WANG, YING-CHEN, Ph.D. Factor Analytic Models and Cognitive Diagnostic Models: How Comparable Are They?—A Comparison of R-RUM and Compensatory MIRT Model with Respect to Cognitive Feedback. (2009)
Directed by Dr. Terry Ackerman. 99pp

The necessity and importance of cognitive diagnosis is being realized by more and more researchers. As a result, a number of models have been defined for cognitive diagnosis—the IRT-based discrete cognitive diagnosis models (ICDMs) and the traditional continuous latent trait models. However, there is a lack of literature that compares the newly defined ICDMs based on constrained latent class models to more traditional approaches such as a multidimensional factor analytic model. The purpose of this study is to compare the feedback provided to examinees using a multidimensional item response model (MIRT) versus feedback provided using an ICDM. Specifically, a Monte Carlo study was used to compare the diagnostic results from the R-RUM, a noncompensatory model with dichotomous abilities, to diagnoses made based on the 2PL CMIRT model, a compensatory model with continuous abilities. A fully crossed design was used to consider the effects of test quality, Q-matrix structure and inter-attribute correlation on the agreement rates of the diagnostic feedback for examinees between these two models. Given that one of the factors of this study is “test quality”, an initial study was performed to explore the possible relationship between test quality (including estimated model parameters) based on the models used to characterize examinee responses. In addition, because these models provide examinee information in different ways (one discrete and one continuous), a method using logistic regression, which is used to discretize the continuous estimates provided by the 2PL CMIRT, is discussed as a way to maximize

diagnostic agreement between these two models.

The significance of this study is that, if the two models agree consistently across the experimental conditions, model selection for cognitive purposes can be based largely on the preference of the researcher, which is informed by an underlying theory and assessment purposes. However, if the two models do not agree consistently, this study will help (1) to identify situations where the two models agree or disagree consistently and (2) to explore the feasibility of using the MIRT model for classifying examinees cognitively.

The results from the first study demonstrate that the two models define test quality in different ways and that item parameters of the two models are weakly associated. Therefore, subsequent comparisons are made within each model after estimating the R-RUM and the 2PL CMIRT, using common datasets. The results from the final study indicate that (1) the two models agree more consistently under the R-RUM generation, (2) there is a higher agreement rate between the two models under most scenarios of simple structure, (3) there is more error for both models under the MIRT generation, and (4) the MIRT model does not appear to be as successful at classification decisions as the R-RUM. Possible future directions are discussed.

FACTOR ANALYTIC MODELS AND COGNITIVE DIAGNOSTIC MODELS: HOW COMPARABLE ARE THEY? —A COMPARISON OF R-RUM AND COMPENSATORY MIRT MODEL WITH RESPECT TO COGNITIVE FEEDBACK

by

Ying-chen Wang

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2009

Approved by

_____
Committee Co-Chair

_____
Committee Co-Chair

To my husband, my son and my sisters

In memory of my parents

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of Graduate School at The University of North Carolina at Greensboro.

Committee Co-Chair _____

Committee Co-Chair _____

Committee Members _____

_____

_____

_____

April 27th, 2009_____
Date of Acceptance by Committee

February 3rd, 2009_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

I cannot be where I am without professional help from faculty of Educational Research Methodology (ERM) Department. It would be impossible for me to enjoy measurement and learn so much if I had not transferred to this program. I am very grateful to their help. The ERM professors are: Dr. Terry Ackerman, Dr. Richard Luecht, Dr. Robert Henson, Dr. Rick Morgan, Dr. John Willse and Deborah Bartz.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Traditionally, testing industries have focused on constructing measures to assess a single dimension. The test is assumed to measure only one latent or unobserved ability or skill via the measured variables or items. Each examinee is rank ordered based on the total item scores or a single continuous latent ability and therefore only a single score is reported. Such reports have been widely used for high-stake decisions such as college admissions, scholarship awards and even graduation. As a result, researchers and practitioners have applied various statistical tools to verify that only one latent ability is present in the data structure.

Despite its parsimonious nature, traditional scaling of examinees has some limitations. Most psychological and educational tests measure multiple skills and the unidimensionality assumption cannot be met under these circumstances (Hambleton & Swaminathan, 1985). In addition, it falls short of cognitive psychology in the twentieth century. Cognitive psychometrics involves measurement models assessing high-order thinking, which is related to a set of skills. It is commonly agreed that research in high-order thinking is fundamental to the testing industry, as many tests are based on cognitive problem-solving skills (Gierl, Leighton, & Hunka, 2000). As a summative assessment model, traditional modeling, such as unidimensional item response theory (IRT) models, might be appropriate. However, traditional assessment is limited in its ability to provide any formative feedback for improving instruction,

learning and curriculum development. Principals, teachers and educators need more informative reports for classroom instructions and intervention programs. This urgent public demand is culminated in the No Child Left Behind Act (2001), which explicitly calls for 'interpretive, descriptive and diagnostic reports' and the use of assessment results for improving students' academic achievements. Whereas both forms of assessment are necessary, one during the learning and teaching process and the other at the end of the instruction, formative assessments are more useful diagnostically at the classroom level throughout the course of instruction. In the simplest case, formative assessments should determine mastery or non-mastery for a set of *K* skills.

Recently, a variety of probabilistic latent class models have been developed for cognitive diagnostic purposes. These models assume that classes are defined by a set of discrete latent abilities, either binary or multicategorical. Each of these IRT-based cognitive diagnostic models (ICDMs) has an item response function (IRF) that predicts the probability of the correct response for each item, given the attribute status of each examinee on each skill. As in IRT, the use of an IRF enables researchers to evaluate the quality of test items through the evaluation of the item parameters. Once an appropriate model is selected, each examinee's profile is produced.

As an alternative for cognitive diagnosis, some researchers have pointed out that other IRT-based continuous latent models parallel the above discrete ICDMs. Contrary to the discrete ICDMs, these models place each of the underlying ability distributions on a continuum. DiBello, Roussos and Stout (2007) and Stout (2007) discussed these continuous models as possible psychometric models for cognitive diagnosis. Among these models, the application of multidimensional item response

theory (MIRT) models is common in research. For instance, *Applied Psychological Measurement* devoted the winter issue of 1996 to research in MIRT models. Instead of providing an estimate of a profile defining which attributes (or skills) have been mastered (i.e., a mastery profile), MIRT models produce factor scores. Therefore, if one were interested in determining which skills should be improved, further research must be performed to choose some factor score for each skill, at and above which the examinees are classified as masters and below which the examinees are classified as nonmasters. Consequently, if research or assessment is based on the factor scores from MIRT models, it is important to research how these conclusions about cognitive status of examinees compare to those from the ICDMs.

Both types of models, MIRT models or ICDMS, can be classified according to skill interactions into compensatory models and noncompensatory or conjunctive models. *Compensation* means that higher values on one skill can offset the lower values on other skills when calculating the probability of the correct response to an item. The extreme case of a compensatory model is the disjunctive model, which means a certain minimum on ONLY one of the relevant attributes is necessary to compensate for the lack of ability on all other skills for the correct response of the item. *Noncompensation or conjunction* means certain minimums on all skills are necessary for a high chance of a correct answer of the item. Anyone not having a minimum ability for at least one attribute will lack the ability to answer the item correctly. Having a higher ability in one attribute is NOT sufficient to compensate for the lower ability in other attribute(s) and to answer the item correctly (see Chapter II for more details).

The vast arrays of the psychometric models for cognitive diagnosis and their different ways to express cognitive complexity (e.g, underlying latent distributions, skills interaction, etc) make model selection difficult for accurate formative assessments. If the selection is to be made among models differing only in scale assumptions, this might only pose the challenge of selecting some set of some factor scores from MIRT models to evaluate the examinees cognitively. If the selection is made among models differing only in skill interactions, this might only pose the challenge of determining the type of skill interactions to provide cognitive feedback. If the selection is to be made among models differing in both scale assumptions and skill interaction (compensatory or noncompensatory), this would pose the challenge of determining the type of skill interactions for cognitive evaluation of examinees in addition to the challenge of determining a reasonable set of cut points. In the latter case, it is expected that the cognitive evaluation of examinees will be different with a noncompensatory ICDM versus a compensatory MIRT or a compensatory ICDM versus a noncompensatory MIRT.

It is always difficult to select a reasonable psychometric model because of the challenge of identifying how the skills interact with each other—across items, individuals, groups and forms. In addition, it is not always clear whether the true underlying distributions of abilities are discrete or continuous. However, if in application, final decisions based on cognitive feedback are similar even when using different models, then model selection may be based on an underlying theory without a focus on how these decisions will impact ultimate decisions for examinees. Due to the recency of the cognitive diagnosis, there has been limited research concerning the

comparison of the ICDMs and MIRT models for cognitive diagnostic purpose. Therefore, it is the research goal of this study to compare the two types of models and investigate if model selection can influence final decisions that may be made for an examinee.

For the purpose of the current study, two models with different scale assumptions and different skill interactions—one compensatory MIRT model and one noncompensatory ICDM model—were chosen (see Chapter II). The purpose of the current study is to determine how comparable the two models are with respect to the cognitive evaluation of the examinees. The two models have different assumptions about attribute scale and skill interactions. Therefore, it is necessary to identify what technique is most appropriate to compare the two different models. In chapter III, a technique is described such that the two models yield the most consistent evaluation of the examinees. Next, based on this technique, the models are compared with respect to how much the two models agree for cognitive diagnostic purposes.

To address these goals, a simulation study was performed. Three factors—test quality, the Q-matrix (Tatsuoka, 1983) structure and the correlation between the attributes—were chosen in the simulation study. However, as the ICDMs are recently developed, its relationship with MIRT models is still unclear. Therefore, a preliminary simulation study must be performed to investigate the relationship between the two models. The relationship between the two models means (1) if they define test quality in the same way and (2) what the relationship between the item parameters of the two models is. It is possible that the two models differ in their definitions of test quality, but the item parameters of the two models might be associated with each other.

Chapter III describes in detail the questions and methodologies about the initial simulation study used to establish a definition of test quality of the ICDMs and MIRT models so that these two methodologies can be fairly compared on the final research goals. Two flowcharts (Figure 3 and Figure 4) are provided to illustrate the simulation procedures. Chapter IV discusses the initial study and chapter V addresses the final research goals.

The answers to the initial study will facilitate the understanding of the relationship between the ICDMs and MIRT models, which will be used to ensure a fair comparison between the models based on test quality. The answers to the final research goal will provide information about the importance of model selection for cognitive feedback. As the demand and the need for cognitive assessment are increasing rapidly, model selection is becoming more and more crucial for formative assessment to be popular (DiBello & Stout, 2007; Bolt, 2007). If model selection does not impact the outcome related to examinees' cognitive status, it is possible for popular models to be used without affecting the results. If model selection does impact the outcome, the study is helpful to identify situations where the two models agree or disagree consistently. The results from the final research goal will also provide insight into the feasibility of using MIRT models for cognitive classification of examinees.

Chapter II provides a discussion of the ICDMs and traditional analytic models including the MIRT models. The review on different skill interaction is discussed and the comparison of the two selected models is provided. Chapter III discusses the questions, methodologies and statistics of each simulation study. Chapter IV deals

with the preliminary study and the final research goal of the study. Chapter V ends the study with conclusions and future directions.

CHAPTER II

LITERATURE REVIEW

Cognitive diagnosis, skill assessment or skill profiling refers to the partitioning the latent multidimensionality into discrete latent attributes and evaluating the examinees with respect to their status of mastery of each attribute (Hartz, Roussos & Stout, 2002). In the literature on cognition, 'attribute' is used interchangeably with 'dimension', 'factor', 'skill', 'subskill' and 'latent ability'. In this study, the ICDMs refer only to the stochastic models recently developed. All of these models assume that attributes are discrete and are discussed in detail in Section 2.1. The traditional continuous latent variable models, referred as traditional factor analytic models, are presented in Section 2.2. In both sections, conjunctive models and compensatory models are discussed. Section 2.3 includes the definitions and literature review of compensation and noncompensation. The last section presents the comparison of the selected models.

2.1 IRT-based Cognitive Diagnostic Models

IRT-based cognitive diagnostic models (ICDMs) recently developed all define the probability of correctly answering an item as a function of a set of discrete attributes measured by the item. In addition, the models require that a Q-matrix has been defined with elements q_{ik}, where 1 indicates that the k^{th} attribute is required by the i^{th} item and 0 otherwise. In most cases, the Q-matrix is assumed as fixed and is determined by content experts. In addition, most ICDMs assume that only mastery of

those attributes specified by the Q-matrix is necessary for the correct responses. These ICDMs can be classified according to skill interaction into noncompensatory or conjunctive and compensatory models. The conjunctive models are presented first and the compensatory models are presented next.

Conjunctive Models

Reparameterized Unified Model (RUM, Hartz et al, 2002, also referred to as the Fusion model) was defined based on the Unified Model (DiBello, Stout & Roussos, 1995). The Unified Model is among the first cognitive models to acknowledge that the Q-matrix is an incomplete representation of all the cognitive requirements for the test, thus differentiating the Unified Model from most early cognitive diagnosis models. Specifically, the Unified model includes $P_{C_i}(\theta_j)$, where θ_j is a single continuous ability parameter as a unidimensional projection of examinee *j*'s relevant attributes outside those defined in the Q matrix (using a Rasch model with different parameters—c_i). The problem with the Unified Model is that it is not estimable because there are $2k_i+3$ parameters (k = the number of attributes required by the item) for each item *i* and thus, the parameters are not identifiable.

Hartz (2002) developed the RUM (Fusion Model) out of the Unified Model. She reparameterized the Unified model so that it was estimable and she retained the interpretability of the parameters. The reparameterized model has $2+K_i$ parameters per item, where K_i represents the total number of required attributes for an item. The R-RUM defines the probability of a correct response $P(X_{ij}=1/\alpha_j,\theta_j)$ as:

$$P(X_{ij} = 1/\alpha_j, \theta_j) = [\pi_i^* \prod_{k=1}^{K} r_{ik}^{*(1-\alpha_{jk})q_{ik}}] P_{c_i}(\theta_j) \quad (2.1)$$

where $\pi_i^* = \prod_{k=1}^{K} \pi_{ik}^{q_{ik}}$

=*P*(correctly applying all item *i* required attributes given q_{ik}=1 for all item required attributes), which is the probability of giving a correct answer to all the attributes given that an examinee *j* is a master of all the traits (*k=1,…K*) related to item *i*.

$$r_{ik}^* = \frac{P(Y_{ijk} = 1/\alpha_{jk} = 0)}{P(Y_{ijk} = 1/\alpha_{jk} = 1)}$$

which is interpretable as item *i* discrimination parameter for attribute *k* or the penalty for not mastering attribute *k*

c_i = the amount that correct item performance requires θ_j, in addition to the required Q attributes; referred to as the completeness index for item *i*.

The ranges of the parameters are $0 \leq \pi_i^* \leq 1$, $0 \leq r_{ik}^* \leq 1$, $0 < c_i < 3$. For the discrimination parameter, r_{ik}^* is 1 when the item does not require the k^{th} attribute and 0 when the discrimination is maximum. The additional ability, θ_j, is assumed to be continuous, ranging from $-\infty$ to $+\infty$. As the value of θ_j approaches infinity, $P_{C_l}(\theta_j)$ approaches to 1 for all values of c_i. When the value of c_i is approximately 0, the different values of $P_{C_l}(\theta_j)$ will influence the item response function. The estimation of the RUM was solved using a Markov Chain Monte Carlo (MCMC) algorithm and a stepwise parameter selection procedure.

The RUM is among the most common ICDMs studied (e.g, Jang, 2005). Hartz (2002) applied the model to PSAT/NMQT for the purpose of improving students' performance on SAT. Jang (2005) also applied the RUM comprehensively to ETS-TOEFL standardized testing. Jang constructed the Q-matrix by combining the characteristics of the items with the results from DIMTEST and DETECT. The insignificant item parameters were eliminated and the program for the RUM was rerun on the data, using the modified Q-matrix. The follow-up study, surveys and interviews, was conducted on a sample of 28 students and two teachers, to cross-validate the diagnostic reports. Roussos, Hartz and Stout (2003) applied the RUM to the math section of American College Testing's assessment.

