%
0001	52	2.58%
1110	37	1.83%
1000	31	1.54%
1010	29	1.44%
0100	24	1.19%
0101	22	1.09%
1001	22	1.09%
1101	9	0.45%
1100	7	0.35%

Figure 7 illustrates a sample skill profile report for an MELAB examinee about his/her mastery of the reading skills. The horizontal axis shows the PPMs of each skill, and a vertical line indicates the cut-off point of 0.5. As can be seen, this examinee is a master of skills 1(vocabulary), 3(extracting explicit information), and 4 (understanding implicit information), but not skill 2 (syntax). Providing such a skill profile report can help the examinee identify his/her weakness, and thus more efficiently facilitate the learning and teaching of reading comprehension.

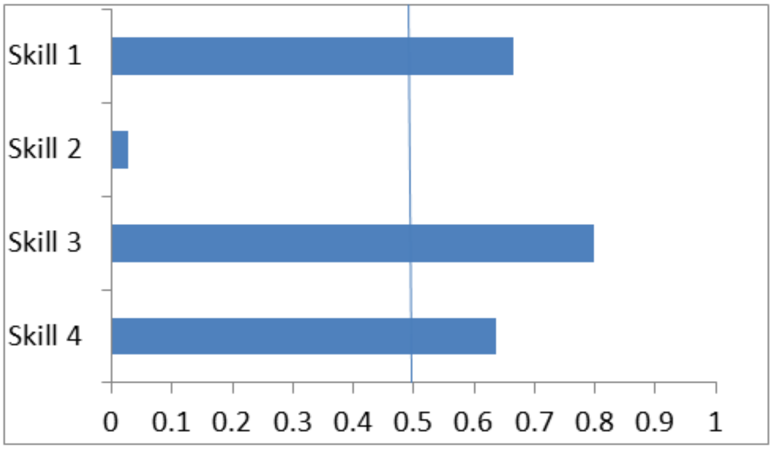


Figure 7. Sample Skill Profile.

Discussion

The present study yielded useful diagnostic information to MELAB examinees than are currently available to. However, cognitive diagnostic modeling is new to the field of language assessment, and many issues need further investigation.

Retrofitting the MELAB Reading Test with CDMs

As suggested 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 for existing tests. Retrofitting the MELAB reading test has produced more diagnostic information than if only a total score is provided. It also deepens our understanding of the MELAB reading test so as to accumulate empirical evidence for further diagnostic assessment and test development. However, retrofitting with preexisting tests involves a time-consuming process of Q-matrix construction, which may yield results that are not optimal.

A noticeable indeterminacy involved in the retrofitting 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 the test. Two major factors considered were the modeling capacity and the meaningfulness of the skill mastery profile. Gao (2006) suggested that ten reading skill components underlie the MELAB test. However, given the fact that this test consists of only 20 items, the present study only involved five subskills: vocabulary, syntax, extracting explicit information, synthesizing and connecting, and making inferences. However, only three items were initially identified as requiring skill 5 (making inferences). In order to have more information for parameter estimation, skill 5 (making inferences) and skill 4 (connecting and synthesizing) were collapsed into the more general skill of understanding implicit information through connecting ideas and making inferences. One important implication for the test developers is, therefore, to keep balance between the number of subskills being measured and the numbers of items in the test, i.e. more items should be included if more fine-grained diagnostic information is of interest.

Jang (2009) and Sawaki, Kim, and Gentile (2009) also commented on the skill granularity issue. For the same TOEFL iBT 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. It is also very important to note that 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 present study shows that it is possible to extract richer diagnostic information than the MELAB reading test was designed to elicit. However, retrofitting CDMs with existing tests is by no means an optimal approach for diagnostic assessment. 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). Thus, a successful cognitive diagnostic assessment of reading comprehension largely depends on test development, which again depends on more insightful understanding of the cognitive processes underlying reading comprehension.

Selecting Diagnostic Models for Reading Tests

With a large number of CDMs available, the question is which one to choose for reading tests. 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. Non-compensatory models have been preferred for cognitive diagnostic analysis, as they can generate more fine-grained diagnostic information. 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 TOEFL iBT reading and listening data, including the non-compensatory Fusion Model, the non-compensatory Latent Class Model (Gitomer &Yamamoto, 1991), and the compensatory General Diagnostic Model (von Davier, 2005). They found that “the examinee classification results were highly similar across the three cognitive diagnostic models (p. 258). Jang (2005) also found that reading skills involved in the TOEFL iBT appeared to be a mixture of non-compensatory and compensatory interactions.

In the literature on reading, Stanovich (1980) proposed a compensatory-interactive model. 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 another equally influential model, 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.

Also, it seems that the relationships between the subskills may depend on the relative difficulty levels of the subskills needed for solving a particular item. Thus this relationship may vary across items. If interested, one 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 for diagnostic analysis of reading tests.

Potential Use of Scale Scores for Diagnostic Assessment

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. The spearman rho correlation 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 (understanding implicit information) were respectively 0.956, 0.887, 894, and 0.878. Likewise, Jang (2005) also found high correlations between the scale scores and PPMs in her study with TOEFL reading, which was regarded evidence for the validity of the Fusion Model calibration.

The Fusion Model as well as other IRT-based CDMs have the advantages 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.

Acknowledgments

I would like to express my appreciation to the English Language Institute at the University of Michigan for funding this project and providing the data. I am also grateful for the funding from the TOEFL program of the Educational Testing Service. I especially wish to thank my advisor Dr. Hoi K. Suen at the Pennsylvania State University for his insightful comments and suggestions.

References

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, Madison, MI.

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. Tatsuoka, K. & Kostin, I. (1997). The sub-skills of reading: rule space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47* (3), 423–466.

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.

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

ELI-UM (2010). *The MELAB information and registration bulletin*. Retrieved from http://www.lsa.umich.edu/UMICH/eli/Home/Test%20Programs/MELAB/Officers%20&%20Professionals/MELABInfoBulletin.pdf

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

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

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.

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.

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.

Garc*ía*, G. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic students. *Reading Research Quarterly, 26*, 371–392.

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., & 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.

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

Gough, P. B., & Tunmer, W. E. (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.

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.

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, IL.

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., 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.

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, IL.

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

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, IL.

Kline, R. B. (2005) *Principles and practice of structural equation modeling*. (2nd Edition) Guilford Press.

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., 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.

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

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.

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

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

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.

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

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

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.

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.

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, IL.

Sheehan, K. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement*, *34*(4), 333–352.

Sinharay, S. (2004). Experiences with Markov Chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics, 29*, 461–488.

Stanovich, K.E. (1980). Toward an interactive compensatory model of individual differences in the development of reading proficiency. *Reading Research Quarterly, 16*(1), 32–71.

Tatsuoka, K.K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Templin, J. L. (2005). *Generalized linear mixed proficiency models for cognitive diagnosis*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana, 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.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2006). Multidimensional latent trait modelling (MDLTM) [Software program]. Princeton, NJ: Educational Testing Service.

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.

Yan, D., Almond, R. G., & Mislevy, R. J. (2004). *Comparisons of cognitive diagnostic models* (Research Report, RR-04-02), Princeton, NJ: Educational Testing Service.

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, Pennsylvania State University, State College, PA.