COMBINING SCALING AND CLASSIFICATION: A PSYCHOMETRIC MODEL FOR SCALING ABILITY AND DIAGNOSING MISCONCEPTIONS

by

LAINE P. BRADSHAW

(Under the Direction of Jonathan Templin and Karen Samuelsen)

ABSTRACT

The Scaling Individuals and Classifying Misconceptions (SICM) model is presented as a combination of an item response theory (IRT) model and a diagnostic classification model (DCM). Common modeling and testing procedures utilize unidimensional IRT to provide an estimate of a student's overall ability. Recent advances in psychometrics have focused on measuring multiple dimensions to provide more detailed feedback for students, teachers, and other stakeholders. DCMs provide multidimensional feedback by using multiple categorical variables that represent skills underlying a test that students may or may not have mastered. The SICM model combines an IRT model with a DCM model that uses categorical variables that represent misconceptions instead of skills. In addition to the type of information common testing procedures provide about an examinee — an overall continuous ability, the SICM model also is able to provide multidimensional, diagnostic feedback in the form of statistical estimates of misconceptions. This additional feedback can be used by stakeholders to tailor instruction for students' needs. Results of a simulation study demonstrate that the SICM MCMC estimation algorithm yields reasonably accurate estimates under large-scale testing conditions. Results of an empirical data analysis highlight the need to address statistical considerations of the model from the onset of the assessment development process.

INDEX WORDS: diagnostic classification models, item response theory, nominal response, diagnosing student misconceptions, multidimensional

COMBINING SCALING AND CLASSIFICATION: A PSYCHOMETRIC MODEL FOR SCALING ABILITY AND DIAGNOSING MISCONCEPTIONS

by

LAINE P. BRADSHAW

B.S., University of Georgia, 2007

M. Ed., University of Georgia, 2007

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2011

COMBINING SCALING AND CLASSIFICATION: A PSYCHOMETRIC MODEL FOR SCALING ABILITY AND DIAGNOSING MISCONCEPTIONS

by

LAINE P. BRADSHAW

Major Professor: Jonathan Templin
Karen Samuelsen

Committee: Allan Cohen
Andrew Izsák

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2011

DEDICATION

To my parents, Brink and Patty Bradshaw.

ACKNOWLEDGEMENTS

I would like to express deep appreciation to my advisor, Jonathan Templin. I am thankful to learn from someone who has such a cohesive understanding of the field and a great enthusiasm for both teaching and learning. Who knew Fortran could be so exciting?! He did, and thankfully now so do I. This dissertation would not have been possible without the time he took to teach me the world of psychometrics as he understands it and to guide me in all aspects of becoming a scholar myself.

I would also like to thank Allan Cohen, Andrew Izsák, and Karen Samuelsen for the guidance and encouragement they provided as mentors. I appreciate the insights they contributed to this dissertation. Their suggestions, comments and edits greatly improved this document.

I am grateful to my family who offered constant love and support. I have confidence to “shoot the three” because they believe in me. Thank you Mama, Daddy, Granny Lane, Brinkley, and Mallory Jo. And you, too, Daddy Joe.

I would also like to thank the many other relatives, friends, and colleagues that supported me along this journey. Whether you stayed up all night finishing a project with me, fed me a home-cooked meal, or emailed me a stat comic to say you were thinking of me, thank you.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Although often considered stagnant or slow to change (Buros, 1977), over the past 30 years the field of psychometrics has undergone radical changes in the way mental traits are represented and estimated from test data. Recent advances have put multidimensional measurement models at the forefront of psychometric research because they promise detailed feedback for students, teachers, and other stake holders. Diagnostic classification models (DCMs; e.g., Rupp, Templin, & Henson, 2010) provide one approach to the measurement of multiple dimensions. DCMs use categorical latent *attributes* to represent skills underlying a test that students may or may not have mastered. Given the current reliance of testing on measuring an overall ability, DCMs, while useful for providing detailed feedback, may not fulfill all the needs of policy-driven assessment systems. The methods developed in this dissertation are designed to combine two different psychometric approaches for the purposes of garnering information about a student's overall ability and detailing the areas where a student might need more help. This chapter motivates combining these two approaches with a real-world perspective that balances the demands of educational policy and the utility of improving instruction.

The standards-based movement in education was initiated by the National Council of Teachers of Mathematics (NCTM) with the release of the *Curriculum and Evaluation Standards for School Mathematics* (1989). The use of standards then spread to other content areas, and standards are now the focus of K-12 teaching and learning. Changes in teachers' instructional and pedagogical practices were necessarily an early initiative for standards-based curriculum implementation. More recently, an emphasis has been upon creating multidimensional tests (also called assessments throughout) to complement the alignment of instruction with standards.

Stakeholders at the state, district, school and classroom levels need feedback from tests that provides information about students' performance with respect to these multidimensional performance standards. Presently, most tests provide overall measures of students' achievement levels with respect to general content areas (e.g., math or science), not at the performance standard level. The dominance of composite (or unidimensional) test scores in educational assessment can be seen as a reflection of traditions in psychometric theory. These traditions, coupled with the perception of stagnation in psychometric research and practice (Buros, 1977), seem to have influenced current educational policy centered on accountability to the point where such composite scores are the focal point of assessment and accountability.

The *No Child Left Behind Act* emphasized the need for more fine-grained information from tests that can be used to understand students' strengths and weakness with respect to performance standards (2001, Section 111 [b][3][c][xiii]). Specifically, the act mandated that state governments provide diagnostic score reports that highlight students' weaknesses. Furthermore, the Common Core Standards, a set of national standards presently adopted by 41 states, provided a further emphasis on the need for assessments that provide diagnostic information (National Research Council, 2010). In spite of the need for multidimensional diagnostic feedback, most state-level tests have been (and continue to be) built to provide an overall composite score. Consequentially, state-level tests have been used to provide such "diagnostic" information only from post-hoc analyses, often using summed subscores on sections of a test. Subscores often lack of reliability and do not provide an added-value over a total score (Haberman, Sinharay, & Puhan, 2009; Sinharay & Haberman, 2007). Haberman et al. (2009) and Sinharay and Haberman (2007) caution the use of subscores at the examinee or school level. Huff and Goodman (2007) report that at the classroom level United States teachers welcome feedback from assessments other than an overall measure of achievement. Furthermore, these results revealed teachers' desires to receive diagnostic feedback from large-scale tests at the classroom and student levels. In all levels of education, from policy and modern curriculum development to the classroom, diagnostic feedback is essential to educational progress.