The Reduced RUM (R-RUM, Hartz et al, 2002, Henson & Douglas, 2005; Fu, 2005) The R-RUM is a simplified version of the RUM with the additional ability, θ_j, removed. With the non-Q attributes ($P_{C_l}(\theta_j)$) removed, it is implicitly acknowledged that the Q-matrix is a complete representation of the skills required for the test or the non-Q attributes are insignificant. The interpretations of the remaining parameters are the same as in the RUM and thus the probability of a correct response is defined as:

$$P(X_{ij} = 1 / \alpha_j) = \pi_i^* \prod_{k=1}^{K} r_{ik}^{(1-\alpha_{jk})q_{ik}} \qquad (2.2)$$

Henson & Douglas (2005) applied this model in the study on the ICDM test discrimination indices.

The NIDA Model (noisy inputs, deterministic "and" gate, Junker and Sijstma, 2001; Maris, 1999) In the NIDA model, the probability of a correct response is:

$$P(X_{ij}=1/\alpha_j, \mathbf{s}, \mathbf{g}) = \prod_{k=1}^{K}[(1-s_k)^{\alpha_{jk}} g_k^{1-\alpha_{jk}}]^{q_{ik}} \qquad (2.3)$$

Where $s_k = P(\eta_{ijk}=0/\alpha_{jk}=1, q_{ik}=1)$, a slipping parameter

$g_k = P(\eta_{ijk}=1/\alpha_{jk}=0, q_{ik}=1)$, a guessing parameter

η_{ijk}, a latent variable defined at attribute level, with 1 indicating the examinee *j* has correctly applied attribute *k* on item *i* and 0 otherwise.

The NIDA model predicts the probability of giving a correct response as the product of slipping and guessing parameters. In the model, s_k is an error probability that an examinee incorrectly applies attribute *k* when in fact, he or she is a master of that attribute and g_k is the probability that an examinee correctly applies attribute *k* when he or she is a non-master of that attribute. Because the slipping and guessing parameters are defined at the attribute level, only the Q-matrix distinguishes difference among items and no item specific parameters are defined. Maris (1999) gives another version of the NIDA model with the parameters estimated for each item and so the probability of a correct response is defined as:

$$P(X_{ij}=1/\alpha_j, \mathbf{s}, \mathbf{g}) = \prod_{k=1}^{K}[(1-s_{ik})^{\alpha_{jk}} g_{ik}^{1-\alpha_{jk}}]^{q_{ik}} \qquad (2.4)$$

However, like the Unified Model, this model is not identified.

de la Torre and Douglas (2004) applied the NIDA model for assessing the skills used in mixed number subtraction. Based on the content and the problem-solving characteristics of the 20-item test, they identified an eight-skill Q matrix for fraction subtraction.

The DINA Model (deterministic inputs, noisy "and" gate, Junker & Sijstma, 2001; Macready & Dayton, 1977; Haertel, 1989). The DINA model defines the probability of a correct response as a function of two probabilities based on whether the examinee has mastered the required attributes for the i^{th} item. Specifically,

$$P(X_{ij} = 1/\xi_{ij}, s_j, g_j) = (1 - s_i)^{\xi_{ij}} g_i^{(1-\xi_{ij})} \quad (2.5)$$

Where $\xi_{ij} = \prod_{ki=1}^{K} \alpha_{jk}^{q_{ik}}$, which is an indicator of whether examinee *j* has mastered all the required attributes for item *i*, with 1 indicating the mastery of all of the item's required attributes and 0 nonmastery of at least one attribute;

$s_i = P(X_{ij} = 0 / \xi_{ij} = 1)$, a slipping parameter; defining the probability that the examinee *j*, a master of all traits, incorrectly responds to the item.

$g_i = P(X_{ij} = 1/\xi_{ij} = 0)$, a guessing parameter, meaning that a nonmaster of at least one attribute, 'guesses' and correctly responds to the item.

The DINA model constrains $(1 - s_i)$ to be greater than g_i. The model simplifies examinees into two groups—masters and non-masters. In the non-master group, the examinees missing one attribute are equivalent to those missing all the attributes.

Zhang (2006) applied the DINA model for differential item functioning (DIF) study. In the study, Zhang manipulated the item parameters for the different groups and completed a DIF analysis on simulated data and using real data. In addition to the NIDA model, de la Torre and Douglas (2004) also applied the DINA model for the cognitive diagnosis of the skills used in mixed number subtraction. Recently, based on

real data, de la Torre and Lee (2007) used the DINA model to explore the relationship between the ICDMs, classical testing theory and IRT indices.

Compensatory Models

In the following section, examples of compensatory models are introduced. They include the compensatory RUM (Hartz, 2002), NIDO (Templin, Henson, Douglas, 2006) and a disjunctive model—DINO model (Templin & Henson, 2006). As defined in the previous chapter, a disjunctive model is an extreme case of the compensatory model in the sense that the competency on ONLY one skill is enough for the correct answer of the item. Last are the LCDM (Henson, Templin, & Willse, 2008) and the GDM (von Davier, 2005), the two general versions of compensatory and noncompensatory model as was shown by Henson, Templin and Willse (2008) through their introduction of the log-linear cognitive diagnostic model (LCDM).

Compensatory RUM (Hartz, 2002). The compensatory RUM is a compensatory version of the R-RUM, where the probability of a correct response is defined as:

$$P(X = 1/\beta_i, a, q_i, \gamma_i) = \frac{\exp[\beta_i + \sum_{k=1}^{K} \gamma_{ik} q_{ik} a_{jk}]}{1 + \exp[\beta_i + \sum_{k=1}^{K} \gamma_{ik} q_{ik} a_{jk}]} \qquad (2.6)$$

where β_i = the intercept parameter interpreted as the baseline log-odds of getting the item correct for examinees not mastering the skill.

γ_{ik} =the increased log-odds of getting the item correct for each mastered Q-matrix indicated skill

Therefore, for those who are nonmasters of all the Q-matrix specified attributes, the probability of the correct response is a function of the intercept parameter. This

model was later defined as a special case of the generalized diagnostic model (GDM, to be discussed, von Davier, 2005) and was applied to TOEFL test (von Davier, 2005).

The NIDO Model (noise input deterministic 'or' gate, Templin, Henson &Douglas, 2006) Based on NIDA model, Templin, Henson and Douglas (2006) developed a compensatory model so that the probability of a correct response:

$$P(X_{ij}=1/\alpha_j, q_{ik}) = \frac{\exp[\sum_{k=1}^{K}(\beta_k + \gamma_k \alpha_{jk})q_{ik}]}{1+\exp[\sum_{k=1}^{K}(\beta_k + \gamma_k \alpha_{jk})q_{ik}]} \quad (2.7)$$

where β_k = the threshold of getting the skill correct for examinees not mastering the skill;

γ_k = the skill level discrimination parameter

Notice that the NIDO model defines the probability of a correct response using only two parameters per skill. Like the NIDA model, this model does not have parameters at the item level and so the item parameters will have identical values within the same skill. As a result, the probability of getting the item correct will be identical for items with an identical Q-matrix entry.

The DINO Model (deterministic input noise 'or' gate, Templin & Henson, 2006) Based on the DINA model, Templin and Henson (2006) developed a disjunctive model. Similar to the notation ξ_{ij} in the DINA model, the notation ω_{ij} is used to divide examinees into two groups: those who have mastered at least one attribute of the Q-matrix (ω_{ij}=1) and those who have not mastered any Q-matrix specified entries (ω_{ij}=0) for the i^{th} item. Specifically:

$$\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{jk})^{q_{ik}} \quad (2.8)$$

Incorporating this notation into the DINA model, the conjunctive model now becomes a disjunctive model, predicting the probability of a correct response as a function of the slip and guessing parameters:

$$P(X_{ij} = 1/\omega_{ij}) = (1 - s_i)^{\omega_{ij}} g_i^{(1-\omega_{ij})} \quad (2.9)$$

where $(1 - s_i) > g_i$. Templin and Henson (2006) applied the DINO model to evaluate and diagnose the pathological gamblers.

The Log-linear Cognitive Diagnostic Model (LCDM, Henson, Templin & Willse, 2008) The LCDM is a flexible log-linear model that can fit many of the noncompensatory or compensatory models discussed above. First, give a general model when the number of attributes is 2 (*K*=2). The LCDM predicts the probability of correct response as:

$$P(X_{ij} = 1/\alpha) = \frac{\exp(\gamma_{i1}\alpha_1 + \gamma_{i2}\alpha_2 + \gamma_{i12}\alpha_1\alpha_2 - \beta_i)}{1 + \exp(\gamma_{i1}\alpha_1 + \gamma_{i2}\alpha_2 + \gamma_{i12}\alpha_1\alpha_2 - \beta_i)} \quad (2.10)$$

where γ_{i12} represents skill interactions with a value greater than 0 indicating the noncompensation and 0 or less indicating compensation.

γ_{ik} is the discrimination parameter for each attribute related to item *i*.

β_i is the intercept parameter interpreted as the probability of a correct response for those who are nonmasters of the required skills.

Notice this is a model for dichotomous data. Using examples, Henson, Templin and Willse (2008) demonstrated how the LCDM could fit compensatory RUM, DINA,

DINO and reduced RUM. Perhaps more importantly, the LCDM provides a parameterization for assessing the differences between each model and thus can be used to identify a reduced model such as the models previously described. The authors also performed MCMC estimations on a real dataset. The results from the LCDM estimation indicated that some items were consistent with the DINA, one item was consistent with the DINO and some items were consistent with compensatory RUM.

The Generalized Diagnosis Model (GDM, von Davier, 2005) The GDM is a general and flexible version of the ICDMs. The GDM can provide parameter estimates for multiple item types (dichotomous and ordered responses) with multiple latent ability types (either dichotomous or approximately continuous). With the GDM, the Q-matrix entries can be either dichotomous or polytomous skills. Within the class of the GDM, both compensatory and noncompensatory ICDMs may be specified (Henson et al, 2008). The GDM predicts the probability of correct responses by:

$$P(X = x / \beta_i, a, q_i, \gamma_i) = \frac{\exp[\beta_{xi} + \gamma_{xi.}^T h(q_{ik}, a_{jk})]}{1 + \sum_{y=1}^{m_i} \exp[\beta_{yi} + \gamma_{yi.}^T h(q_{ik}, a_{jk})]} \qquad (2.11)$$

where $h(q_i, a) = (h_1(q_i, a), ..., h_k(q_i, a))$ is a vector of functions

$\gamma_{xi} = (\gamma_{xi1}, ..., \gamma_{xiK})$, ($2^k$-1) dimensional slope parameters to determine the contribution of each non-zero Q-matrix entry.

β_{xi}, the real-valued difficulty parameters

When $h(q_i, a) = \alpha_{ik} \times q_{jk}$, the compensatory RUM is a special case of the GDM. With the exception of the RUM, all the above ICDMs can be modeled with the GDM

(Henson et al, 2008). However, the GDM can approximate the RUM (Henson et al, 2008). Notice when *k* in equation 2.11 is 1 and α_j is defined as a continuous latent variable with normal distribution, the GDM is an expression for the two-parameter logistic IRT model.

The GDM was applied to both the simulated data and the real data (von Davier, 2005). For the simulated data, the classification accuracy across four skills using Cohen's kappa was above .85 across five different replications. The application was done on TOEFL Internet-based testing pilot data with two forms (Form A and B) and two sections (Reading and Listening). The Q-matrices were supplied by the experts. Seven out of eight skills were strongly related to the overall ability obtained using the traditional 2PL IRT model. The skill profile indicated four highly correlated skill classifications for the Listening section and the three highly correlated skill classifications for the Reading section.

The popular ICDMs in the literature have been commonly conjunctive models, such as the RUM and DINA. These ICDMs are IRT-based in the sense that they share some similarities with the IRT models in their assumptions. The ICDMs assume local independence conditional on the latent ability (i.e., α_j). Specifically, they assume that after conditioning on an examinee's abilities, the responses of an examinee to different items will not influence each other and that examinees from the same group (i.e., the same α_j) should have the same expected response pattern. In the ICDMs, monotonicity means that the probability of correctly responding to an item is non-decreasing in each coordinate of the attributes with all other coordinates held

fixed (Junker & Sijtsma, 2001).

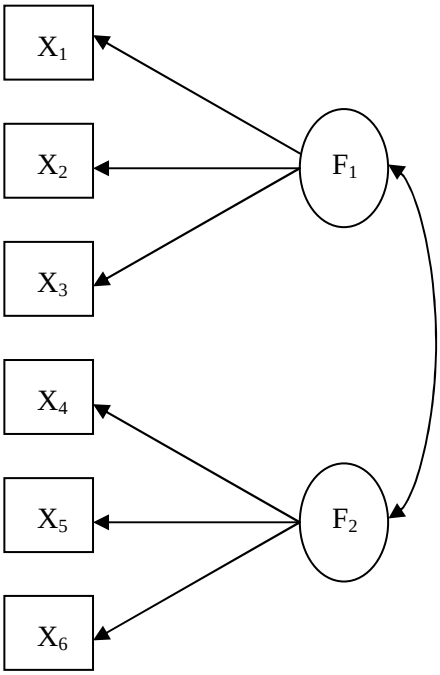
2.2 Traditional Factor Analytic Models

Linear Factor Model

Factor analysis started with Charles Spearman (1904). He proposed the one-factor theory, which assumed the test measured one general factor in common, *g*, general intelligence. He suggested that all human intellectual activities have this general factor in common. In addition, the more two tests have in common with the general factor, the higher their correlation would be. He also proposed a second factor, *the specific factor*. This factor was only specific to a single activity or variable and not correlated with the general factor. Its presence could reduce the correlation between the tests. Therefore, within a test, it is the general factor, a factor universal to a person's ability, that accounts for the correlation among the items.

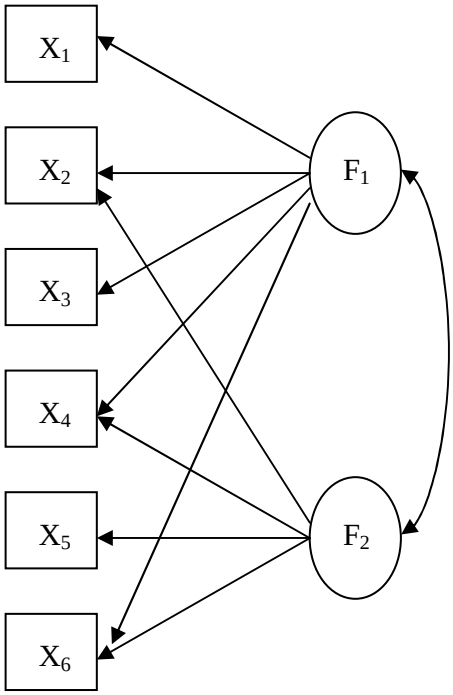
Some researchers did not agree with the one-factor model. Thurstone (1938) is one of the famous proponents of the multiple factors. Analyzing the responses from 240 volunteer students on fifty-six tests, he identified nine independent factors. Later, Thurstone (1941) completed a second study and found the same factors present. It was Thurstone who put forward the concept of 'simple structure', a very important concept in factor analysis. *Simple structure* describes a test where each item loads on only one dimension. Graphically, simple structure can be represented as follows:

Figure 1. Simple Structure



As opposed to simple structure, a test is factorially complex when a measured variable is related to more than one factor or an item is measured by more than one factor (refer to Figure 2).

Figure 2. Factorially Complex Structure



Generally, for each person, the factor model may be expressed:

$$x_i = \mu_i + \Lambda_{ik} f_k + \varepsilon_i \qquad (2.12)$$

In this model, $\boldsymbol{x_i}$ is a column vector of the measured variable i, or responses to items. The constant $\boldsymbol{\mu_i}$ represents the i^{th} item's difficulty. $\boldsymbol{\Lambda}$ is a ($i \times k$) matrix of factor loadings, representing the amount of information that each item contains about each factor k related to item i. Factor loading describes discriminating power of the item. For standardized data, factor loadings range from 0 to 1 with 1 indicating maximum discrimination and 0 indicating no relation with the factor. $\boldsymbol{f_k}$ is a column vector of latent variables and $\boldsymbol{\epsilon_i}$ is a column vector of unique factors. When $K>1$, it is a multi-factor model. When $K=1$, $\boldsymbol{\Lambda}$ is a column vector and the equation (2.12) is the

expression for classical testing theory (CTT) ($\boldsymbol{f}$ corresponds to T, unobservable true score in CTT).

Item Response Theory Models

In the above linear factor models (equation 2.12), the observed variable is predicted based on a linear combination of a set of latent variables. However, equation 2.12 is not appropriate for dichotomous item responses. When equation 2.12 is a one-factor model, the model has the following limitations. First, the assumption of linearity between the item and the latent factor cannot be met (McDonald, 1999). It is possible that equation (2.12) yields a probability less than 0 if the factor score is too small, and a probability greater than one if the factor score is large enough. Second, it assumes that error and factor are independent of each other and that the error variance is constant across all values of factors. When K in equation 2.12 is greater than 1, the linear factor model is a multiple-factor model. When applying the linear multiple-factor model to educational measurement, the same limitations associated with the linear one-factor model still exist except that each factor has its constant error variance across the values of the latent ability.

In educational measurement, one method to overcome these limitations is by using a nonlinear transformation such as is commonly used the popular IRT models. IRT models have some favorable features—such as the invariance of both item parameter estimates and ability estimates and the ability to predict the probability of the correct response for an examinee to an item given the item parameter(s). In addition, the standard error of measurement, that is the inverse of square root of information, varies across ability. The relationship between the probability of a correct

response and the latent ability is monotonic, that is, as ability increases, the probability of the correct response increases. In IRT models, the common models are either logistic models or the normal orgive models (Lord, 1952) and they differ approximately by a constant, but the logistic IRT models are more popular due to their simplicity in computation. IRT models can be classified into three-parameter (3PL) model (Birnbaum, 1968), two-parameter (2PL) model (Birnbaum, 1968) and one-parameter (1PL) model (Rasch, 1961). Because the focus of the current study is about cognitive diagnosis, only the multidimensional item response theory (MIRT) models are discussed.

Multi-dimensional IRT models The multidimensional IRT (MIRT) models predict the probability of the correct response for an item as a function of a set of item parameters as well as a vector of the given ability levels. In MIRT, there are two classes of popular models—the compensatory MIRT models (CMIRT, Reckase & McKinley, 1991) and the noncompensatory MIRT models (NCMIRT, Sympson, 1978).

Noncompensatory Multidimensional IRT (NCMIRT) Model (Sympson, 1978) Each dimension in the NCMIRT has its own difficulty parameter (d_{ik}) and its own discrimination parameter, a_{ik} , for the k^{th} trait related to item i . Higher values of the difficulty parameters indicate more difficult items and lower values indicate easy items. The multiplicative nature of the noncompensatory models prohibits an examinee from compensating for a low ability on one dimension by having a high ability on another or the other dimension(s). The most complex model of this family of NCMIRT is the 3PL NCMIRT, where the probability of a correct response is:

$$P(x_{ij} = 1 / \Theta_j, a_{ik}, d_{ik}) = c_i + (1 - c_i) \prod_{k=1}^{K} \frac{e^{(a_{ik}\theta_{jk} - d_{ik})}}{1 + e^{(a_{ik}\theta_{jk} - d_{ik})}} \qquad (2.13)$$

The 2PL NCMIRT model is a simpler version of this 3PL model with c_i constrained to zero for *i*=1,…,I. The 1PL NCMIRT model is the simplest version of equation (2.13) with the discrimination parameters constrained to unity and guessing fixed at zero.

Compensatory Multidimensional IRT (CMIRT) Model (Reckase & McKinley, 1991). Unlike the noncompensatory model, the CMIRT model has a vector of discrimination parameters, one difficulty parameter and one guessing parameter per item. The negative values of the difficulty parameter (d_i) indicate the more difficult items while the positive values suggest the easier items. Regardless of the number of dimensions, there is only one item difficulty parameter and one item guessing parameter. The 3PL CMIRT model, as is indicated, includes the discrimination parameter *a* for each skill *k* related to item *i*, a guessing parameter (c_i), and a difficulty parameter (d_i) for all dimensions. Specifically, the 3PL multidimensional logistic model is:

$$P(x_{ij} = 1 / \Theta_j, a_{ik}, d_i) = c_i + (1 - c_i) \frac{e^{\sum_{k=1}^{K} a_{ik}\theta_{jk} + d_i}}{1 + e^{\sum_{k=1}^{K} a_{ik}\theta_{jk} + d_i}} \qquad (2.14)$$

The discrimination parameters in (2.14) are constrained to be positive and the length of the item vector is equal to the amount of multidimensional discrimination (Ackerman, 1994; Reckase & McKinley, 1991). Due to the additive nature of the elements in the exponent, the examinees having a low ability on one dimension can

benefit from having a high ability on another or other dimension(s).

As to the 2PL CMIRT (Reckase, 1985), the guessing parameter is set to zero. Thus the model becomes:

$$P(x_{ij} = 1 / \Theta_j) = \frac{e^{\sum_{k=1}^{K} a_{ik}\theta_{jk} + d_i}}{1 + e^{\sum_{k=1}^{K} a_{ik}\theta_{jk} + d_i}} \qquad (2.15)$$

Note that this is equivalent to the nonlinear factor analysis with a logit link as previously described (Christoffersson, 1975; McDonald, 1967).

With the 1PL or the Rasch CMIRT model, the guessing parameters are set to zero and the discrimination parameters are constrained to unity.

These two types of the models can be rewritten as a generalized multidimensional item response theory (GMIRT) model (Ackerman & Bolt, 1995):

$$P(x_{ij} / \Theta_j, a_{ik}, b_{ik}, \mu) = \frac{e^{\sum_{k=1}^{K} f_{ijk}}}{[1 + e^{\sum_{k=1}^{K} f_{ijk}}] + \mu[\sum_{k=1}^{K} e^{\sum f_{ijk}}]} \qquad (2.16)$$

where $f_{ijk} = a_{ik}(\theta_{jk} - d_{ik})$. In equation (2.16), u is a weight with 0 representing fully compensatory model and 1 fully noncompensatory model, but any value between 0 and 1 indicates the varying degree of compensation required by the attributes. This model may be viewed as a general expression of the MIRT models and the unidimensional IRT models. In addition, a guessing parameter could be included to define a three-parameter model.

In educational measurement, the nonlinear factor model and the MIRT models, are more popular. The 1996 winter issue of *Applied Psychological Measurement* was

devoted to research of MIRT models. As shown in the next section, a large amount of research has been completed using MIRT models in educational measurement. As members of the IRT family, the relationship between MIRT models and the linear factor analysis has been established (Christoffersson, 1975). Due to its popularity, there may be some circumstances where the MIRT model would be selected to provide diagnostic information as to the ICDMs. Therefore, it is the goal of the current study to compare these two types of models to investigate how consistent the two models are with respect to cognitive diagnostic and to identify the situations where they are comparable.

2.3 Literature on Compensation and Noncompensation

The concepts of compensation and the noncompensation or conjunction was first introduced by Coombs (1964), Coombs and Kao (1955) and Johnson (1935). Under conjunctive model, the joint abilities of all attributes are necessary for answering the item correctly. Anyone lacking the ability in one attribute will lack sufficient knowledge to answer the item correctly and so will most likely miss the item. That is, having a higher ability on one attribute is NOT sufficient for compensating for the lower ability in other attribute(s) and answering the item correctly.

In contrast, compensatory models allow for a higher ability on one attribute to compensate for the lower ability on other attribute(s), thus increasing the probability of getting the item correct. Popular compensatory models include the linear factor models and some MIRT models with additive properties. Unlike equation 2.13, which is multiplicative across dimensions, equation 2.14 to equation 2.15 are additive across

the dimensions. Although additive models in the literature assume a compensatory relationship between the latent abilities and the response holds, other models, such as a disjunctive model, can also be considered compensatory. Disjunctive model require that a minimum competency on ONLY one attribute is enough for the correct answer. Apart from disjunctive model, disjunctive processing may also be represented by the negative interaction term (Henson, Templin, & Willse, 2008).

The compensatory and noncompensatory models are different from each other in the nature of cognition. The implied cognitive assumption of compensation is that the complete mastery of the Q-matrix skills is not necessary for the correct answer of the item. Instead, an ability at or above a minimum level on any of the relevant skills plays a dominant role in answering the item correctly (in the disjunctive case, it is enough to have a minimum on one skill for the correct response of the item). The cognitive assumption of noncompensation is that all the skills relevant to the item are necessary for the correct response of the item. Empirical evidence supports both types of models.

Some research found compensation outperformed noncompensation while other research found compensation and noncompensation were comparable or noncompensation was superior. For example, Simpson (2005) used the GMIRT model to investigate the relationship between noncompensatory processing and the task of matrix completion. She found u, an indicator of the degree of compensation, in the GMIRT model, was greater than 0, supporting the compensatory processing in the cognitive solution of matrix completion. Mislevy et al. (2002) found that compared with the conjunctive model, the compensatory model produced relatively high

reduction in posterior variance, indicating the compensatory model is a better fit. Comparing the compensatory model with the noncompensatory model, Van Leeuwe & Roskam (1991) found that a compensatory MIRT model provided better fit to LSAT data than a noncompensatory MIRT model.

Hambleton and Slater (1997) compared a compensatory policy with a policy combining compensatory and conjunctive components with respect to standard setting. Their results demonstrated that the compensatory policy increased the levels of decision consistency and the levels of decision accuracy whereas the policy combining both compensatory and conjunctive components lowered the levels of decision consistency and the levels of decision accuracy. Under the policy with the conjunctive components, the candidates failed at a very high rate. Consistent with Hambleton and Slater's results, Haladyna and Hess (1999) found compensatory strategies outperformed conjunctive strategies decisively in terms of reliability and rater consistency. Richter and Späth (2006), in their study of decision-making, found that people integrated information with other types of task-relevant knowledge in judgment and decision making, which was an indication of compensatory decision-making.

On the other hand, some research does find both models are comparable or support the noncompensatory model. Way, Ansley and Forsyth (1988) simulated data using both compensatory and noncompensatory models. Their independent variable was the correlation between the dimensions and the dependent variable was the ability estimates. Their results showed that the observed score distributions for each model were comparable and the θ estimates were most highly related to the average of the

two θ parameters. In a study of the success of the graduate students (Nelson, Nelson & Malone, 2000), both the compensatory term and the conjunctive term were found to be significant predictors. Investigating geometric analogy solution as a function of systematic variations in information structure of the item, Mulholland, Pellgegrino and Glaser (1980) found that the best-fitting function was a nonadditive model (a conjunctive model) instead of a simple additive model (a compensatory model). In the study of teacher licensure, Mehrens and Phillips (1989) found that the conjunctive model was more appropriate when the purpose was to set a cut-off value for the minimal competence instead of predicting the degree of success. To study Korean high school students' decision-making process, Hong & Chang (2004) conducted their study using 'think-aloud', tape-recording and observations and concluded that students preferred the non-compensatory rules instead of the compensatory rules which allowed the trade-off among alternative strategies.

With the complexity of cognition, it is impossible for one model to be the best for all scenarios. Apart from cognition, many factors might influence which type of skill interaction might occur. These factors include assessment purposes, content areas, test designs, attribute structures, or different target populations. Skill interactions might vary across items, skills, test structures, individuals, groups and populations. It is quite possible that some data might be a mixture of compensation and conjunction.

2.4 Comparison of the R-RUM and the 2PL CMIRT

A common saying may depict the dilemma of psychometricians very precisely: "A person with one watch knows what time it is; a person with two watches is never quite sure." The challenge becomes greater when there are many models available.

That is, models will have to be selected based on a compromise of model fit, the purpose of the models and some additional factors such as the assessment purpose and the way of reporting the cognitive status. However, when the measurement from two different models yields a similar interpretation, then one can make a selection based on personal preference, software availability or/and the assessment purposes. Thus, the goal of the current study is to investigate the effect of two different models on the final cognitive diagnosis of the examinees.

To make such a comparison, two models were selected—R-RUM and 2PL CMIRT model. When choosing the models, four factors were taken into consideration—model popularity, the substantive item parameter interpretations, skill interactions and attribute scales. Among the ICDMs, the conjunctive models are more commonly used such as the RUM, the R-RUM and the DINA (e.g. Hartz et al, 2002; Jang, 2005; Henson and Douglas, 2005). Among the traditional MIRT models, the CMIRT models are more often found to outperform the NCMIRT models (e.g., Bolt &.Lall, 2003; Mislevy et al, 2002). The R-RUM shares similar item parameter interpretations as the 2PL MIRT model. π_i^* in the R-RUM, ranging from 0 to 1, can be interpreted as the conditional item difficulty parameter based on Q-matrix. It is closer to d_i, item difficulty parameter in the 2PL MIRT models. In the R-RUM, r_{ik}^* is interpretable as item *i* discrimination parameter for attribute *k*, with 0 indicating the maximum discrimination and 1 indicating no discrimination. This is somewhat similar to a_{ik}, discrimination parameter in the 2PL MIRT models. The rest of the ICDMs do not share the similar item parameter interpretations with MIRT models as the R-RUM.

When selecting models for comparison, all underlying assumptions were also considered. The R-RUM is a conjunctive model and the 2PL CMIRT is a compensatory model. The R-RUM assumes the underlying distributions are discrete while the 2PL CMIRT assumes each of the distributions is on a continuum. The 2PL CMIRT and the R-RUM aggregate all different assumptions and are, therefore, chosen for the research goal. If these two models can yield a similar interpretation about the cognitive status of the examinees, then the challenge of selecting a cognitive diagnostic model can be based on whichever model the psychometricians prefer (maybe, the customers prefer), what software is available, or/and whichever model fit the assessment purposes.

However, an initial challenge must be overcome before directly comparing the R-RUM with the 2PL CMIRT model with cognitive feedback. The R-RUM is newly developed and its relationship with the traditional MIRT models is unknown. A preliminary study is necessary to address the relationship between the two models. Two questions are related to the relationship between the two models: (1) how do the two models define test quality? (2) What is the relationship between the item parameters of the two models?

In Chapter III, Figure 3 is the flowchart to address the initial challenge regarding the relationship of the two models with two specific questions. Notice that the results from the test quality of the two models will influence the comparability of these two models. Figure 4 provides the detailed simulation procedures to investigate if the two models can produce a similar interpretation of the cognitive status of the examinees. Included are also the research questions, the methods and the statistics

used in each simulation study.

CHAPTER III

METHODOLOGY

The purpose of the current study is to find out how comparable the ICDMs and the traditional MIRT models are with respect to cognitive feedback of examinees. For this purpose, the R-RUM and the 2PL CMIRT model are selected. The R-RUM is a noncompensatory model with discrete attributes and the 2PL CMIRT model is a compensatory model with continuous attributes. If these two models yield the similar results about the cognitive status of the examinees consistently across experimental conditions, then model selection can be based on the preference of the researchers or/and the clients in addition to software availability. However, unlike the R-RUM, which yields the probability of mastering each skill, the MIRT model produces continuous factor scores, and thus classification of examinees into masters and non-masters does not exist for the MIRT model. Therefore, first, a methodology is defined to identify a point, or a cut-off, for the factor scores so that examinees at or above this point are masters and examinees below this point are nonmasters. Specifically, assume that a common dataset is collected and fit by both the R-RUM and the 2PL CMIRT model. The R-RUM analysis of this data will result in estimates that can be directly used to classify examinees as a master of each attribute whereas the results from the 2PL CMIRT model for each attribute will be continuous scores for each examinee, with no direct way of determining how to transform the continuous

scores of the MIRT into dichotomous estimates of mastery/nonmastery. Therefore a method is described to determine a cutoff on the scale of the MIRT continuous abilities such that the agreement of mastery/nonmastery of the two models, when using the same dataset, is maximized.