The purpose of this dissertation is to develop a new psychometric model that, when coupled with feasible test design refinements, can enable more statistically optimal diagnostic score reports *in addition* to providing an overall measure of ability. The diagnostic reports reflect statistical estimates of student misconceptions. The model therefore bridges the overall measure of student ability that is predominant in educational measurement with efficient statistical measures for providing the diagnostic feedback called for by policy and stakeholders. The remainder of this chapter synthesizes the dual approaches of psychometrics: the traditional focus on unidimensional scoring and the newer focus on multidimensional methods.

Traditional Approach to the Modeling of Item Responses

The psychometric model used for test analysis drives decisions made during the test development process. Educational assessments are not typically designed to provide multidimensional feedback with respect to components of a content domain; rather they are designed to determine the overall ability of a student across that domain. In education, unidimensional item response theory models (IRT; e.g., Hambleton, Swaminathan, & Rogers,

1991) arguably represent the dominant statistical modeling approach. All models within the unidimensional IRT framework locate a student's ability along a single latent continuum representing overall ability. IRT models estimate the probability of an item response as a function of a student's ability. Items are screened in a way that those providing the most information about a student's ability are selected and those not providing as much information are dropped from a test (regardless of the content or quality of these items). When unidimensional IRT is used, items that are multidimensional (i.e., measure more than just a single ability) often show misfit because they violate the statistical assumptions underlying the unidimensional model. Because of this assumption, these items are often excluded from a test.

IRT methods were developed to extend Classical Test Theory (CTT; e.g., Crocker & Algina, 1986). IRT models shifted the focus of the analysis from the test-level (under CTT) to the item-level (under IRT). Under CTT, the total score of a test represents the estimate of a student's ability, with all items of the test contributing equally to the total score. In contrast, for IRT models other than the Rasch model, the estimate of a student's ability depends on *which* items they answer correctly as items contribute differentially to the ability estimate (i.e., measure the trait with differential precision). Thus when creating a test using IRT models, items that are strongly related to the continuous trait are said to have more information and are preferred to items less strongly related to the continuous trait. These items can better distinguish among students at different levels of the trait.

Additionally, unlike CTT scores, a student's ability estimate from a test calibrated with an IRT model is independent of the set of items answered by the student. The characteristics of the items themselves (i.e., item parameters such as difficulty and discrimination) can be equated to allow a student’s ability to be estimated with different sets of items, enabling testing programs to use different forms of a test across or within administrations (e.g., Kolen & Brennan, 2009).

The unidimensional estimates of student ability provided by IRT models allow stakeholders to rank students with respect to their composite ability. Ranking students informs decision-making processes regarding which students have a higher ability and should thus be preferentially selected for admissions, awards, or leadership and employment positions. Because composite ability is measured along a single continuum, differences in ability can be examined with respect to students, schools or districts for comparative purposes and for accountability purposes. Differences can be examined both cross-sectionally and longitudinally.

Although IRT methods provide information about a student’s ability, the score provides no diagnostic information regarding the concepts that a student has mastered or has yet to master. This single score indicates who knows *more* or who is performing at a *higher* level. The use of CTT or IRT methods to provide a single composite score where *higher* is *better* molds our language about educational performance, achievement, and success. The meaning of *higher* or *better* is implicit and often not scrutinized. For instance, Parks (2011) describes hierarchical metaphors that are dominant in discourse surrounding students’ abilities in mathematics. These metaphors “draw on a notion of children ordered in physical space, whether horizontally (‘far ahead’) or vertically (‘low’)” and are so common they are not typically regarded as metaphors in conversations (p. 85). Central to Parks’ observations is that the metaphors portray mathematics achievement as a rigid, linear path along which students can only move in one direction, an action indicated with discourse such as getting “ahead”, keeping “up”, and progressing “forward.” These descriptions mirror the types of feedback offered by traditional assessment practices and provide the lens through which ability is seen.

In light of the prevailing psychometric methodology, remembering what composes ability is important, for the components of ability are what teachers teach and students learn. In practice, mastery of different sets of these components can lead to the same IRT score, so the interpretation of a score with respect to these multiple components of ability can vary by student. Two students with the same score may have learned widely different sets of components. Because learning and instruction are component-driven, composite scores of ability play an inconsequential role in molding instruction or in improving student achievement.

In contrast, diagnostic classification models (DCMs) provide an alternative framework with which to conceptualize and assess ability. Under DCMs, ability is characterized by understanding multiple components that compose a given content domain. The focus of the analysis is shifted to identify which content components each student understands, thereby providing insight into *why* a student may not be performing well instead of identifying *which* students are performing well. Single-word summaries of overall student achievement such as high or low are not meaningful in this context. The DCM framework provides a new language to describe student achievement and to communicate feedback from assessments.

Diagnostic Approach to the Modeling of Item Responses

The purpose of a test designed to be modeled with a DCM differs greatly from one designed to be modeled with unidimensional IRT. DCMs measure a student’s mastery status on multiple skills (or components of ability) as represented by categorical latent variables referred to as *attributes*. This approach is in contrast to unidimensional IRT or CTT where the ability of a student is measured by a single continuous latent variable. Under the DCM framework test developers write items to measure these attributes. Items that are highly discriminating between mastery statuses of students are preferred. Further, test developers seek to build a collection of items that measures or captures the multidimensionality of the test. This practice is in contrast to the item screening process of test construction under IRT which commonly discards items measuring multiple dimensions. The goal of a DCM is to statistically classify students according to the combination of attributes they have mastered, so tests are developed to maximize information to correctly make these classifications (Henson & Douglas, 2005).

The utility of a student's estimated multidimensional mastery status from a DCM is that attributes students have mastered can be viewed as strengths or areas in which students do not need further instruction. Similarly, if not more importantly, attributes that students have not mastered indicate areas in which instruction or remediation should be focused. Thus, the attribute mastery status can provide feedback with respect to the more fine-grained components of a content area, which then can be used to tailor instruction to students' specific needs.

A Blended Approach to the Modeling of Item Responses

IRT models and DCMs each have distinct purposes; however, a problem in choosing between the methods arises when both types of information are needed. Each method makes an assumption with questionable tenability in exchange for desired information. Commonly used IRT models estimate a student's composite continuous ability, requiring a strong assumption of unidimensionality of the construct. If ability is thought of as lying on a continuum, then it is a simple extension to expect that multiple continuous abilities can be estimated from a single test. In practice, however, estimating multiple continuous abilities for a student under IRT is not common because it is a task requiring impractical data demands and yielding unreliable ability estimates (Templin & Bradshaw, in press). In contrast to IRT models, DCMs are able to provide reliable classifications with respect to these multiple dimensions within the construct. The trade-off of using the DCM instead of an IRT approach is that the DCM makes coarse approximations of each dimension, treating each dimension as categorical (and often dichotomous). In reality, a given construct is likely to be multidimensional (e.g., the components of mathematics as compared to overall math ability), yet each dimension likely exists to some degree, as opposed to being present or absent. In comparison, IRT makes more fine-grained measurements with respect to a coarser construct whereas DCMs make more coarse measurements with respect to more fine-grained constructs. The continuous trait estimated by IRT models is desired to scale a student's ability for comparative or accountability purposes, yet the multiple traits estimated with DCMs provide useful information in guiding decision-making with respect to instruction.