Among the statistical tools, binomial logistic regression (thence referred as logistic regression) is used to convert the continuous values of the MIRT model to dichotomous outcomes. In logistic regression, independent variables can be interval, nominal or categorical, or a combination of all these and the dependent variable is dichotomous. Logistic regression can be used to predict the likelihood of having or not having the expected outcome given the independent variable(s). The property of logistic regression is that it is either monotonic increasing or monotonic decreasing. In the current study, the independent variable is the estimated continuous factor scores from the MIRT model and the dependent variable is the estimated mastery status (either master or nonmaster) when the R-RUM has been estimated using the same dataset. Thus, an examinee will be classified as a master on one θ when the predicted probability of the logistic regression is equal to or greater than .50. As the estimated continuous factor scores increase, the expected likelihood of being a master (i.e., the predicted probability of the dependent variable equaling 1 in the logistic regression) increases monotonically. Using logistic regression, the predicted probability for the mastery status of each skill will be obtained given each continuous factor score. Those having a predicted probability at or above .50 are classified as masters and those below .50 are classified as nonmasters. Because the cut-off values

(i.e. .50) from logistic regression yield the most consistent cognitive evaluations of examinees between the two models, they are referred as ‘optimal’.

Provided that the previously described method will be used to compare the two models, the following paragraphs provide an explanation of the conditions selected to compare them in a simulation study. Because this is a simulation-based study about how comparable the two models are with respect to cognitive feedback given to examinees, factors in this study are considered if they are expected to affect the estimation of the examinees’ profiles (either continuous or dichotomous) either directly or indirectly. Section 3.1 discusses these conditions in detail.

3.1 Experimental Conditions

As was discussed, factors of the simulation studies are selected that are expected to affect the cognitive feedback of examinees. One important factor affecting the estimation of examinees’ cognitive status is test quality. Test quality directly influences the ability of a test to accurately estimate examinees’ profile, either continuous or dichotomous. Henson and Douglas (2005) redefined the test reliability or the test quality in cognitive diagnosis to be the accuracy of classification of examinees. Item discrimination, in the cognitive diagnostic models, measures the extent that an item provides information about the classification of each attribute. Items with high discrimination are more reliable at classifying examinees as masters or nonmasters. Simulation studies (Hartz et al, 2002; Henson & Douglas, 2005) showed that test quality directly affects the correct classification rate of the examinees. A high-quality test has a higher correct classification rate. In contrast, a low-quality test has a higher misclassification rate. When test quality is low, two parallel tests will

not agree even if the true model is applied and so the agreement rate in this case must be low if two different models are compared when calibrated using the same dataset. If and only if the two models define test quality in the same way, the estimated factor scores of a master will be consistently higher than those of a nonmaster. On the contrary, if the two models define test quality differently, the implication is that one model is more reliable at classifying examinees. Therefore, comparisons cannot be made across the datasets simulated using the two different models. Comparisons can only be made on the datasets simulated using each model after running the estimation programs of the two models on the common datasets.

In this study, different test qualities—high, medium and low—are replicated. In the R-RUM, the items with high π_i^* and low r_{ik}^* are more informative about the attributes (Hartz et al, 2002; Henson, Douglas, 2005; Templin, Henson & Templin**,** 2008). To be more specific, Henson and Douglas (2005) defined high, medium and low quality tests in the R-RUM as follows:

1. High quality test: $\pi_i^* \sim (.85, .95)$ and $r_{ik}^* \sim (.10, .30)$
2. Medium quality test: $\pi_i^* \sim (.75, .95)$ and $r_{ik}^* \sim (.10, .90)$
3. Low quality test: $\pi_i^* \sim (.75, .85)$ and $r_{ik}^* \sim (.40, .90)$

In MIRT models, the test quality is related to the composite discrimination index, which is $a_c = \sqrt{\sum_{k=}^{K} a_{ik}^2}$, where a_{ik} are from equation 2.13 to equation 2.16 (Ackerman, 1994). Higher values of a_c indicate the item is good at differentiating the abilities among examinees. Following the definition of test quality in cognitive

diagnosis, a good item in MIRT models, when applied for cognitive purposes, should be more able to discriminate among examinees' continuous traits to answer an item correctly. Similarly, tests constructed with MIRT models according to the different definitions of test quality should differ in their ability at discriminating examinees along the continuous traits. For the 2PL CMIRT model, high, medium and low quality tests will be defined as (personal communication with Dr. Terry Ackerman):

1. High quality test: a_c ~ (1.30, 1.80)
2. Medium quality test: a_c ~(.70-, 1.20)
3. Low quality test: a_c ~ (.30, .70)

Table 1 summarizes the definitions of test quality of the two selected models and the definitions in this table are applicable to both simulation studies:

Table 1. Test Quality Table

Models	R-RUM		2PL CMIRT Model
Parameter / Quality	π_i^*	r_{ik}^*	a_c
High Quality	.85~.95	.10~.30	1.30~1.80
Medium Quality	.75~.95	.10~.90	.70~1.20
Low Quality	.75~.85	.40~.90	.30~.70

Next, the number of attributes per form is fixed at 4. A test can be constructed such that an item only measures one skill, which is referred to as 'simple structure' in factor analytic model (Figure 1). Alternatively, an item can be complex and measures more than one skill, which is referred to as 'factorially complex structure' (i.e., complex structure) in factor analytic model (Figure 2). In the simple structure, the

sum of each row in the Q-matrix equals to one. The sum of each row in the Q-matrix, under the complex structure, is greater than 1 and will be set between 2 and 4 in this study. The data structure is important because the effect of skill interaction on the probability of correct response is absent when the data structure is simple and so it is expected that these conditions are when the two models (the R-RUM and 2PL MIRT) would agree the most. The opposite is true when the data structure is complex. In this dissertation, both simple structure and complex structure are going to be generated:

1. Simple structure: the sum of each row is 1
2. Complex structure: the sum of each row is between 2 and 4

Last, the inter-attribute correlation is selected because inter-attribute correlation affects the dimensionality of the data structure. As the inter-attribute correlation approaches unity for all attribute pairs, the structure of the data approaches unidimensionality. The dimensionality of the data structure has potential influence on the estimation of the examinees' cognitive status. Therefore, the inter-attribute correlation is selected as the third experimental condition and the inter-attribute correlations in this study are capped at .20, .50 and .90 to replicate the possible range for correlated attributes in the real world.

In addition to the factors mentioned, the sample size for all conditions of this study is 2000 and the test length is 40. For each experimental condition, there are ten replications. Altogether, there are $3 \times 2 \times 3 \times 10$ datasets and they are replicated in both simulation study 1 and study 2.

Table 2. Experimental Conditions for Simulation Study

Data Structure / Test Quality	Simple Structure	Complex Structure
Quality Test	*r*=.20, .50, .90	*r*=.20, .50, .90
Normal Test	*r*=.20, .50, .90	*r*=.20, .50, .90
Poor Test	*r*=.20, .50, .90	*r*=.20, .50, .90

r=correlation

Notice that test quality could play a central role in that it directly impacts the ability to estimate examinees' ability. One challenge arises when comparing model performance for the R-RUM and the 2PL CMIRT because it is unknown whether the two models define test quality in the same way and whether the item parameters of the two models are related to each other. As far as this topic is concerned, research is limited. de la Torre and Lee (2007) explored the relationship between classical test theory (CTT), item response theory (IRT) and the ICDMs, using the DINA model and real data. Therefore, an initial study is completed to explore the relationship between the R-RUM and the traditional 2PL CMIRT model in terms of test quality and item parameters. There are two possible outcomes with the initial study. The most desirable outcome is that two models define test quality in the same way, i.e., same amount of reliability regarding the estimation of examinees' ability. The least desirable outcome is that they do not define test quality in the same way, meaning that one model is more reliable at estimating the examinees' cognitive profile in a nonsystematic way. Thus, as was reiterated in the section on test quality, the results of the initial study determine the methodological framework of the second simulation study. Section 3.2 gives the details for the initial study, specific questions and statistics.

3.2 Simulation Study 1: A Comparison of Test Quality and Item Parameters between the R-RUM and the CMIRT

Research Questions

As was discussed previously, test quality is central because it directly affects how reliably the abilities of examinees (either continuous or discrete) are estimated. When the two tests define test quality in the same way, the two models are 'equally' reliable with cognitive diagnosis, yielding the same amount of correct classification rate with the truth. Comparison can be made via simulating datasets separately using the two models and making a comparison across the results from the two models. Otherwise, if they define test quality differently, then the two models cannot be compared directly across the simulation conditions using two different models. Thus, comparison has to be made via simulating datasets separately with each model and estimating the examinees' profiles, both continuous and dichotomous, on the common datasets. In addition, in both circumstances, the agreement rate of the two models should be in line with test quality regardless of data structure. That is, the agreement rate is higher under high-quality test, mediocre under medium-quality test and lower under low-quality test. Therefore, the first question in simulation study 1 is: "Do the two models define test quality in the same way, i.e., are they symmetric in terms of test quality?"

Both the R-RUM and the 2PL CMIRT define test quality using discrimination parameters. The item parameter related to test quality is mostly r_{ik}^{*} in the R-RUM and a_c in the 2PL CMIRT model. Apart from test quality, it is also necessary to

explore the relationship between other item parameters of the two models. Such parameters include π_i^* versus d_i and r_{ik}^* versus a_{ik}. The question is how strongly the item parameters of the two models are related to each other? Specifically, the question is: are item parameters of one model recoverable given that the item parameters of another model are known, i.e., are they symmetric in terms of item parameters? If the item parameters of one model are recoverable, it is hypothesized that π_i^* in the R-RUM and d_i in the MRIT should be positively correlated to a high degree. On the other hand, r_{ik}^* in the R-RUM and a_{ik} in the MIRT model should be negatively correlated at a high degree. In addition, the association and the differences between the item parameters should exhibit a consistent pattern across the experimental conditions (specified in Table 2).

The recoverability of item parameters of one model using another model means that (1) one model is used to generate data (e.g., the R-RUM); (2) both models are applied to the data and the item parameters of the two models are estimated (first estimation); (3) data are generated using the second model (e.g., the 2PL CMIRT model) assuming the item parameters for the second model from the first estimation are the true parameters; (4) data generated from the previous step (step 3) are estimated using the first model (e.g., the R-RUM) (second estimation). If the item parameters of the first model are recoverable using the second model, i.e., the two models are symmetric in terms of item parameters, then the estimated item parameters for the first model from the second estimation should be associated at least moderately with the estimated item parameters of the first model from the first

estimation. The association and any differences between the two sets of estimated item parameters should also show a consistent pattern across different experimental conditions (specified in Table 2). However, if the recovered item parameters of the first model from the two estimations are only associated moderately, but the association and/or the differences between the two models do not display any consistent pattern across the conditions (specified in Table 2), the two models are only associated in terms of item parameters.

To briefly summarize the questions in the first simulation study, the question is: are the two models symmetric?

1. Are the two models symmetric in term of test quality? That is, do they define test quality in the same way?
2. Are the two models symmetric in terms of item parameters? This question is expressed in two specific questions:
    a. Are the item parameters of the two models associated with each other? Do the association of the item parameters and the differences of the item parameters show a consistent pattern across experimental conditions (specified in Table 2)?
    b. Are the item parameters of one model recoverable if another model is used?

Simulation Procedures

Figure 3 describes the procedures for data generation in study 1. First, R-RUM datasets were generated (using the program ‘CDM.EXE’ compiled in FORTRAN):

1. The first step is to generate the Q-matrices and the multivariate normal distributions.
    a. Randomly generate the test Q-matrix, 40-item exams with 4 attributes. To generate the Q-matrix for each test, a random (40 x 4) 0/1 matrix is generated such that the sum of each row is greater than 0 and less than or equal to 4 (i.e., all items must measure 1-4 attributes for the complex design and for the simple design; the total of each row was 1). The sum for each column is greater than 5 (i.e., for any given test each attribute must be measured by at least 5 items).
    b. Randomly generate four attributes, i.e., multivariate normal distributions with means of 0, standard deviations of 1 and a correlation structure of ρ. $\rho\sim$ uniform (.20, .50, .90). The sample size is 2000.
2. A cut-off value is set at 0 for the θs to dichotomize the latent distributions into the attribute patterns.
3. Randomly generate the item parameters (π_i^*, r_{ik}^*) for the R-RUM. The item parameters, π_i^*, r_{ik}^*, are simulated using random uniform distributions with lower bounds and upper bounds defined to replicate the different qualities of the test (as specified in Table 1).
4. Randomly simulate the examinees' responses using the R-RUM (equation 2.2).
5. Estimate both (a) item parameters and (b) person parameters of the MIRT model on the R-RUM datasets (using a FORTRAN program 'MIRT.EXE', to

be discussed in the section 3.4). This was a transitory step for MIRT generation.

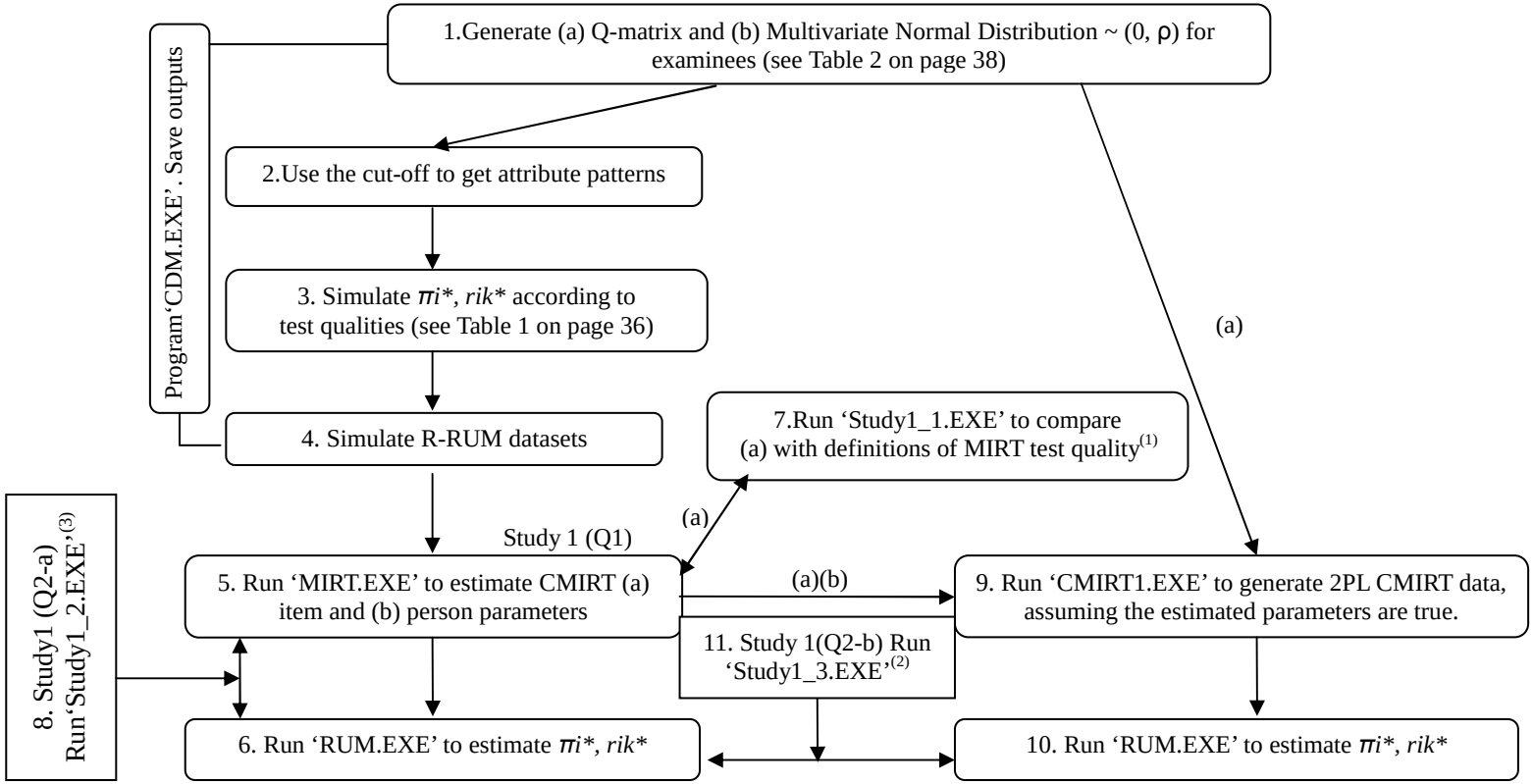
6. Estimate π_i^*, r_{ik}^* for the R-RUM datasets (using a FORTRAN program 'RUM.EXE', to be discussed in the section 3.4).
7. Obtain the maximums, minimums and averages of the estimated item parameters (including item difficulty, discrimination parameters and a_c, the composite discrimination index) from Step 5 (after running the first descriptive FORTRAN program called 'Study1_1.EXE').
8. Obtain the maximum, minimum and average differences, standard error of differences of the estimated π_i^*, r_{ik}^* from Step 6 and the estimated CMIRT model item parameters from Step 5 (after running the second descriptive FORTRAN program called 'Study 1_2.EXE'). The correlations were averaged across different datasets within each condition, assuming that the tests were measuring the same set of attributes.

Next is the 2PL CMIRT data generation (using the program 'CMIRT1.EXE' compiled in FORTRAN):

9. Randomly generate the MIRT datasets, assuming the estimated item (Step 5, a) and person parameters (Step 5, b) are the true parameters for the 2PL CMIRT model and using the same Q-matrices from 1 (a). The model used in this step of data generation is expressed in equation (2.16).
10. Estimate π_i^*, r_{ik}^* on the datasets generated in Step 9 (using the FORTRAN program 'RUM.EXE').

11. Obtain the maximum, minimum and average differences, standard error of differences of the estimated π_i^*, r_{ik}^* from Step 6 and from Step 10 (using the third descriptive FORTRAN program called 'Study 1_3. EXE'). Similarly, the correlations are averaged across different datasets within each condition, assuming that the tests are measuring the same set of attributes.

Figure 3. Flow Chart for Simulation Study 1:



Single arrows indicate transitions. Double arrows indicate comparisons. The letters in parentheses beside arrows mean the outputs with the same letters from the previous step are passed onto the next step. (1) Outputs include the grand mean, maximum and minimum, standard error for the item parameters and composite *a*. (2)(3) Outputs also include standard error of differences and averaged correlations.

Research Analyses

The reported descriptive statistics included mean, standard deviation and reliability indices for the score distributions. To examine if the test quality of one model corresponds to the respective test quality of another model, the means, minimum and maximum values and standard deviations (obtained from Step 7) were listed in tables for the estimated CMIRT item parameters (item difficulty, discrimination, and composite discrimination parameter, i.e., $a_c = \sqrt{a_1^2 + a_2^2 + a_3^2 + a_4^2}$) after running MIRT.EXE on the R-RUM datasets (Step 6 a). The results were compared with the test quality definition of the MIRT model specified in Table 1 of Section 3.1.

To investigate if the estimated item parameters of the two models were associated with each other, the following statistics were reported: minimum differences, maximum differences, standard error of the mean differences along with the average correlations between the estimated π_i^*s, r_{ik}^*s (from Step 5) and the estimated MIRT *a* and *b* parameters (from Step 6) were reported. To examine if the item parameters of one model are recoverable using another model, the reported statistics also included the grand mean differences, minimum differences, maximum differences, standard error of the mean differences along with the average correlations between the two estimated π_i^*s, r_{ik}^*s (one from Step 5 and the other from Step 10). The average correlations were calculated across the different datasets within each experimental condition, assuming that each form within each condition measured the same set of skills repeatedly. If the differences are small, standard errors are small and

the associations are at least moderate, the two models are at least associated.

3.3 Simulation Study 2: How Comparable Are the Two Models with Respect to Cognitive Feedback?

In this section, research questions related to the final goals are given first. Next, detailed simulation procedures and a flowchart are given for the second study. Because the results from the first simulation study (see section 2 of Chapter IV) showed clearly that the two models define test quality differently, simulation study 2 is performed according to Figure 4.

Research Questions

Two specific questions related to the final goal are:

1. How much do the two models agree and disagree with cognitive diagnosis of examinees?
2. What are the correct classification rates with the true attribute profiles associated with each model?

Simulation Procedure

At the beginning of the current chapter, logistic regression was identified as the appropriate technique from which the optimal cut-off values can be obtained given each estimated factor score. A program for logistic regression (called ‘Logistic.EXE’) was compiled in FORTRAN to obtain the expected likelihood of mastery for each given factor score. In addition, a number of small programs were compiled in FORTRAN for the second simulation study. ‘Alpha.EXE’ is a program compiled to dichotomize the estimated attribute profiles from the R-RUM program, ‘RUM.EXE’. Last, ‘Consistency.EXE’ was compiled in FORTRAN to cross-tabulate the agreement

rate with the estimated cognitive status of examinees of the two models and to calculate the correct classification rate of the estimated $\alpha's$ with the true $\alpha's$.

Each of the programs is listed in the specific step of the data generation procedures (see Figure 4 for the flowchart). In the left-hand part of the chart, Step 1 to Step 5 are the same as in the first simulation study. Therefore, only the remaining simulation steps are described. In the right-hand of the chart, in addition, Step 1 is the same as in the first study. Thus, the description starts with the second step.

For the R-RUM model:

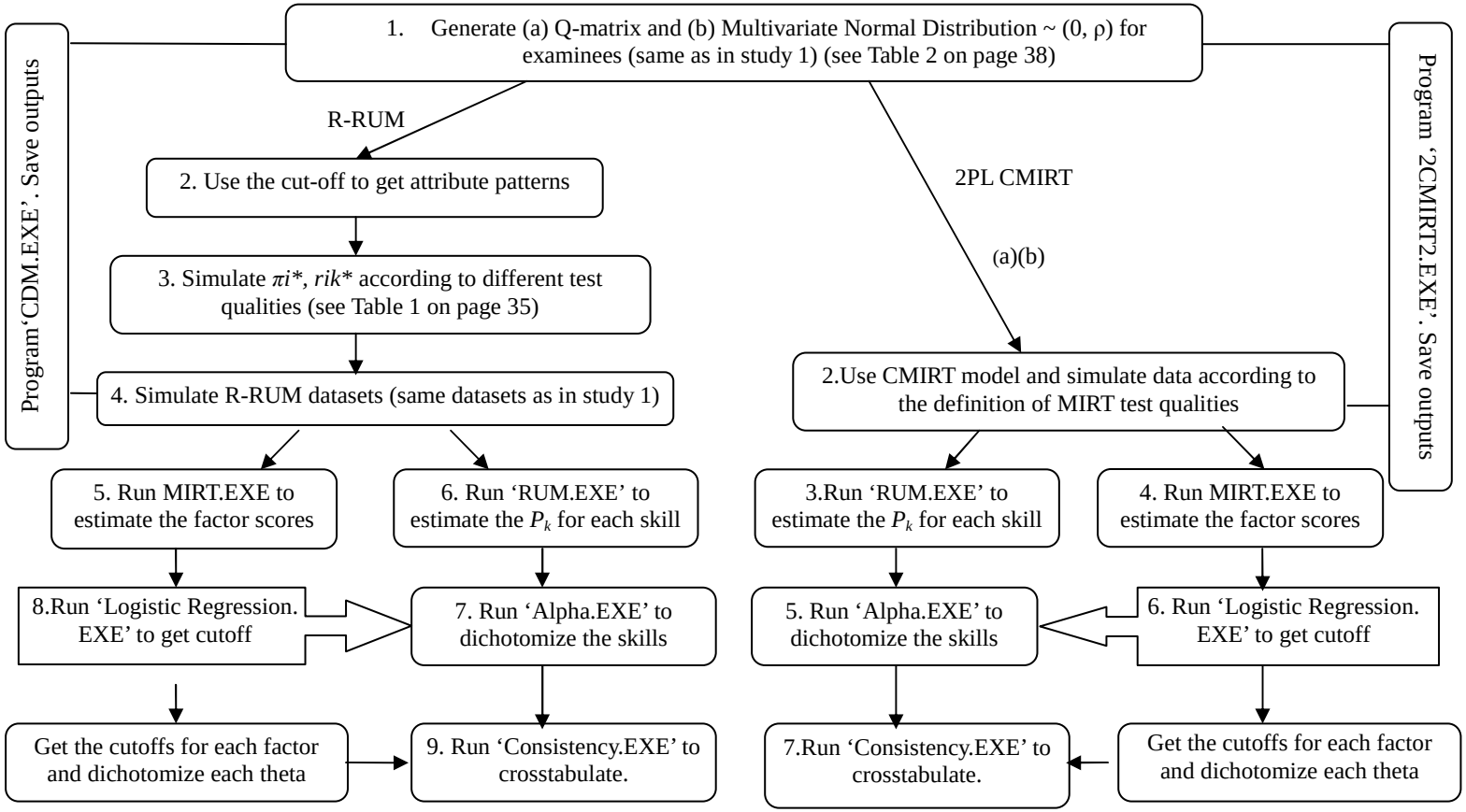
6. Estimate the probability of being a master on the R-RUM datasets (using 'RUM.EXE').
7. Dichotomize the attribute estimates from step 5 (using 'Alpha.EXE').
8. Obtain the predicted likelihood of being a master for each given factor scores (using 'Logistic.EXE').
9. Crosstabulate the agreement rates between the estimated cognitive status of examinees of the two models and the correct classification rates with the truth (using 'Consistency.EXE').

For the MIRT model:

2. Generate the MIRT dataset according to the definitions of test quality in the MIRT model (using CMIRT2.EXE, complied for the 2PL CMIRT generation in this study).
3. Estimate the probability of being a master on the MIRT datasets using the R-RUM (using 'RUM.EXE').
4. Estimate the factor scores on the MIRT datasets (using 'MIRT.EXE').

5. Dichotomize the estimated alphas from the (using ‘Alpha.EXE’).
6. Obtain the predicted likelihood of being a master (using ‘Logistic.EXE’) to obtain the cutoff point for the estimated factors.
7. Calculate the agreement rates between the estimated cognitive status of examinees of the two models and the correct classification rates with the truth (using ‘Consistency.EXE’)

Figure 4. Flowchart for Simulation Study 2



Single arrows indicate transitions. Double arrows indicate comparisons. The bigger arrow points to the dependent variables. The letters in parentheses beside arrows mean the outputs with the same letters from the previous step are passed onto the next step.

Research Analyses

The analyses were completed after obtaining the results from the above procedure. Comparison can be made within each model to examine the agreement rate and the correct classification rates with the truth for each model. Statistics included the raw agreement rate and Cohen's *Kappa*. Cohen's *Kappa* was included because the raw agreement rate is a chance-dependent statistics.

3.4 Estimation Method

In the current study, Markov chain Monte Carlo (MCMC) estimation was used to estimate the two models (R-RUM and 2PL CMIRT). MCMC has become an increasingly popular method of estimation in educational measurement for IRT models (e.g., Bolt and Lall, 2003; Bradlow, Wainer & Wang, 1999; Patz & Junker, 1999, Yao & Boughton, 2007) as well as for ICDM (e.g., Hartz et al, 2003; Henson & Douglas, 2005; Templin & Henson, 2006).

MCMC incorporates the principles of Bayesian inference by simulating random samples from a theoretical distribution, specially, the posterior distribution so that the features of the theoretical distribution can be estimated using the random samples (Patz and Junker, 1999). For measurement models, the joint posterior density for a measurement model, $f(\theta,\beta \mid X)$, can be expressed using Bayesian theorem as:

$$f(\theta,\beta \mid X) = f(X \mid \theta,\beta) * f(\theta,\beta) / \left[\int_{\theta,\beta} f(X \mid \theta,\beta) * f(\theta,\beta) d(\theta,\beta)\right] \qquad (3.1)$$

Where X represents the response data

θ denotes person parameters (either continuous or dichotomous, either unidimensional or multidimensional) in the measurement model

β denotes item parameters in the model (either a_{ik}, d_i in 2PL CMIRT or r_{ik}^*, π_i^* in R-RUM)

$f(X \mid \theta, \beta)$ is the likelihood of the item response given all the person and item parameters.

$f(\theta, \beta)$ is the prior density of the model parameters.

Note that the quantity in the denominator is the marginal distribution of the data X and this is a normalizing constant

Essentially, MCMC defines a Markov chain, M_0, M_1, M_2, …, with states $M_k = (\theta^k, \beta^k)$, where k is the total number of states. Observations (i.e., states) are sampled from the Markov chain. The way the Markov chain moves from one state to the next is determined by the transition kernel (Patz & Junker, 1999):

$$t[(\theta^0, \beta^0), (\theta^1, \beta^1)] = P[M_{K+1} = (\theta^1, \beta^1) \mid M_K = (\theta^0, \beta^0)] \tag{3.2}$$

The stationary distribution $f(\theta, \beta)$ satisfies

$$\int_{\theta,\beta} t[(\theta^0, \beta^0), (\theta^1, \beta^1)] f(\theta^0, \beta^0) d(\theta^0, \beta^0) = f\left(\theta^1, \beta^1\right) \tag{3.3}$$

Unlike maximum likelihood estimation (MLE), where the goal is to obtain point estimates of interest, sampled values under MCMC converge to distributions expressed in the left hand of (3.1) (i.e, the posterior distribution). After convergence, the initial set of draws (the burn-in) is ignored, leaving a stationary distribution, $f(\theta, \beta)$. Researchers can obtain either the averages of the posterior (expected a posteriori, EAP) or locate the maximum values (Maximum a posteriori, MAP) for the model parameters. Standard error of the posterior can also be estimated

using the standard deviation of the random draws from the Markov Chain.

In MCMC, the specification of the prior is necessary for all item and person parameters. Ideally, selected priors are conjugate priors. Conjugate priors are the priors that return posterior distributions from the same family of distributions as the prior, thus rendering MCMC more efficient. When conjugate priors are not available, it is possible to specify priors with known properties to make MCMC sampling more efficient (Kim & Bolt, 2007).

Once the priors are specified, a model is specified for the response data, the choice of sampling mechanism is an important step because the integration for the posterior is either impossible or too burdensome computationally. Two popular sampling procedures are Gibbs sampling and Metropolis-Hasting within Gibbs (MHwG).

The Gibbs sampler is a mechanism to simulate draws from the joint posterior distribution when the conditional distribution of each variable is known. For Gibbs sampling, Markov chains with transition kernels are constructed in (Geman and Geman, 1984):

$$t_G[(\theta^0,\beta^0),(\theta^1,\beta^1)] = p(\theta^1 \mid \beta^0, X)\, p(\beta^1 \mid \theta^1, X) \qquad (3.3)$$

The Gibbs sampling algorithm generates each parameter (θ^K, β^K) repeatedly with respect to its conditional distribution, conditioning on other variables. Two transition steps are taken from one state $(\theta^{k-1}, \beta^{k-1})$ to the next (θ^k, β^k):

1. Draw $\theta^k \sim p(\theta \mid X, \beta^{k-1})$;
2. Draw $\beta^k \sim p(\theta \mid X, \theta^k)$

The known conditional distributions make Gibbs sampling easy to implement and the value is always accepted ($\alpha = 1$). As is shown in the following discussion, it is a special case of Metroplis-Hasting.

The algorithm of MHwG uses a proposal distribution. It is an algorithm when samples from the complete conditionals can not be drawn according to the Gibbs algorithms. Unlike Gibbs sampling, the conditional distributions are unknown for this algorithm. Similar to Gibbs sampling, Metroplis-Hasting algorithm uses separate proposal distributions $q_\theta(\theta^0, \theta^1)$ and $q_\beta(\beta^0, \beta^1)$. After the proposal distribution is drawn, it is accepted or rejected (Patz & Junker, 1999):

1. Draw $\theta^k \sim p(\theta \mid X, \beta^{k-1})$:

   (a) Draw $\theta^* \sim q_\theta(\theta^{k-1}, \theta)$

   (b) Accept $\theta^k = \theta^*$ with probability

   $$\alpha(\theta^{k-1}, \theta^*) = \min\left\{ \frac{p(X \mid \theta^*, \beta^{k-1})\, p(\theta^*, \beta^{k-1}) q_\theta(\theta^*, \theta^{k-1})}{p(X \mid \theta^{k-1}, \beta^{k-1})\, p(\theta^{k-1}, \beta^{k-1}) q_\theta(\theta^*, \theta^{k-1})}, 1 \right\} \qquad (3.4)$$

   Otherwise, set $\theta^k = \theta^{k-1}$

2. Draw $\beta^k \sim p(\theta \mid X, \theta^k)$:

   (a) Draw $\beta^* \sim q_\beta(\beta^{k-1}, \beta)$

   (b) Accept $\beta^k = \beta^*$ with probability

   $$\alpha(\beta^{k-1}, \beta^*) = \min\left\{ \frac{p(X \mid \theta^k, \beta^*)\, p(\theta^k, \beta^*) q_\beta(\beta^*, \beta^{k-1})}{p(X \mid \theta^k, \beta^{k-1})\, p(\theta^k, \beta^{k-1}) q_\theta(\beta^*, \beta^{k-1})}, 1 \right\} \qquad (3.5)$$

   Otherwise, set $\beta^k = \beta^{k-1}$

where (θ^*, β^*) is the candidate step in the Markov chain.