Overview of Dissertation and Chapters

This dissertation formulates, examines, and tests the performance of a nominal response psychometric model combining the functionalities of the IRT and DCM frameworks. The Scaling Individuals and Classifying Misconceptions (SICM) model measures a student's continuous ability using the correct/incorrect nature of his or her responses to multiple-choice items, while simultaneously classifying the student according to dichotomous attributes, defined as errors or misconceptions. In the SICM model, the misconceptions (attributes) are indicated by the nature of *which* incorrect alternative is selected, assuming a test is constructed with misconceptions playing into each response alternative. The SICM model yields the traditional IRT ability score, but it also provides multidimensional feedback using DCM estimates of errors.

The SICM model combines the IRT and DCM frameworks to capitalize on the salient features of each. The IRT and DCM frameworks can complement each other because each can provide a type of information that the other cannot. IRT models can describe a student's overall ability in a content area, yet cannot offer information that provides insights into what makes a student's ability high or low (i.e., where the student is lacking in understanding). IRT models identify *which* students are performing well. In contrast, DCMs are designed to measure skills or attributes composing ability but provide no information about the overall composite ability level of the student. DCMs identify *why* a student is not performing well. Coupled together through the SICM model, the IRT and DCM components can provide a more thorough description of traits affecting students' item responses on an assessment. The model describes a measure of composite ability impacting the correctness of the response and identifies distinct errors in understanding that yield specific incorrect responses, thus providing both *which* students are performing well and *why* students are not performing well.

This introductory chapter provided the rationale and motivation for developing the SICM model. Chapter 2 provides a review of existing psychometric models that embody the foundations of the proposed SICM model. Chapter 3 specifies and describes the SICM model and how it can be estimated. Chapter 4 describes the design and results of a simulation study conducted to assess the performance of the new model. Chapter 5 describes an empirical study conducted to illustrate the model's practical utility through an application to an operational assessment created for the purpose of diagnosing misconceptions.

CHAPTER 2

THEORETICAL BACKGROUND

This chapter describes relevant models from the item response theory (IRT) and diagnostic classification model (DCM) frameworks that provide the foundation for a new psychometric model that combines these two frameworks: the Scaling Individuals and Classifying Misconceptions (SICM) model. The purpose of this chapter is to explain the formulations and foundations of a set of latent variable models that are relevant to and instrumental in developing the SICM model. Therefore, this chapter serves the purposes of (a) comparing and contrasting features of these models, and (b) explaining IRT models and DCMs separately before combining them in the next chapter through the SICM model.

To acknowledge that the psychometric models discussed in this dissertation (including the SICM model) can be applied in a variety of contexts, the term *examinee* will be used from this point forward to more generally describe the individual taking the test. Although examinees commonly are students in an educational setting, examinees may also be, for example, teachers in an educational setting, or test takers from other settings, such as patients in psychology.

Latent Variable Modeling

Latent variable measurement models examine the relationship between an underlying construct or set of constructs (the latent variable(s)) and a set of measured variables (the observed variables). A common assumption in latent variable models is that differences in the quality or magnitude of the latent trait(s) are manifested in observed differences in the measured variables. By quantifying these observations, we are able to posit and test the existence and nature of the relationships between latent and measured variables via statistical models.

Although conceptually akin, IRT models and DCMS have been developed separately and thought of rather distinctly. The remainder of this chapter explains the structure and assumptions of a selection of models from the IRT and DCM frameworks. Understanding the differences in the models helps one select the appropriate model for the type of data to be analyzed and the type of conclusions to be made.

Item Response Theory Models

In education, where binary data in the form of scored responses (i.e., correct or incorrect) to test items are prevalent, a common statistical modeling approach utilizes IRT models. The IRT perspective models the linear relationship of the log-odds of the expected value of a response and a continuous latent trait. Put differently, the relationship between an underlying trait and the probability of observed item responses is modeled using a non-linear generalized linear mixed models (GLMM) approach. The parameters of the model describing the items form the fixed effects, and the parameters of the model describing the examinees form the random effects. For binary items (e.g., items scored correct or incorrect), a logistic link function relates the conditional probability an examinee answers an item correctly given their level of the latent trait (i.e., ability level) to a linear predictor or *kernel* of the GLMM.

For the one parameter logistic (1-PL) IRT model, the kernel quantifies the distance between an examinee's ability and the item's difficulty on a continuum. Specifically, the model predicts the logit (or log-odds) of a correct response to an item, as expressed by:

$$\text{logit}\,(X_{ei} = 1|\theta_e) = \theta_e - \beta_i\,, \qquad (1)$$

where the expected value of $(X_{ei} = 1|\theta_e)$ is the conditional probability examinee e provides a correct response to item *i* $(X_{ei} = 1)$ given examinee e's latent ability level (θ_e). This conditional probability is denoted $\pi_{i|\theta_e}$, and the log-odds of a correct response is equivalently expressed as $\ln\left(\frac{\pi_{i|\theta_e}}{1-\pi_{i|\theta_e}}\right)$. The parameter β_i is interpreted as a difficulty parameter. As β_i increases, the probability of a correct response decreases when holding ability constant. Thus, the greater the value of β_i, the more difficult the item is to answer correctly.

The conditional probability of a correct response in Equation (1) can be re-expressed as the one-parameter logistic (1-PL or Rasch) model (Rasch, 1960):

$$\pi_{i|\theta_e} = P(X_{ei} = 1|\theta_e) = \frac{\exp(\theta_e - \beta_i)}{1 + \exp(\theta_e - \beta_i)} \tag{2}$$

where $\pi_{i|\theta_e}$is the probability an examinee e provides a correct response ($X_{ei} = 1$) to item i given his or her ability level (θ_e), and exp (·) denotes the exponential function where $e \approx 2.718$ is raised to the power (·). When an examinee's ability is equivalent to the item's difficulty, the conditional probability of a correct response is .50.

The functional relationship between the probability an examinee provides the correct response and the latent ability is assumed to follow the shape of the item characteristic curve (ICC) or *trace line* (Lazarsfeld, 1950) as specified by the model in Equation (2). The ICC is a smooth *S*-shaped curve that displays the assumed monotonically increasing relationship between ability and the probability of a correct response. For the Rasch model this functional form assumption requires that the lower asymptote of each ICC equals 0 and that the slope of the ICC at any point along the ability continuum is equivalent for each item.