The resulting Markov chain has the stationary distribution $f(\theta, \beta) = p(\theta, \beta \mid X) \propto p(X \mid \theta, \beta) p(\theta, \beta)$, indicating the joint posterior is proportional to the product of $p(X \mid \theta, \beta) p(\theta, \beta)$

It should be noted that the convergence of Markov chain is crucial. Consequently, it is important to evaluate MCMC convergence. Time-series plot is an efficient way to check the convergence of the chain. The time-series plots in the current study showed that the MCMC algorithm converged very well for all experimental conditions.

Computer Programs The two computer programs that use MCMC algorithm are RUM.EXE (Henson, 2005) and MIRT.EXE (Henson, 2006). RUM.EXE is a program compiled for the R-RUM parameter estimation. MIRT.EXE was complied in FORTRAN to estimate factor scores. Jiang (2005), in her simulation study, found that the correlations between the true and the estimated thetas were around .80 for the mixed structure (i.e., some items measured only one skill and some measured more than one) when the number of dimensions was 5 and the number of items was 45. For the same number of dimensions and items with complex structure, the FORTRAN program used in this study recovered the ability parameters quite efficiently with the average correlation being .85.

Chapter 4 contains the results for the two simulation studies and Chapter 5 discusses the results and future direction.

CHAPTER IV

RESULTS

The present chapter presents the results from (1) the symmetry of the two models and (2) the comparison of the two models with respect to cognitive feedback. The first question can be written in two parts:

1. Are the two models symmetric in term of test quality?
2. Are the two models symmetric in terms of item parameters?

The second question, which is the goal of the study, focuses on how comparable the two models are with respect to cognitive feedback. Specific questions include:

1. How much do the two models agree and disagree with cognitive diagnosis of examinees?
2. What are the correct classification rates with the truth associated with each model?

The first section contains the descriptive statistics of the datasets. The second section contains the results on the symmetry of the two models in terms of test quality and item parameters. The last section of the chapter includes the results for comparing the two models with respect of cognitive feedback of examinees.

4.1 Initial Descriptive Statistics

Table 3 contains the descriptive statistics (mean, standard deviation, reliability—KR 20) for test quality for the R-RUM. From Table 3, it is evident that test quality plays an important role in determining the magnitude of mean, standard

deviation and KR-20. The most predominant trend is that as test quality dropped, the tests became easier. For each data structure, the higher test quality, as typically defined, was associated with more difficult tests. Holding the test quality constant, tests became more variable as inter-attribute correlations increased. Holding inter-attribute correlation constant, tests with simple structure were less variable than tests with complex structure for the same test quality. Compared with complex structure, test with simple structure was easier and more homogeneous because there was limited higher-order thinking involved for each item. These indicate that test quality and data structure will have an impact on the performance of examinees. High-quality tests with complex structure are more able to discriminate among examinees, thus decreasing the variability of tests. The traditional reliability index showed that reliability decreased as test quality dropped and it increased within the same test quality as inter-attribute correlation increased because higher inter-attribute correlation creates more dependency and tests tend to measure the same thing.

Table 3. Descriptive Statistics for the R-RUM

			Mean	SD	KR20
Complex Structure	High-Quality	r=.2	14.255	10.150	.941
		r=.5	14.961	12.265	.965
		r=.9	17.786	14.981	.981
	Medium Quality	r=.2	19.511	7.644	.865
		r=.5	19.625	9.279	.915
		r=.9	20.913	11.050	.945
	Low Quality	r=.2	21.840	6.136	.767
		r=.5	21.880	7.141	.834
		r=.9	22.294	8.815	.899
Simple Structure	High-Quality	r=.2	21.568	8.900	.897
		r=.5	21.513	10.471	.933
		r=.9	21.615	12.866	.964
	Medium Quality	r=.2	25.069	5.949	.766
		r=.5	25.936	6.509	.816
		r=.9	25.622	7.849	.878
	Low Quality	r=.2	26.322	4.426	.560
		r=.5	26.408	4.877	.643
		r=.9	26.308	5.771	.752

Table 4 contains the descriptive statistics (mean, standard deviation, reliability—KR20) for the MIRT model. A similar, although different pattern, was observed in Table 4 for the traditional MIRT model. The mean did not exhibit a clear pattern, but rather it fluctuated. This can be attributed to the fact that the difficulty parameter in the MIRT model generation ranged from +3 to -3. Because of the randomness and the wider range, the threshold values might move up or down within the range for datasets, thus creating a certain amount of fluctuations among the means. Holding test quality constant, simple structure produced less variable forms than

complex structure. As in the datasets generated using R-RUM, within the same test quality, forms became more variable as the inter-attribute correlation went up. As in Table 3, KR-20 indexes also increased as the inter-attribute correlation went up within each test quality.

Table 4. Descriptive Statistics for the 2PL CMIRT Model

			Mean	SD	KR20
Complex Structure	High-Quality	r=.2	20.071	8.929	.908
		r=.5	20.116	10.509	.941
		r=.9	19.805	11.861	.958
	Medium Quality	r=.2	19.408	6.894	.835
		r=.5	19.898	8.084	.887
		r=.9	20.388	9.104	.917
	Low Quality	r=.2	20.491	4.651	.624
		r=.5	19.701	5.384	.726
		r=.9	20.181	6.251	.802
Simple Structure	High-Quality	r=.2	19.928	6.906	.831
		r=.5	20.020	8.353	.892
		r=.9	19.845	10.102	.934
	Medium Quality	r=.2	19.542	5.277	.711
		r=.5	19.699	6.250	.800
		r=.9	20.370	7.342	.862
	Low Quality	r=.2	20.092	3.889	.444
		r=.5	20.223	4.454	.589
		r=.9	20.356	4.938	.668

4.2 Symmetry of the Two Models

Are the two models symmetric in terms of test quality? The estimated $a_c's$, the item discrimination and difficulty parameters are displayed in Table 5 to Table 13. The

reported statistics include mean, minimum, maximum and standard deviation. To evaluate if the two models are symmetric in terms of test quality, the baseline comparison is set up to be the $a_c's$ in test quality definitions of the MIRT model (see Table 1). Comparisons were made between the values in the criterion table and the estimated $a_c's$ based on the R-RUM datasets. The size of composite *a*'s estimated using the R-RUM datasets for each test quality condition was much larger than their counterparts in the criterion table. When test quality was high or medium, the size of the maximum $a_c's$ was between 4 to 5, about 2.5 times or larger than their counterparts in the MIRT model definition. More importantly, the estimated mean of $a_c's$ was approximately 2.7 for high-quality test, 1.5 for the medium-quality test and .80 for the low-quality test. All these indicated that, if interpreted in a traditional way, the discrimination indices for the R-RUM were much more discriminating between masters and nonmasters than their counterparts in the traditional MIRT model. The means for item difficulty and $a_c's$ revealed that, as test quality dropped, tests became easier, thus less discriminating. Comparing complex structure with simple structure, the mean for item difficulty clearly showed that tests were harder for complex structure than for simple structure. Complex structure involves high-order thinking of more than one skill per item; therefore, it is harder.

For the complex structure, the means of $a_c's$ increased as inter-attribute correlations increased, holding test quality constant. This phenomenon was also reported by Smith (2007), who demonstrated via simulation studies and mathematical formula that when the data structure was complex, the $a_c's$ became larger as the

inter-attribute correlation increased because the inter-attribute correlation played a part in the magnitudes of the $a_c's$. The standard deviation for $a_c's$ also increased as the inter-attribute correlations increased, indicating the discriminating power of the tests is more and more variable as tests become more unidimensional. The mean and standard deviation for item difficulty decreased as test quality decreased. Following the definition of test quality in cognitive diagnosis, the above phenomena indicates that item parameters tended to be more homogeneous, i.e., less able to discriminate between masters and nonmasters as test quality dropped. That is, items do not discriminate between masters and nonmasters very well as test quality drops. Thus, it can be inferred that item difficulty in MIRT is not only correlated with π_i^*, but also with r_{ik}^* in the R-RUM.

For the simple structure, if test quality was held constant, the inverse occurred with the mean $a_c's$ (in this case, it is a_{ik}, depending on which trait the item measures), which decreased as the inter-attribute correlations increased. The only exception occurred for low quality test with high inter-attribute correlation. There were only ten replications per condition. Had more replications been performed, more phenomena due to randomness would have disappeared. Regardless of data structure, standard deviations for item difficulty and $a_c's$ of the medium-quality tests were the highest, compared with those of the high and low quality tests. It occurs because the estimated MIRT item parameters were based on the R-RUM datasets and true item parameters for the R-RUM datasets have the widest range for the medium-quality tests.

It can be concluded from the magnitudes of the means of $a_c's$ that the two

models do not have a symmetric relationship as far as test quality is concerned. Had the two models been symmetric in terms of test quality, the two models can be compared across test quality simulation conditions. Because the models are not symmetric in term of test quality, the comparison of the two models with respect of cognitive diagnosis must be made within each model after estimating the R-RUM and the 2PL CMIRT model (running both RUM.EXE and MIRT.EXE) on the common datasets. Under this scenario, the assumption is that the test is built separately using each model. However, another model is selected and the subsequent analyses are still be very informative about how much the two models agree and disagree. The next question of model symmetry is: are the two models symmetric in terms of item parameters?

Table 5. Descriptive Statistics for Test Quality Definition for High-quality Test When r=.20

Structure	Complex Structure			Simple Structure		
	Difficulty	Discrimination	Composite α	Difficulty	Discrimination	Composite α
Mean	-1.177	1.883	2.890	.381	2.704	2.704
Minimum	-2.977	.812	1.662	-.283	1.681	1.681
Maximum	1.045	5.163	5.163	1.002	5.048	5.048
Standard Deviation	1.191	.542	.487	.266	.566	.566

Table 6. Descriptive Statistics for Test Quality Definition for High-quality Test When r=.50

Structure	Complex Structure			Simple Structure		
	Difficulty	Discrimination	Composite α	Difficulty	Discrimination	Composite α
Mean	-1.192	1.923	2.897	.369	2.627	2.627
Minimum	-2.966	1.020	1.684	-.359	1.590	1.590
Maximum	.875	4.457	4.457	1.046	4.535	4.535
Standard Deviation	1.081	.497	.500	.254	.514	.514

Table 7. Descriptive Statistics for Test Quality Definition for High-quality Test When r=.90

Structure	Complex Structure			Simple Structure		
	Difficulty	Discrimination	Composite α	Difficulty	Discrimination	Composite α
Mean	-.881	2.064	3.004	.391	2.557	2.557
Minimum	-2.930	1.070	1.612	-.324	1.551	1.551
Maximum	.911	4.909	5.476	1.048	4.616	4.616
Standard Deviation	1.002	.517	.588	.269	.530	.530

Table 8. Descriptive Statistics for Test Quality Definition for Medium-quality Test When r=.20

Structure	Complex Structure			Simple Structure		
	Difficulty	Discrimination	Composite α	Difficulty	Discrimination	Composite α
Mean	-.180	.993	1.540	.673	1.464	1.464
Minimum	-2.860	.077	.191	-.631	.173	.173
Maximum	1.810	3.469	4.083	2.371	6.426	6.426
Standard Deviation	.954	.602	.684	.567	.902	.902

Table 9. Descriptive Statistics for Test Quality Definition for Medium-quality Test When r=.50

Structure	Complex Structure			Simple Structure		
	Difficulty	Discrimination	Composite α	Difficulty	Discrimination	Composite α
Mean	-.204	1.105	1.698	.795	1.431	1.431
Minimum	-2.908	.078	.252	-.621	.143	.143
Maximum	2.405	4.092	5.025	2.265	6.467	6.467
Standard Deviation	.933	.641	.787	.571	.939	.939

Table 10. Descriptive Statistics for Test Quality Definition for Medium-quality Test When r=.90

Structure	Complex Structure			Simple Structure		
	Difficulty	Discrimination	Composite α	Difficulty	Discrimination	Composite α
Mean	-.013	1.152	1.713	.743	1.371	1.371
Minimum	-2.639	.171	.289	-.605	.165	.165
Maximum	2.262	5.237	5.237	2.217	7.455	7.455
Standard Deviation	.818	.628	.812	.523	.923	.923

Table 11. Descriptive Statistics for Test Quality Definition for Low-quality Test When r=.20

Structure	Complex Structure			Simple Structure		
	Difficulty	Discrimination	Composite α	Difficulty	Discrimination	Composite α
Mean	.204	.582	.895	.740	.792	.792
Minimum	-1.237	.075	.167	.121	.114	.114
Maximum	1.319	1.489	1.569	2.012	8.311	8.311
Standard Deviation	.529	.276	.295	.278	.493	.493

Table 12. Descriptive Statistics for Test Quality Definition for Low-quality Test When r=.50

Structure	Complex Structure			Simple Structure		
	Difficulty	Discrimination	Composite α	Difficulty	Discrimination	Composite α
Mean	.206	.645	.962	.744	.756	.756
Minimum	-1.122	.121	.144	.112	.082	.082
Maximum	1.379	1.453	1.909	1.389	1.776	1.776
Standard Deviation	.505	.264	.326	.274	.346	.346

Table 13. Descriptive Statistics for Test Quality Definition for Low-quality Test When r=.90

Structure	Complex Structure			Simple Structure		
	Difficulty	Discrimination	Composite α	Difficulty	Discrimination	Composite α
Mean	.267	.705	1.061	.736	.774	.774
Minimum	-1.085	.134	.160	.094	.060	.060
Maximum	1.337	1.731	1.826	1.446	6.674	6.674
Standard Deviation	.428	.249	.344	0.278	.443	.443

Are the two models symmetric in terms of item parameters? The two questions were asked about the symmetry of the item parameters of the two models. The first question is about the estimated item parameters of the two models (both obtained on the R-RUM datasets). Are the estimated item parameters of the two models associated? Do the association and the differences of the item parameters of the two models show a consistent pattern across experimental conditions? As pointed out earlier, symmetry means that the item parameters, either between the two models or recovered using another model, are not only associated with each other, but also the patterns of association and differences are consistent across all experimental conditions.

Table 14 to Table 19 display descriptive statistics on the symmetry of two models in terms of item parameters. The reported statistics include grand mean differences, minimum differences, maximum differences, standard errors of mean difference and average correlations. However, the results in the tables (Table 14 to Table 19) demonstrated that there was no consistent pattern across the experimental conditions. First of all, the correlation between π_i^* and d_i was positive and the correlation between r_{ik}^* and a_{ik} was negative across all experimental conditions. The patterns of association and differences changed with inter-attribute correlations, data structure and test quality. For the complex structure, π_i^*'s and d_i's were weakly associated for different test qualities and different inter-attribute correlations whereas the associations between r_{ik}^* and a_{ik} were moderate or high. For the simple structure, the item parameters, either between π_i^* and d_i or between r_{ik}^* and a_{ik}, are

moderately or highly correlated.

When test structure was complex and test quality was held constant, the association between π_i^* and d_i became weaker and then stronger as inter-skill correlation increased from .20 to .90. The association between r_{ik}^* and a_{ik} became stronger and then weaker as inter-skill correlation increased from .20 to .90. Different patterns were observed for the simple structure. When data structure was simple and inter-attribute correlation was held constant, the association between π_i^* and d_i reduced as test quality dropped except when test quality was low and inter-attribute correlation was .50. There was an outlier (.10) in the association between π_i^* and d_i, the lowest correlation between π_i^* and d_i among the simple structure, thus decreasing the average correlation for this condition. On the contrary, the opposite was observed for the correlation between r_{ik}^* and a_{ik} –the association increased as test quality dropped. Comparing simple structure with complex structure, the size of correlations between r_{ik}^* and a_{ik}, that between π_i^* and d_i, was larger for simple structure than for complex structure with inter-attribute correlation and test quality held constant.

As far as the mean difference is concerned, the magnitudes of mean differences dropped as test quality dropped if the inter-attribute correlations were held constant. The declining pattern was observed both in the mean difference between r_{ik}^* and a_{ik} as well as in the mean difference between π_i^* and d_i. The mean difference between π_i^* and d_i was smaller for simple structure than for complex structure. With

inter-attribute correlation being fixed, the magnitude in the mean differences between r_{ik}^{*} and a_{ik} was larger for simple structure than for complex structure. The general pattern for mean differences indicated that the item parameters, π_i^{*} and d_i as well as r_{ik}^{*} and a_{ik}, tended to get closer as test quality dropped. It is consistent with the previous observation when the size of mean composite *a* decreased from about 2.7 to 1.5 and from about 1.5 to about .80 as test quality decreased.

As far as standard error of mean differences was concerned, the size of the standard error of mean differences was quite consistent within the same test quality. A comparison of simple structure with complex structure indicated that this statistic was larger for complex structure than for simple structure. For the complex structure, standard error of mean differences showed a systematic decrease between π_i^{*} and d_i and between r_{ik}^{*} and a_{ik} as test quality dropped. For the simple structure, standard error of mean differences associated with the differences between r_{ik}^{*} and a_{ik} showed the same systematic decrease as test quality dropped. However, in case of simple structure, standard error of mean differences associated with the differences between π_i^{*} and d_i was larger for medium-quality test.

Table 14. Descriptive Statistics for the Relation between Item Parameters of the Two Models in Case of High-quality Test, Complex Structure

	r=.20		r=.50		r=.90	
	π_i^*, d_i	r_{ik}^*, a_{ik}	π_i^*, d_i	r_{ik}^*, a_{ik}	π_i^*, d_i	r_{ik}^*, a_{ik}
Grand Mean Difference	-2.076	1.683	-2.086	1.72	-1.78	1.858
Minimum Difference	-.002	.427	-.081	.699	-.030	.690
Maximum Difference	-3.902	5.066	-3.913	4.351	-3.878	4.809
SEMD	1.187	1.257	1.079	1.252	.998	1.356
Average Correlation	.127	-.508	.061	-.570	.134	-.528

SEMD=standard error of mean difference

Table 15. Descriptive Statistics for the Relation between Item Parameters of the Two Models in Case of Medium-quality Test, Complex Structure

	r=.20		r=.50		r=.90	
	π_i^*, d_i	r_{ik}^*, a_{ik}	π_i^*, d_i	r_{ik}^*, a_{ik}	π_i^*, d_i	r_{ik}^*, a_{ik}
Grand Mean Difference	-1.024	.487	-1.049	.609	-.857	.646
Minimum Difference	-.007	-.002	-.001	.000	-.004	-.001
Maximum Difference	-3.777	3.313	-3.825	3.951	-3.578	5.129
SEMD	.945	.870	.925	.955	.806	.945
Average Correlation	.198	-.892	.165	-.904	.271	-.859

SEMD=standard error of mean difference

Table 16. Descriptive Statistics for the Relation between Item Parameters of the Two Models in Case of Low-quality Test, Complex Structure

	r=.20		r=.50		r=.90	
	π_i^*, d_i	r_{ik}^*, a_{ik}	π_i^*, d_i	r_{ik}^*, a_{ik}	π_i^*, d_i	r_{ik}^*, a_{ik}
Grand Mean Difference	-.596	-.075	-.592	-.004	-.533	.046
Minimum Difference	.001	.000	.001	.000	.001	.001
Maximum Difference	-1.992	1.067	-1.908	1.061	-1.873	1.336
SEMD	.520	.425	.500	.407	.421	.383
Average Correlation	.293	-.922	.188	-.938	.238	-.853

SEMD=standard error of mean difference

Table 17. Descriptive Statistics for the Relation between Item Parameters of the Two Models in Case of High-quality Test, Simple Structure

	r=.20		r=.50		r=.90	
	π_i^*, d_i	r_{ik}^*, a_{ik}	π_i^*, d_i	r_{ik}^*, a_{ik}	π_i^*, d_i	r_{ik}^*, a_{ik}
Grand Mean Difference	-.520	2.502	-.529	2.425	-.510	2.348
Minimum Difference	.004	1.356	.007	1.280	.003	1.227
Maximum Difference	-1.141	4.956	-1.203	4.413	-1.156	4.519
SEMD	.246	1.822	.234	1.641	.247	1.691
Average Correlation	.717	-.850	.735	-.816	.748	-.806

SEMD=standard error of mean difference

Table 18. Descriptive Statistics for the Relation between Item Parameters of the Two Models in Case of Medium-quality Test, Simple Structure

	r=.20		r=.50		r=.90	
	π_i^*, d_i	r_{ik}^*, a_{ik}	π_i^*, d_i	r_{ik}^*, a_{ik}	π_i^*, d_i	r_{ik}^*, a_{ik}
Grand Mean Difference	-.173	.984	-.061	.918	-.105	.859
Minimum Difference	.004	.010	.000	-.006	.000	.000
Maximum Difference	1.443	6.126	-1.391	6.364	-1.389	7.314
SEMD	.537	1.255	.539	1.293	.496	1.257
Average Correlation	.571	-.908	.556	-.911	.488	-.910

SEMD=standard error of mean difference

Table 19. Descriptive Statistics for the Relation between Item Parameters of the Two Models in Case of Low-quality Test, Simple Structure

	r=.20		r=.50		r=.90	
	π_i^*, d_i	r_{ik}^*, a_{ik}	π_i^*, d_i	r_{ik}^*, a_{ik}	π_i^*, d_i	r_{ik}^*, a_{ik}
Grand Mean Difference	-.060	.151	-.052	.104	-.062	.127
Minimum Difference	-.002	.001	.002	-.001	.000	.000
Maximum Difference	1.061	7.863	-.650	1.302	-.656	6.228
SEMD	.263	.606	.263	.500	.264	.568
Average Correlation	.512	-.970	.397	-.963	.509	-.952

SEMD=standard error of mean difference

The tables (Table 20 to Table 25) reported grand mean difference, minimum difference, maximum difference, standard error of mean difference and average correlation for the recoverability of item parameters. These tables showed the estimated item parameters between π^*'s and between r^*'s (obtained after running 'RUM.EXE' on both R-RUM datasets and MIRT datasets) were only moderately or lowly correlated. For both data structures, the associations between the estimated π^*'s and r^*'s were strongest at the medium quality test. Next were the associations between the estimated item parameters at low quality test. The associations between the item parameters were weakest for high quality test. The pattern of association strength can be attributed to the range of item parameters defined in the test quality. The range for the medium-quality test is widest, for the high-quality test is the narrowest. In measurement, restricting the range will restrict the correlations.

The grand mean difference became smaller for both the estimated π^*'s and the estimated r^*'s as test quality dropped. The trend was observed both with simple structure and with the complex structure. It can be explained partly by the fact that the estimated CMIRT data for this question were generated based on the estimated parameters from the R-RUM datasets. The finding is also consistent with the two previous observations. The first observation was with the decrease in the mean composite *a* as test quality dropped. The second observation was with the decrease in the grand mean difference of the estimated item parameters between the two models (first question related to the symmetry of item parameters).

The standard error of mean difference for π^* became smaller as test quality

dropped for both simple structure and complex structure, holding the inter-attribute correlation constant. An exception occurred at simple structure with .90 inter-attribute correlation. However, the difference between the standard error of mean difference for high quality test and for medium quality test was only .01, thus, it is negligible. The decrease in the standard error of mean difference happened because the size of the mean differences reduced as test quality dropped. For complex structure, the same trend was observed for the standard error of mean difference. As test quality dropped, the standard error of mean difference for π^* and for r^* became smaller.

Table 20. Descriptive Statistics for Recoverability of Item Parameters of the Two Models in Case of High-quality Test, Complex Structure

	r=.20		r=.50		r=.90	
	π^*	r^*	π^*	r^*	π^*	r^*
Grand Mean Difference	-.241	.250	-.229	.264	-.176	.261
Minimum Difference	.000	.000	.000	.000	.000	.000
Maximum Difference	-.793	.870	-.716	.884	-.711	.877
SEMD	.197	.297	.186	.307	.172	.316
Average Correlation	.134	.438	.181	.349	.237	.366

SEMD=Standard Error of Mean Difference

Table 21. Descriptive Statistics for Recoverability of Item Parameters of the Two Models in Case of Medium-quality Test, Complex Structure

	r=.20		r=.50		r=.90	
	π^*	r^*	π^*	r^*	π^*	r^*
Grand Mean Difference	-.136	.119	-.135	.144	-.124	.126
Minimum Difference	.000	.000	.000	.000	.000	.000
Maximum Difference	-.710	.835	-.568	.855	-.500	.834
SEMD	.146	.217	.125	.246	.117	.247
Average Correlation	.361	.603	.380	.601	.435	.529

SEMD=Standard Error of Mean Difference

Table 22. Descriptive Statistics for Recoverability of Item Parameters of the Two Models in Case of Low-quality Test, Complex Structure

	r=.20		r=.50		r=.90	
	π^*	r^*	π^*	r^*	π^*	r^*
Grand Mean Difference	-.133	-.033	-.101	-.013	-.094	-.011
Minimum Difference	.001	.000	.000	.000	.000	.000
Maximum Difference	-.449	-.578	-.390	-.572	-.332	-.560
SEMD	.100	.154	.094	.162	.077	.172
Average Correlation	.383	.497	.266	.485	.314	.406

SEMD=Standard Error of Mean Difference

Table 23. Descriptive Statistics for Recoverability of Item Parameters of the Two Models in Case of High-quality Test, Simple Structure

	r=.20		r=.50		r=.90	
	π^*	r^*	π^*	r^*	π^*	r^*
Grand Mean Difference	-.079	.298	-.085	.315	-.077	.278
Minimum Difference	.000	.000	.000	.000	.000	.000
Maximum Difference	-.396	.804	-.406	.777	-.407	.795
SEMD	.093	.299	.097	.301	.090	.296
Average Correlation	.277	.521	.220	.569	.317	.629

SEMD=Standard Error of Mean Difference

Table 24. Descriptive Statistics for Recoverability of Item Parameters of the Two Models in Case of Medium-quality Test, Simple Structure

	r=.20		r=.50		r=.90	
	π^*	r^*	π^*	r^*	π^*	r^*
Grand Mean Difference	-.066	.158	-.070	.114	-.070	.103
Minimum Difference	.000	.000	.000	.000	.000	.000
Maximum Difference	-.388	.797	-.400	.825	-.379	.694
SEMD	.083	.227	.091	.224	.091	.231
Average Correlation	.547	.733	.565	.681	.482	.664

SEMD=Standard Error of Mean Difference

Table 25. Descriptive Statistics for Recoverability of Item Parameters of the Two Models in Case of Low-quality Test, Simple Structure

	r=.20		r=.50		r=.90	
	π^*	r^*	π^*	r^*	π^*	r^*
Grand Mean Difference	-.048	.051	-.047	.061	-.039	.070
Minimum Difference	.000	.000	.000	.000	.000	.000
Maximum Difference	-.233	.494	-.222	.479	-.227	.501
SEMD	.064	.153	.060	.163	.063	.165
Average Correlation	.412	.634	.352	.582	.401	.573

SEMD=Standard Error of Mean Difference

The results of the first simulation study demonstrated that the two models are not symmetric, either in test quality or in item parameters. However, evidence was strong that the item parameters of the two models are only weakly associated with each other because the associations between the item parameters are very low for some experimental conditions. The next section focuses on the comparison of the two models in terms of cognitive diagnosis. Based on the results of this simulation study, the comparison with regards to final goal must be made within each model after running the programs of both models on the common datasets.

4.3 How Comparable Are the Two Models with Cognitive Feedback?

The final research goal of this study is to investigate if the two models are comparable with respect to cognitive feedback. If the two different models yield the same amount of disagreement with each other and with the true attribute patterns, the application of one model versus another does not influence the cognitive feedback. The application of one model versus another is relevant if the two models yield different amounts of agreement with each other and with the truth. For the final goal, there are two specific goals:

1. How much do the two models agree and disagree with cognitive diagnosis of examinees?
2. How much do the estimated α's for each model agree with the true α's?

How much do the two models agree and disagree? Both the raw agreement (Table 26) and *Kappa* statistic (Table 27) were reported. As *Kappa* statistic is not chance-dependent, the interpretation based on *Kappa* will be more appropriate. *Kappa* statistic showed that the agreement rates of the two models were higher in the case

when the R-RUM was used to generate the data when compared to those in the MIRT generation. The phenomenon can be attributed to the fact that the MIRT model assumes continuous distributions; therefore, the MIRT model is insensitive to the classification of examinees into either masters or nonmasters. It is also observed that, as test quality dropped, tests became less discriminating at classifying examinees into masters or nonmasters. Consequently, the agreements between the two models decreased. As the inter-attribute correlation went up, test became more unidimensional and the agreement between the two models decreased. The agreements between the two models in case of low inter-attribute correlation were higher than those in case of medium and high-attribute correlation. Simple structure outperformed complex structure across all experimental conditions so far as the agreement rates between the two models are concerned.

Table 26. Percentage of Raw Agreement between the Two Models

			R-RUM Generation	MIRT Generation
Complex Structure	High-Quality	*r*=.20	.987	.928
		r=.50	.984	.917
		r=.90	.978	.913
	Medium Quality	*r*=.20	.972	.937
		r=.50	.970	.920
		r=.90	.953	.901
	Low Quality	*r*=.20	.961	.942
		r=.50	.946	.916
		r=.90	.914	.896
Simple Structure	High-Quality	*r*=.20	.999	.984
		r=.50	.999	.984
		r=.90	.996	.932
	Medium Quality	*r*=.20	.984	.972
		r=.50	.977	.932
		r=.90	.958	.883
	Low Quality	*r*=.20	.978	.957
		r=.50	.942	.879
		r=.90	.885	.816

Table 27. *Kappa* between the Two Models

			R-RUM Generation	MIRT Generation
Complex Structure	High-Quality	*r*=.20	.975	.853
		r=.50	.967	.832
		r=.90	.955	.826
	Medium Quality	*r*=.20	.943	.872
		r=.50	.940	.839
		r=.90	.905	.801
	Low Quality	*r*=.20	.921	.876
		r=.50	.892	.827
		r=.90	.829	.791
Simple Structure	High-Quality	*r*=.20	.997	.967
		r=.50	.995	.931
		r=.90	.992	.864
	Medium Quality	*r*=.20	.967	.943
		r=.50	.954	.864
		r=.90	.917	.765
	Low Quality	*r*=.20	.955	.913
		r=.50	.884	.754
		r=.90	.770	.631

How much do the estimated α's for each model agree with the truth? Table 28 displayed the raw agreement with the true attribute profile. Table 29 showed *Kappa*-based agreement with the true attribute patterns. *Kappa* indexes showed that there was higher agreement with true attribute profile under the R-RUM generation. *Kappa* indexes indicated that fitting the MIRT model to the R-RUM data yielded higher agreement with the truth than fitting the R-RUM to the MIRT data or fitting the MIRT model to the MIRT datasets. This is because the underlying distributions of the R-RUM are discrete. When the MIRT model was fit to the R-RUM datasets, the

estimated thetas were pulled to extremes. Consequently, the agreement with the truth under the R-RUM generation was higher.

Under the MIRT generation, the agreement with the truth was lower. Unlike the R-RUM, the MIRT model assumes underlying distributions were on a continuum. In most cases, fitting the R-RUM to the MIRT data yielded higher agreement with the truth than the true model, the MIRT model. The only exception occurred at high quality test for inter-attribute correlation of .20. The continuous distribution characteristics make the MIRT model insensitive to classification purposes.

The general trend is that the R-RUM yields higher agreement with the truth across the conditions. The R-RUM is more sensitive to classification purposes. In most cases, the amount of agreement increased as inter-attribute correlations went up. Comparing complex structure with simple structure, it is obvious that simple structure recovered the true attribute profile better than complex structure. Test quality does affect the correct classification rate. The higher quality the test has, the higher agreement it produces.