This IRT model also makes other assumptions. The response variable is assumed to follow a Bernoulli distribution and the continuous trait is assumed be normally distributed. Responses to items are assumed to be conditionally independent given an examinee's location along the ability continuum. The model relies on the assumption of unidimensionality which states that a single latent trait, as opposed to multiple latent traits, causes differences in responses to the items. These assumptions are illustrated using a conventional path diagram for a unidimensional IRT model in Figure 1. In this figure, the single-headed arrows point from a single, continuous latent variable (θ) towards each observed item (X_1, X_2, X_3, X_4) to indicate that the response to each item is due to presence of the latent variable. Responses to items are only related to each other through indirect paths through the latent variable. Because responses to items are conditionally independent, the probability of an examinee's response pattern can be expressed as the product of conditional item response probabilities:

$$P(\boldsymbol{X}_e = \boldsymbol{x}_e|\theta_e) = \prod_{i=1}^{I}(\pi_{i|\theta_e})^{x_{ei}}\,(1-\pi_{i|\theta_e})^{1-x_{ei}} \qquad (3)$$

where $\boldsymbol{x}_e$ is vector of examinee e's observed responses to all I items.

The Rasch model can be extended to additionally make the response probability a function of the item's ability to discriminate among examinees with different levels of ability. This model is referred to as the two-parameter logistic (2-PL) model (Birnbaum, 1968). Essentially the 2-PL IRT model adds a weight for the latent trait (or factor, θ). This weight is called the discrimination parameter in an IRT context and is referred to as a factor loading in other contexts. The 2-PL IRT model is expressed as:

$$P(X_{ei} = 1|\theta_e) = \pi_{i|\theta_e} = \frac{\exp(\alpha_i(\theta_e - \beta_i))}{1 + \exp(\alpha_i(\theta_e - \beta_i))}. \qquad (4)$$

The intercept, or difficulty, of the item is $-\alpha_i\beta_i$. The discrimination of the item, or the loading for ability, is α_i. The 2-PL IRT model has the same assumptions as the Rasch model, but the model allows ICCs for different items to have different slopes at given ability levels.

A Model for Nominal Responses in Item Response Theory

Variations of IRT models have been developed for analyzing nominal response data. Nominal response options consist of categories with no inherent order, in contrast to ordinal response options in which responses exist in some relative degree to each other. An example of a nominal response type of item is to ask an examinee which baseball team, given three choices: a) Atlanta Braves, b) Oakland Athletics, or c) Boston Red Sox, won the World Series in 1995. The options are unrelated categories. In contrast, if you asked an examinee to choose a category indicating how much he or she likes the Braves (e.g., not at all, somewhat, or very much), those responses have an inherent order and are thus ordinal-type responses.

Multiple-choice tests in education provide nominal response data. However, in practice nominal responses are commonly scored as correct or incorrect, and the resulting binary data are analyzed. Instead of scoring item responses, the nominal item response can be modeled directly. Dichotomizing the responses into two categories–correct or incorrect–collapses all of the incorrect alternatives into one category and fails to preserve the uniqueness of each incorrect alternative. Such a dichotomization can be viewed as an incomplete modeling of the item response (Thissen & Steinberg, 1984). If the characteristics of the incorrect alternatives present variations in the item response, then those characteristics should be modeled in the item response function (van der Linden & Hambleton, 1997). Modeling responses to the alternatives directly, in addition to providing a more complete model, can also provide a means of evaluating item alternatives in the test-development process.

A nominal response type of model provides a conditional probability that an examinee will select alternative n_{ij} among the set of J_i alternatives for item i given his or her ability level

(θ_e). A multinomial logit parameterization can be used to model this conditional probability and can be expressed as (Bock, 1972)

$$\pi_{n_{ij|\theta_e}} = P\left(X_{ei} = n_{ij}|\theta_e\right) = \frac{\exp(c_{n_{ij}} + a_{n_{ij}}(\theta_e))}{\sum_{j=1}^{J_i} \exp(c_{n_{ij}} + a_{n_{ij}}(\theta_e))}. \qquad (5)$$

The term $a_{n_{ij}}$ is an alternative-specific loading (or discrimination parameter) for θ_e, and $c_{n_{ij}}$ is an alternative-specific intercept. To identify the model either (a) the sum of the discrimination parameters and the sum of the intercept parameters across alternatives within an item must be set to equal zero, or (b) the parameters corresponding the baseline option must all be set to equal zero. For dichotomous items with only two alternatives, this nominal response IRT (NR IRT) model is equivalent to the 2-PL IRT model.

As in the 2-PL IRT model, the probability of selecting the correct response monotonically increases as θ_e increases, assuming the largest positive discrimination parameter is for the correct alternative. The probability of selecting the alternative with the largest negative discrimination parameter will monotonically decrease as θ_e increases. Thus the probability of selecting this alternative as θ_e approaches negative infinity goes to one; in turn, the probability of selecting any other alternative is zero. Thissen & Steinberg (1984) noted this feature of Bock's NR IRT model is not plausible and extended the model to yield the Multiple-choice (MC) model. This model adds a parameter to the model that represents an examinee's propensity to guess on a given item. Accounting for guessing in the MC model, however, results in a different implausible feature: the probability of selecting a correct response is not always monotonically increasing with θ_e.

Multidimensional Item Response Theory Model

The IRT models presented up to this point have been unidimensional IRT models. Although IRT is commonly used to model a unidimensional latent space, multidimensional IRT (MIRT) models estimate multiple continuous latent variables instead of one. If there are inherently multiple dimensions underlying the construct of interest, the unidimensional assumption–and, in turn, the local independence assumption–used for unidimensional IRT is violated. If an IRT model is used when the assumption of local independence is violated, parameter estimates may be biased, resulting in standard errors that are too small (Ackerman, 1992). Under these circumstances, a MIRT models are more appropriate. They provide a means to model multiple continuous dimensions and yield an estimate of an examinee's location on each dimension.

The multidimensional 2-PL IRT model, shown here measuring *f = 1,…, F* dimensions (or abilities), can be expressed as (Reckase, 1997)

$$P(X_{ei} = 1|\boldsymbol{\theta}_e) = \frac{\exp(\boldsymbol{\alpha}_i\boldsymbol{\theta}_e - \delta_i)}{1 + \exp(\boldsymbol{\alpha}_i\boldsymbol{\theta}_e - \delta_i)} \tag{6}$$

such that $\boldsymbol{\theta}_e$ represents a row vector of F ability parameters associated with each examinee. Similar to unidimensional IRT where θ_e is assumed to be normally distributed, for MIRT models, $\boldsymbol{\theta}_e$ has a multivariate normal distribution. The term $\boldsymbol{\alpha}_i$ represents a column vector of *F* discrimination parameters for item i. The parameter δ_i represents an item's threshold (-1 times an intercept). When δ_i equals the sum of the linear combination of ability parameters and item discriminations, the item has a 50% chance of being answered correctly. Because this MIRT model is compensatory, there are many ways the linear combination of $\boldsymbol{\alpha}_i\boldsymbol{\theta}_e$ can be equal to δ_i.

Equation (6) specifies the measurement component of the MIRT model: the relationships between the observed items and the latent abilities. The structural component of the model describes the relationships, or correlations, between the abilities. The distribution of ability is multivariate standard normal ($\sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$), where $\boldsymbol{\mu}$ is the vector of F ability means, each equal to zero, and $\boldsymbol{\Sigma}$ is the ability variance/covariance matrix whose diagonal entries equal one. Figure 2 illustrates a MIRT model where two dimensions are measured by a set of eight items. The figure shows an example of a MIRT model with simple structure, which means each item measures a single ability. However, MIRT models do not restrict items to only measure one dimension. The measurement components of the model are depicted by the single-headed arrows relating the latent abilities to the observed variables, and the model's structural components include the variances and covariances of the latent abilities.

Because they model multiple continuous latent variables, MIRT models provide a parallel approach to DCMs within the IRT framework. Estimating multiple abilities yields more detailed, multidimensional feedback for stakeholders, as preferred (NCLB, 2001; Huff & Goodman, 2007). However, reliably estimating multiple *continuous* abilities, the goal of a MIRT model, requires a large number of items (e.g., Templin & Bradshaw, in press). The large number of items needed may render the MIRT model difficult to use in practice because time is limited when it comes to how long an examinee can concentrate on a given test or how many days a school is willing to sacrifice instructional time for testing.

IRT Summary

The primary outcome of either uni- or multi- dimensional IRT is the location of an examinee along a continuum of a latent ability or along multiple continua for a set of abilities.

One can use this type of information to rank examinees according to the measured trait, which, among other things, may be useful in selecting among candidates for awards, scholarships, promotions, or admissions. However, examinee estimates from IRT provide limited information for other decision-making common in educational and psychological settings. For example, in education, teachers must decide whether a student has mastered a content objective or needs to show improvement with respect to the objective. These types of diagnoses shape how to differentiate instruction for students. In clinical diagnoses, decisions are required regarding whether or not a patient or client has a psychological disorder before treatment strategies can be tailored for the patient or client.

Diagnostic Classification Models

Diagnostic classification models (DCMs) can help statistically make diagnoses frequently encountered in educational and psychological settings. DCMs classify examinees according to sets of latent attributes that examinees have or have not mastered. In education, where a latent trait may be a skill or the understanding of a concept, an examinee who has the trait is said to have *mastery* of the skill or concept. DCMs can be used, for example, to classify examinees according to the set of content objectives they have mastered. In psychology, where a trait may be a characteristic of a disorder, an examinee who has the trait is said to *possess* the characteristic. DCMs can be used to classify examinees according to the set of characteristics they possess. For DCMs, latent traits are referred to as *attributes*. In this dissertation, mastery of an attribute and possession of an attribute have the same meaning.

Instead of locating an examinee along a latent continuum as in IRT, DCMs seek to place an examinee into a latent group. IRT models assume traits are continuous and focus on determining “how much” of a trait an examinee has. DCMs commonly assume traits are dichotomous and focus on determining whether or not the examinee has the trait. DCMs typically measure multiple traits, and the pattern of traits an examinee possesses determines the latent group into which an examinee is classified: each pattern of traits creates a latent group.

Because IRT estimates do not classify examinees into groups, when classifications are needed, they must be provided by subsequent analyses. For example, federally-mandated end-of-course educational tests administered by states often use IRT to provide an examinee’s ability estimate on a continuum. However, to evaluate Adequate Yearly Progress (AYP; NCLB, 2001), examinees need to be classified into proficiency categories (e.g., below basic, basic, proficient, above proficient). The score yielded by the IRT model does not provide sufficient information to make a decision about which proficiency level a student is performing, nor do IRT methods seek to optimally, or statistically, separate students into groups. Classifications of students are typically products of cut scores made along the continuum. The cut scores are established via standard setting methods that use expert human judgment to determine points along the continuum that represent boundaries of proficiency categories. Therefore, when using IRT, the examinee’s score is determined by optimal statistical methods, but the examinee’s classification is not.

As providing classifications are DCMs primary purpose, the models seek to statistically separate students into a small number of groups with respect to each latent trait. Often DCMs use two groups: masters and non-masters of a skill. However, more groups could be used, such as four groups representing the proficiency categories described above. IRT focuses on locating a person on a fine continuum instead of classifying a person into a coarse group. DCMs can more reliability place examinees into groups than IRT models can locate an examinee along a continuum (Templin & Bradshaw, in press). If classification is the purpose of the assessment, much of the information garnered from IRT is not necessary or useful for classifying examinees. DCMs may be more appropriate models to use in cases where classification is the purpose, or when multidimensional feedback is desired. DCMs sacrifice fine-grained measurement to provide multidimensional measurements. Instead of finely locating each examinee along a set of continuous traits as a MIRT model does, DCMs coarsely classify each examinee with respect to each trait. This trade-off enables DCMs to provide diagnostic, multidimensional feedback in practical testing settings. The number of items needed to reliability estimate a single dimension in an IRT model, which is common practice for large-scale educational assessments, is sufficient for estimating multiple dimensions in a DCM (Templin & Bradshaw, in press).

Because DCMs can measure more dimensions with reasonable data demands, researchers using DCMs have treated multiple components of a trait as the latent variables to estimate. For example, constituents of operating with fractions, reasoning multiplicatively with rational numbers, and exhibiting behaviors of pathological gamblers have been defined as latent variables to be estimated with DCMs (Izsák, Lobato, Orrill, Cohen, & Templin, 2009; Tatsuoka, 1990; Templin & Henson, 2006). It seems fitting that latent traits for DCMs are referred to as attributes as the term attribute connotes a part or characteristic of the whole, rather than the whole itself.