In conclusion, the discrete ICDM is more appropriate for classification purposes. From the results of this study, it is evident that it does not matter which model should be selected when the true underlying distribution is dichotomized. When the assumption about discrete distributions hold, the two models yield pretty consistent results, especially when the data structure is simple. When the underlying distribution is continuous, it is still appropriate to use the cognitive diagnostic models for cognitive evaluation of examinees. If the MIRT model is applied for cognitive diagnosis, alternative ways of reporting high-thinking skills need to be considered.

Table 28. Percentage of Agreement with the True Attribute Patterns

			RUM Generation		MIRT Generation	
			Fit R-RUM	Fit MIRT	Fit R-RUM	Fit MIRT
Complex Structure	High-Quality	*r*=.20	.981	.975	.820	.829
		r=.50	.979	.971	.828	.827
		r=.90	.989	.974	.852	.822
	Medium Quality	*r*=.20	.944	.934	.777	.780
		r=.50	.957	.944	.796	.786
		r=.90	.976	.945	.808	.768
	Low Quality	*r*=.20	.862	.857	.687	.686
		r=.50	.889	.876	.729	.714
		r=.90	.941	.894	.762	.716
Simple Structure	High-Quality	*r*=.20	.996	.996	.861	.860
		r=.50	.996	.995	.867	.864
		r=.90	.997	.994	.892	.866
	Medium Quality	*r*=.20	.950	.946	.801	.799
		r=.50	.958	.951	.811	.803
		r=.90	.973	.950	.854	.805
	Low Quality	*r*=.20	.851	.850	.698	.697
		r=.50	.860	.849	.724	.705
		r=.90	.916	.863	.788	.720

Table 29. *Kappa*-based Agreement with the True Attribute Patterns

			RUM Generation		MIRT Generation	
			Fit R-RUM	Fit MIRT	Fit R-RUM	Fit MIRT
Complex Structure	High-Quality	*r*=.20	.961	.949	.655	.657
		r=.50	.958	.941	.673	.654
		r=.90	.978	.948	.712	.645
	Medium Quality	*r*=.20	.889	.868	.575	.561
		r=.50	.916	.888	.612	.572
		r=.90	.952	.889	.636	.536
	Low Quality	*r*=.20	.731	.713	.421	.372
		r=.50	.784	.751	.493	.429
		r=.90	.884	.788	.548	.433
Simple Structure	High-Quality	*r*=.20	.992	.991	.732	.720
		r=.50	.992	.99	.743	.729
		r=.90	.994	.989	.789	.733
	Medium Quality	*r*=.20	.900	.892	.619	.598
		r=.50	.917	.902	.640	.606
		r=.90	.946	.900	.719	.611
	Low Quality	*r*=.20	.713	.700	.435	.393
		r=.50	.731	.698	.480	.410
		r=.90	.834	.725	.595	.441

CHAPTER V

CONCLUSIONS AND FUTURE DIRECTIONS

This paper compared the R-RUM and the 2PL CMIRT with respect to cognitive diagnosis. The research was carried out in two separate simulation studies. The first simulation study explored the relationship between the two models—whether they are symmetric in terms of test quality and item parameters. Based on the results of the first study, the second study was performed to compare how comparable the two models are with providing examinees with cognitive information. The final chapter discusses the conclusions of the studies and possible future directions.

5.1 Conclusions

The simulation results of the first study clearly indicated that the two models define test quality in different ways and their item parameters are weakly associated. The first study provided a methodological framework within which the second study was conducted.

There are a few phenomena that are worth pointing out. First, data structure plays an important role in determining the agreement rates between the two models as well as the agreement rates of each model with the truth. Results from the second study revealed that, in case of simple structure, they agreed more consistently and yielded the highest correct classification rate. It can be attributed to the fact that each item measures only one attribute, thus eliminating the impact of skill interaction on the correct response. Obviously, when each item of the data measures only one trait, it

does not matter whether the R-RUM or the MIRT model is used. Second, the two models had higher agreement rates when the true model was the R-RUM. Therefore, when the true underlying distributions for the latent variables are dichotomous, the traditional continuous MIRT model can recover the dichotomous traits. In this case, it does not matter much which model is used for cognitive diagnosis. However, when the true underlying distributions are continuous, neither the R-RUM nor the CMIRT perform very well for classifying examinees. Recall the results from the first simulation study. The results clearly show the two models define test quality differently and if interpreted in a traditional way, the R-RUM is more reliable or discriminating as shown in Table 5 to Table 13 (see section 4.2 of Chapter IV). The different definitions of test quality determine to a certain degree that the R-RUM is better able to recover the truth. However, the true underlying distribution plays a more vital role in determining which model is better at cognitive diagnosis and when the two models agree more consistently. Third, as test quality decreased, the agreement rates between the two models decreased. Last, inter-attribute correlation played a role in the agreements rate of the two models with each other as well as with the truth. As test became more unidimensional, the agreement rates between the two models decreased. For datasets with the same test quality, the agreement rates between the estimated α's and the true α's increased as inter-attribute correlation increased, i.e., data approached unidimensionality.

5.2 Future Directions

One important finding of the current study is that the two models do not define test quality in the same way and they do not share one-to-one relationship in terms of

item parameters, but the two models are weakly associated with each other. The results and conclusions were based on the definition of test quality of the two models specified in Table 1. The range of r_{ik}^{*} for medium quality test (.10 to .90) overlaps with high quality test (.10-.30) as well as with low-quality test (.40-.90). The item parameters were generated from random uniform distribution. Because of the characteristics of uniform distribution, it was expected that one third of the parameters fell within the range for high-quality test and one third within the range for low-quality test. However, the effect of the overlapping item parameters on the results of the first simulation study is unknown. Future study is necessary to explore the topic using alternative non-overlapping definitions of test quality after verifying the definitions using simulation study.

There were only ten replications per condition for this study. Some outliers came into being as a result. Future study should include more replications with more examinees. This study only investigated the cases where the cut-off value is uniform. Further research is necessary to include situations where the cut-off value is non-uniform. Due to distributional assumption, it can be expected that the classification purposes of cognitive diagnosis will put the discrete cognitive models at advantage. Therefore, it will be more important to explore other possible ways of reporting the attribute profile when the MIRT model is used. One of the possible ways of reporting the attribute profile is to build a large examinee bank and report the percentile. It is also advisable to consult experts to determine a certain percentile or a certain factor score as a cut-off. It is also important to determine how to report the

attribute profiles. If the decision is to report the discrete profile, perhaps cognitive diagnostic model might be better. If the decision is to take full advantage of the proficiency scale, the MIRT model will be favorable. Therefore, determining how to report the attribute profile is also crucial for model selection (DeBillo, Stout, 2007).

When comparing the R-RUM and the MIRT model for cognitive diagnosis, the results in this study were optimal because cognitive information from both models was available and the cut-off point from logistic regression maximized the agreement rate between the two models. Zero was assumed to be the true cut-off point. In the real world, it is possible that only the MIRT model is used for cognitive diagnosis and zero may not be the desirable cut-off point. Under this scenario, getting a realistic cut-off point is crucial. Standard setting is highly recommended.

The current study simulated 2000 examinees and 40 items. Future study is necessary to address the effect of the number of items and examinees on the correct classification of the examinees' cognitive status (mastery versus nonmastery). The significance of this direction is that it will help to investigate the robustness of each model under the varying number of items and examinees. Thus, it will provide important feedback on which model is robust in case of small number of examinees, small number of items and combinations of both.

It might be equally important to develop some statistical indexes to test if the underlying distribution is discrete or continuous so that the selection of continuous versus discrete models is based on scientific evidence.

Another direction of research might be within the MIRT models. Ackerman and Bolt (1995) proposed the generalized MIRT (GMIRT) model. The GMIRT model may

be modified for cognitive purposes into a discrete version:

$$P(x_{ij} / \alpha_{jk}, \gamma_{ik}, b_{ik}, \mu) = \frac{e^{\sum_{k=1}^{K} f_{ijk}}}{[1 + e^{\sum_{k=1}^{K} f_{ijk}}] + \mu[\sum_{k=1}^{K} e^{f_{ijk}}]} \quad (5.1)$$

where $f_{ijk} = (\gamma_{ik} q_{ik})\alpha_{jk} - b_{ik}$. γ_{ik} represents the discrimination power of k^{th} attribute related to item i. α_{jk} is attribute profile with 1 indicating the examinee is a master and 0 otherwise. u is a weight with 0 representing compensatory model and 1 noncompensatory model, but any value between 0 and 1 indicates the varying degree of compensation required by the attributes. This is analogous to the generalized MIRT (GMIRT) model. The only difference is between θ and α, θ being continuous and α being discrete—either dichotomous or polytomous. This model belongs to item response theory model. The obvious convenience is that the weight, u, can vary across item, assessing the different degree of compensation or noncompension within the test. With this model, it is also possible to do exploratory Q-matrix analyses using NOHARM.

REFERENCES

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, 7(4), 255–278.

Ackerman, T. A.& Bolt, D.A. (1995, April) How different cognitive strategies produce differential item performance. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In : Lord, F.M., Novick, M. R. (Eds.), *Statisitical Theoryies of Mental Test Scores*. Addision-Wesley, Reading, MA.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261–280.

Bolt, D. M. and Lall, T. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using markov chain monte carlo. *Applied Psychological Measurement*, 27(6):395–414.

Bolt, D. M. (2007) The present and future of IRT-based Cognitive Diagnostic Models (ICDMs) and Related Methods. *Journal of Educational Measurement*. 44(4): 377-384.

Bradlow, E.T., Wainer, H., & Wang, X. (1999) A Bayesian random effects model for testlets, *Psychometrika*, 64, 153-168.

Christofferssn, A. (1975). Factor analysis of dichotomized variables. *Pyschometrika*, 40, 5-32

Coombs, C. (1964). *A theory of data*. New York: John Wiley.

Coombs, C. and Kao, C. (1955). Nonmetric factor analysis. *Engineering Research Bulletin*, 38.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353.

de la Torre, J., & Lee, Y. C. (2007) Relationships between Cognitive Diagnosis, CTT and IRT Indices: An Empirical Investigation

DiBello, L. V., Roussos, L.A.& Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S.Sinharay (Eds.), *Handbook of Statistics* (pp. 979-1030). Amsterdam, The Netherlands: Elsevier.

DiBello, L. V., Stout, W. F., & Roussos, L.A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-389). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

DiBello, L. V., & Stout, W. F. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement*. 44(4): 285-292

Fu, J. (2005) *A polytomous extension of the Fusion Model and its Bayesian*

parameter estimation. Unpublished doctoral dissertation. Madison, WI: University of Wisconsin-Madison.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions a Pattern Analysis and Machine Intelligence*, 6, 721-741.

Gierl, M.J., Leighton, J.P., Hunka, S.M. (2000). Exploring the logic of Tatsuoka's Rule-Space Model for test development and analysis. *An NCME Instructional Module.*

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301-321.

Haladyna, T & Hess, R. (1999). An Evaluation of Conjunctive and Compensatory Standard-Setting Strategies for Test Decisions. *Educational Assessment*, 6(2): 129-153

Hambleton, R. K. & Slater, S. C. (1997) Reliability of credentialing examinations and the impact of scoring models and standard-setting policies, *Applied Measurement in Education*, 10(1):19-38

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff

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

Hartz, S., Roussos, L., and Stout, W. (2002). Prime assessment: Skills diagnosis theory and practice. Unpublished Manuscript.

Henson, R. Douglas, J. (2005) Test construction for cognitive diagnosis, *Applied Psychological Measurement*, 29(4): 262-277.

Henson, R. (2006). MIRT.EXE: a computer program for the multidimensional item response model.

Henson, R., Templin, J., & Willse, J. (2008). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2): 190-210

Hong. J. & Chang, N. (2004). Analysis of Korean high school students' decision-making process in solving a problem involving biological knowledge. *Research in Science Education*, 34: 97-111.

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, Champaign, IL: University of Illinois.

Johnson, H. M. Some neglected principles in aptitude-testing. *American Journal of Psychology*, 1935, 47: 159-165.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258-273.

Jiang, Y (2005) *Estimating parameters for multidimensional item response theory models by MCMC methods.* Unpublished doctoral dissertation, East Lansing, MI: Michigan State University.

Lord, F.M. A theory of test scores. *Psychometric Monograph*, 1952, No. 7

McDonald, R. P. (1967). *Nonlinear factor analysis*. Psychometric Monographs, No. 15. Richmond, VA: William Byrd Press.

McDonald, R. (1997). Normal-ogive multidimensional model. In van der Linden, W. and Hambleton, R., editors, *Handbook of modern item-response theory*, pages 257–269. New York: Springer.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Education Statistics*, 33, 379-416.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64, 187-212.

Mehrens, W.A. & Phillips, S.E. (1989). Using college GPS and test scores in teacher licensure decisions: conjunctive versus compensatory models. *Applied Measurement in Education*, 2(4): 277-288.

Mislevy, R. J., Senturk, D., Almond, R., DiBello, L., Jenkins, F., Steinberg, L., and Yan, D.(2002). Modelling conditional probabilities in complex educational assessments. CSE Technical Report. Research Report No. CSE-TR-580, Center for the Study of Evaluation, CA

Mulholland, T. M., Pellegrino, J. W., and Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, 12:252–284.

Nelson, C.V, Nelson, J. S., & Malone, B. G. (2000) Admission Models for At-Risk Graduate Student in Different Academic Disciplines.

Patz, R. & Junker, B. (1999) A straightforward approach to Markov chain Monte

Carlo methods for item response models, *Journal of Educational and Behavioral Statistics*, 24: 146-178

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In: Neyman, E. (Ed.), In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4, University of California Press, Berkeley, CA

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9 (4):401–412.

Reckase, M. D. and McKinley, R. (1991). The discriminating power of test items that measure more than one ability. *Applied Psychological Measurement*, 15 (4): 361–373.

Richter, T. & Späth , P. (2006). Recognition is used as one cue among others in judgment and decision making, *Journal of Experimental Pyschology,* 32(1): 50-62

Roussos, L. A., Hartz, S. M., & Stout, W. F. (2003). *Real data applications of the Fusion Model system using an improved stepwise algorithm.* Unpublished ETS Project Report, Princeton, NJ.

Simpson, M.A. (2005). Use of a variable compensation item response model to assess the effect of working-memory load on non-compensatory processing in an inductive reasoning task. Unpublished doctoral dissertation. Greensboro, NC: University of North Carolina at Greensboro.

Smith, J. (2007). 'Intuitive' Interpretation of MIRT Model Parameters, Paper presented at 2007 NCME Annual Meeting, Chicago, IL.

Spearman, C. (1904). “General intelligence” objectively determined and measured.

American Journal of Psychology; 15, 201-298

Stout, W. (2007). Skills Diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement*, 44(4), 313-324.

Sympson, J. B. (1978). A model for testing with multidimensional items. In Weiss, D. J.,editor, *Proceedings of the 1977 computerized adaptive testing conference*, pages 82–103.

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 (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Eductional Statistics*, 12, 55-73.

Templin, J & Henson, R. (2006) Measurement of Psychological Disorders Using Cognitive Diagnosis Models, *Psychological Methods,* 11(3): 287-305

Templin, J, Henson, R., & Douglas, J (2006) General theory and estimation of cognitive diagnosis models. Using Mplus to derive model estimates. Manuscript under review.

Templin**,** J. L., Henson, R. A**.**& Templin**,** S.E. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models, Applied Psychological Measurement, 32(7): 559-574

Thurstone, L.L.(1938). Primary mental abilities. *Psychometric Monographs*, No.1

Thurstone, L.L.& Thursone, T.G (1941). Factorial Studies of Intelligence. *Psychometric Monographs*, No.2

U.S. House of Representatives (2001). *Text of No Child Left Behind*.

Van Leeuwe, J. F. J. and Roskam, E. E. (1991). The conjunctive item response model: A probabilistic extension of the Coombs and Kao model. *Methodika*, 5(14-32).

von Davier, M (2005) A General Diagnostic Model Applied to Language Testing Data, *ETS Research Report*, Princeton, NJ: 2005

Way, W. D, Ansley, T. N. & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement*, 12(3), 239-252.

Yao, L. & Boughton, K.A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, 31, 83-105.

Zhang, W. (2006) *Detecting differential item functioning using the DINA model.* Unpublished doctoral dissertation. Greensboro, NC: University of North Carolina at Greensboro.