DCMs require considerable knowledge of the attributes being measured to correctly specify the model. This a priori knowledge makes DCMs confirmatory models in two respects. Prior to any statistical analysis, empirical justifications are used to (a) define and delineate the attribute or set of attributes that are measured by the assessment, and (b) specify the attribute or set of attributes that are measured by each item. The specifications in (b) are described in an item-by-attribute Q-matrix (e.g., Tatsuoka, 1990). The entries in the Q-matrix are denoted by q_{ia}, where $q_{ia} = 1$ if item i measures attribute a and $q_{ia} = 0$ otherwise. However, the Q-matrix may be specified using ordinal or continuous values if needed (Templin, 2004). Correct specification of the Q-matrix ensures the measurement component of the DCM is correct and is imperative to valid interpretations of estimated parameters (Rupp & Templin, 2007). As a result, the classifications produced by DCMs are dependent upon the successful coordination of psychometrics and empirical theories in domain-specific content areas.

Regarding diagnostic assessment, literature has focused on the intersection of cognitive sciences with psychometrics (e.g., Nichols, 1994), although DCMs are no less applicable in other areas, such as behavioral or social sciences (e.g., Templin & Henson, 2006). Connoting this generality in application contexts is the impetus for using the term *diagnostic classification model* (Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010), instead of *cognitive diagnosis model* (e.g., Leighton & Gierl, 2007), although the two terms refer to the same class of models.

Rupp and Templin (2008) describe the aforementioned characteristics of DCMs, as well as others that will be addressed in ensuing sections, in their definition of a diagnostic classification model:

> Diagnostic classification models (DCMs) are probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables. The predictor variables are combined in compensatory and noncompensatory ways to generate latent classes. DCM enable multiple criterion-referenced interpretations and associated feedback for diagnostic purposes, which is typically provided at a relatively fine-grain size. This feedback can be, but does not have to be, based on a theory of response processing grounded in applied cognitive psychology. Some DCMs are further able to handle complex sampling designs for items and examinees, as well as heterogeneity due to strategy use (p. 226).

Diagnostic Classification Models: Types of Latent Class Models

DCMs are able to classify examinees as they are a type of latent class model. As a latent class model, DCMs rely on the premise that, "there exists a set of latent classes, such that the manifest relationship between any two or more items on a test can be accounted for by the existence of these basic classes and by these classes alone" (Stouffer, 1950 as quoted in de Ayala, 2009). As described above, the latent classes or groups for DCMs are defined by patterns of attributes that examinees possess. DCMs are *multiple classification latent class models* (Maris, 1999); they determine whether or not an examinee possesses each individual attribute in order to ultimately classify an examinee by the pattern of attributes they do or do not possess (or have and have not mastered). A pattern of mastered attributes will classify examinees into one class among a set of mutually exclusive set of latent classes that are hypothesized through empirical theory to be an exhaustive set.

The possible patterns are predetermined by the specification of the attributes in the DCM. Given an assessment with A attributes, 2^A possible attribute patterns or distinct latent classes exist, where the latent class or attribute pattern is denoted as p such that $p = (1, \dots, 2^A)$. An attribute pattern p represents the knowledge state of an examinee. For examinee e, the attribute pattern is denoted by a vector of binary indicators; $\boldsymbol{\alpha}_e = [\alpha_1, \alpha_2, \dots, \alpha_A]$, where each element, or attribute α_a, is either present/mastered ($\alpha_a = 1$) or absent/not mastered ($\alpha_a = 0$). Unlike $\boldsymbol{\theta}_e$ in the MIRT model which has a multivariate normal distribution, an examinee's attribute pattern $\boldsymbol{\alpha}_e$ has a multivariate Bernoulli distribution (Maydeu-Olivares & Joe, 2005). The parameters of the multivariate Bernoulli distribution are the $2^A - 1$ probabilities that an examinee possesses each of $2^A - 1$ patterns (the probability of possessing the last attribute pattern is determined, due to the unit sum of the probabilities of possession of all 2^A patterns).

When the number and characteristics of the latent classes are known and defined before the analysis is conducted (as is the case in DCMs), the latent class analysis is labeled as being restricted or confirmatory. The alternative is an unrestricted or exploratory latent class model where researchers seek to empirically find the number of latent classes that exist and subsequently hypothesize theoretical justifications for their existence; similar post hoc interpretation are made in exploratory factor analysis to explain the latent factors uncovered by a factor analysis model. In such instances, interpretations of the resulting classes or factors may be without clarity and, as such, may be statistical artifacts rather than meaningful substantive entities (Alexeev, Templin, & Cohen, in press; Bauer & Curran, 2003).

Assuming responses to items are conditionally independent given an examinee's class membership, the latent class model defines the probability of observing the vector of examinee e's scored item responses to all items (denoted $\boldsymbol{x}_e$) as a function of the attribute pattern p of examinee e ($\boldsymbol{\alpha}_e$) as

$$P(\boldsymbol{X_e} = \boldsymbol{x}_e) = \sum_{p=1}^{2^A} \upsilon_p \prod_{i=1}^{I} \pi_{i|\boldsymbol{\alpha}_e}^{x_{ei}} (1 - \pi_{i|\boldsymbol{\alpha}_e})^{1-x_{ei}}. \tag{7}$$

The term υ_p represents the proportion of examinees that have attribute pattern p, where in DCMs each latent class represents an attribute pattern . The υ_p parameters are probabilities and sum to one ($\sum_{p=1}^{2^A} \upsilon_p$ = 1). These parameters describe the relationships among the attributes (i.e., the correlations) and are thus the structural components of the DCM. The model item parameter,

$\pi_{i|\boldsymbol{\alpha}_e}$ denotes the conditional probability that the examinee provides the correct response for item i given his or her attribute pattern ($\boldsymbol{\alpha}_e$), and x_{ei} is a Bernoulli variable indicating the dichotomous item response ($x_{ei} = 0$ or $x_{ei} = 1$) of examinee e to item i.

The product term specifies the relationship between the observed data and the latent variable and is thus the measurement component of the model. This term expresses the joint probability of the observed responses as the product of the conditional probabilities of each item response, as was the same in Equation (3) for the IRT model. As in IRT, the probabilities can be expressed as a product due to the assumption of conditional independence of item responses given an examinee's measure of the latent variable. In unidimensional IRT that variable is a continuous ability and in DCMs that variable is a latent class.

Log-linear Cognitive Diagnosis Model

The Log-linear Cognitive Diagnosis model (LCDM; Henson, Templin, & Willse, 2009) parameterizes the conditional probability $\pi_{i|\boldsymbol{\alpha}_e}$ in Equation (7) through a logistic link function, as in IRT. However, for the LCDM the latent predictors are binary instead of continuous, resulting in the analysis being closely related to analysis of variance (ANOVA) methods. Binary indicators designate the presence or absence of the latent predictors, or the attributes. Effects of individual attributes (main effects) and effects of combinations of attributes (interaction effects) are modeled in the item response. The LCDM is specified as

$$\pi_{i|\boldsymbol{\alpha}_e} = P(X_{ei} = 1|\boldsymbol{\alpha}_e) = \frac{\exp(\lambda_{i,0} + \boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_i))}{1 + \exp(\lambda_{i,0} + \boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_i))}. \qquad (8)$$

The term $\lambda_{i,0}$ is the intercept that quantifies the log-odds (logit) of a correct response if examinee e has not mastered any of the attributes measured by item i. The term $\boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_i)$ is a linear combination of ANOVA-like main and interaction effects of the model. The main effects and interactions are given in the row vector $\boldsymbol{\lambda}_i^T$, where T represents the transpose. The term $\boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_i)$ is a column vector of indicators used to specify whether the main effects and interactions are present for the examinee and item. The term $\boldsymbol{q}_i = [q_{i1}, q_{i2}, \dots, q_{iA}]^T$and denotes the Q-matrix entries for item i, and $\boldsymbol{\alpha}_e = [\alpha_{e1}, \alpha_{e2}, \dots, \alpha_{eA}]$, the attribute pattern for examinee *e*. Thus, an element of $\boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_i)$ equals one if and only if (a) the item measures the attribute(s) corresponding to the effect $(q_{ia} = 1)$, *and* (b) the examinee possesses the attribute(s) corresponding to the effect $(\alpha_{ea} = 1)$. Otherwise the element equals zero, which discounts any main effect or interaction effect parameter associated with unmeasured attributes for this item or unmastered attributes in an examinee's attribute pattern. Specifically,

$$\boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_i) = \sum_{a=1}^{A} \lambda_{i,1(a)} \, (\alpha_{ea} q_{ia}) + \sum_{a=1}^{A-1} \sum_{b\,=a+1}^{A} \lambda_{i,2(ab)} \, (\alpha_{ea}\alpha_{eb} q_{ia} q_{ib}) + \dots \qquad (9)$$

where $\lambda_{i,1(a)}$ is the main effect for attribute a for item i, $\lambda_{i,2(ab)}$ is the two-way interaction effect between attributes a and b for item i, and the ellipses denotes the third through A^{th}higher-order interactions where $\lambda_{i,A}$ is the A-way interaction effect between all attributes. As in IRT, the linear predictor $(\lambda_{i,0} + \boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_i))$ contains a difficulty parameter in the form of an intercept $(\lambda_{i,0})$ and discrimination parameters in the form of attribute-specific loadings, which are referred to as main effects in the ANOVA literature. A main effect for an attribute provides a measure of discrimination between attribute patterns that do or do not have that attribute.

Unlike IRT, the linear predictor also contains discrimination parameters in the form of interactions between attributes when more than one attribute is measured by an item. A two-way interaction term provides an additional measure of discrimination between attribute patterns that have only one attribute and attribute patterns that have both attributes. For example, consider an item that measures two attributes. As a positive interaction between these two attributes becomes larger, differences in the item response probability between examinees who possess only one attribute and examinees who possess both attributes become larger. Conversely, the larger a negative interaction, the less the discriminating ability the item has because the item response probabilities would be more similar for examinees with one versus both attributes.

To demonstrate the LCDM, consider the DCM depicted in Figure 3. Item 3 is measured by two attributes, Attribute 1 and Attribute 2. For example, Item 3 may be the mathematics problem found in Figure 4. Attribute 1 is the ability to find the area of a rectangle and Attribute 2 is the ability to make conversions among units. An examinee is expected to answer the item correctly (to select Alternative A) if he or she possess both of these attributes. After substituting the Q-matrix entries into the kernel of the LDCM for Item 3, Equation (9) becomes $\boldsymbol{\lambda}_3^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_3) = \lambda_{3,1(1)}(\alpha_{e1}) + \lambda_{3,1(2)}(\alpha_{e2}) + \lambda_{3,2(12)}(\alpha_{e1}\alpha_{e2})$. An examinee may possess both attributes, only Attribute 1, only Attribute 2, or neither attribute; each of these possibilities yields a different value for the kernel of the LCDM. For example, if an examinee has mastered the first two attributes (attribute pattern $\boldsymbol{\alpha}_e = [110]$) or has mastered all three attributes ($\boldsymbol{\alpha}_e = [111]$), then $\boldsymbol{\lambda}_3^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_3) = \lambda_{3,1(1)} + \lambda_{3,1(2)} + \lambda_{3,2(12)}$. The response probability is a function of the main effects for Attribute 1 and 2 and the interaction effect between Attributes 1 and 2. The remaining possible values of $\boldsymbol{\lambda}_3^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_3)$ are given in Table 1 with the corresponding attribute pattern.

This general expression of the LCDM provides a consolidated expression for the family of DCMs. The parameters in the LCDM can be constrained in a number of different ways to yield other DCMs that are either non-compensatory or compensatory with respect to the attributes, as demonstrated by Henson, Templin, & Willse (2009). Non-compensatory models do not allow for mastered attributes to compensate for unmastered attributes. Examples of these models include the Deterministic Inputs Noisy And Gate model (DINA; Haertel, 1989; Junker & Sjitsma, 2001), the Noisy Inputs Deterministic And Gate model (NIDA; Maris, 1999), or the Non-Compensatory Reparameterized Unified Model (NC-RUM; e.g., Hartz, 2002). Compensatory models allow mastered attributes to compensate for unmastered attributes. Examples of compensatory models are the Deterministic Inputs Noisy Or Gate model (DINO; Templin & Henson, 2006), the Noisy Inputs Deterministic Or Gate model (NIDO; Templin, 2006), and the Compensatory Reparameterized Unified Model (C-RUM; Hartz, 2002).

For compensatory models, mastering additional attributes increases the probability of a correct response. One such compensatory model is the C-RUM (Hartz, 2002) presented here to highlight the distinction and parallels between DCMs and MIRT models. In the log-linear framework, the C-RUM is be expressed as

$$\pi_{i|\boldsymbol{\alpha}_e} = P(X_{ie} = 1|\boldsymbol{\alpha}_e) = \frac{\exp(\lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1(a)}\,(\alpha_{ea}q_{ia}))}{1 + \exp(\lambda_{i,0} + \sum_{a=1}^{A} \lambda_{i,1(a)}\,(\alpha_{ea}q_{ia}))}. \qquad (10)$$

There are extensive similarities between Equations (6) (the two-parameter MIRT model) and (10). For both there are no interactions between the latent traits; the kernel is simply a linear combination of an intercept and all relevant traits which are weighted with a discrimination parameter. However, for the C-RUM the latent variable is categorical, whereas in the MIRT model it is continuous. This is also seen in comparing Figure 3 with Figure 2; the difference in the LCDM and the MIRT model is the categorical nature of the traits or attributes in the LCDM, which are depicted by the differential shading of the attributes.

Structural Model for the LCDM

The structural model for the LCDM describes the relationships among the latent attributes and can be specified in a number of different ways. The structural parameter υ_p from Equation (7) expresses the marginal distribution for the attribute pattern (i.e., the probability an examinee has any attribute pattern). The structural model can be parameterized using a log-linear framework, similar to how the conditional probability of an item response was parameterized in Equation (8). Using the log-linear parameterization, the probability an examinee will have attribute pattern p is a function of main effects corresponding to individual attributes mastered in pattern p and interactions corresponding to combinations of attributes mastered in pattern p. This probability is expressed as (Henson & Templin, 2005; Rupp, Templin, & Henson, 2010)

$$\upsilon_p = \frac{\exp(\mathrm{\mu_p})}{1 + \exp(\mathrm{\mu_p})} \tag{11}$$

where

$$\begin{aligned}\mu_p = \gamma_0 + \textstyle\sum_{a=1}^{A} \gamma_{1(a)}\, \alpha_{pa} + \sum_{a=1}^{A-1} \sum_{b=a+1}^{A} \gamma_{2(ab)}\, (\alpha_{pa}, \alpha_{pb}) + \ldots \\ +\ \gamma_{A(ab\ldots)}\ \left(\textstyle\prod_{a=1}^{A} \alpha_{pa}\right).\end{aligned} \tag{12}$$

The term γ_0 is the intercept, the value of the kernel for the attribute pattern with no mastered attributes. The term $\gamma_{1(a)}$ is the main effect for attribute a for attribute pattern p, $\gamma_{2(ab)}$ is the interaction effect between attributes a and b for pattern p, and the ellipses denote the third through A^{th} higher-order interactions where γ_A is the A-way interaction effect between all attributes. A reduced version of this saturated model can also be used to limit the higher-order interactions that are insignificant (Rupp, Templin, & Henson, 2010).

A different structural model may be employed to describe the pairwise associations among attributes by using a high-order factor model that imposes an underlying latent continuum for the attributes to describe their associations with tetrachoric correlations; the structural model defining the tetrachoric correlations is conceptually equivalent to the structural model used in confirmatory factor analysis (Templin & Henson, 2006). This structural approach is useful when associations between attributes are of interest to the researcher. Other specifications of structural models that can be used are described by Rupp, Templin and Henson (2010).

Nominal Response Log-linear Cognitive Diagnosis Model

As the NR IRT model extend the 2-PL IRT model, the LCDM can also be extended to model nominal responses. The latent class model, where $\boldsymbol{x_e}$ is now examinee e's nominal response pattern to all I items on a test, is expressed as

$$P(\boldsymbol{X_e} = \boldsymbol{x_e}) = \sum_{p=1}^{P} \upsilon_p \prod_{i=1}^{I} \prod_{j=1}^{J_i} \pi_{n_{ij}|\boldsymbol{\alpha}_e}^{[x_{ei}=n_{ij}]} \qquad (13)$$

where [·] are Iverson brackets indicating that if $x_{ei} = n_{ij}$, $[x_{ei} = n_{ij}]$ =1; otherwise, $[x_{ei} = n_{ij}]$ = 0. The nominal response LCDM (NR LCDM; Templin & Bradshaw, under review) defines the conditional probability that an examinee will provide a nominal response of n_{ij} among the set of J_i response alternatives for item i, given his or her attribute pattern p, as

$$\pi_{n_{ij}|\boldsymbol{\alpha}_e} = P\left(X_{ei} = n_{ij}\middle|\boldsymbol{\alpha}_e\right) = \frac{\exp(\lambda_{n_{ij},0} + \boldsymbol{\lambda}_{n_{ij}}^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_{n_{ij}}))}{\sum_{j=1}^{J_i} \exp(\lambda_{n_{ij},0} + \boldsymbol{\lambda}_{n_{ij}}^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_{n_{ij}}))} \tag{14}$$

where

$$\boldsymbol{\lambda}_{n_{ij}}^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_{n_{ij}}) = \sum_{a=1}^{A} \lambda_{n_{ij},1(a)} \, (\alpha_{ea} q_{n_{ij}a}) + \sum_{a=1}^{A-1} \sum_{b=a+1}^{A} \lambda_{n_{ij},2\,(ab)} \, (\alpha_{ea} \alpha_{eb} q_{n_{ij}a} q_{n_{ij}b}). \tag{15}$$

Interpretations of the parameters are analogous to the LCDM with a few distinctions. The NR LCDM has alternative-specific, as opposed to item-specific, Q-matrix entries. The vector $\boldsymbol{q}_{n_{ij}} = [q_{n_{ij}1}, q_{n_{ij}2}, \ldots q_{n_{ij}A}]^T$ and denotes the Q-matrix entries for alternative n_{ij} as an $(Ax1)$ column vector. The NR LCDM also has alternative-specific intercepts, main effects and interactions; the effect of the attribute is measured at the alternative level instead of at the item level. The NR LCDM can be identified by constraining the sum of each type of model parameter to be zero. More explicitly, the alternative-specific intercepts within an item sum to zero (i.e., $\sum_{j=1}^{J_i} \lambda_{n_{ij},0} = 0$ for each item), the sum of the alternative-specific main effects for each attribute within an item sum to zero (i.e., $\sum_{j=1}^{J_i} \lambda_{n_{ij},1(a)} = 0$ for each attribute and item), and the sum of the alternative-specific interaction effects for each set of interacting attributes within an item sum to zero (e.g., two way interactions for pair of attributes within an item: $\sum_{j=1}^{J_i} \lambda_{n_{ij},2(aa')} = 0$). Alternately, the NR LCDM can be identified by setting all item parameters equal to zero for a given alternative. This alternative is then referred to as the baseline alternative, where the kernel for this alternative $(\lambda_{n_{ij},0} + \boldsymbol{\lambda}_{n_{ij}}^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_{n_{ij}}))$ equals zero.

The LCDM estimates a conditional probability each examinee will answer the item correctly, whereas the NR LCDM estimates the conditional probability an examinee will select each alternative. Different alternatives may measure different sets of attributes. The NR LCDM can preserve the diagnostic information provided by unique Q-matrix entries for each alternative. When sample sizes are large, the NR LCDM was found to capitalize on information in the incorrect alternatives as demonstrated by greater classification accuracy when compared to the LCDM for dichotomous responses (Templin & Bradshaw, under review).

For example, consider again the item in Figure 4, which was Item 3 for the DCM pictured in Figure 3. The Q-matrix entries for this item, if estimated with the LCDM, are $\boldsymbol{q}_i$= [1,1,0]; the item measures Attributes 1 and 2 but not Attribute 3. The Q-matrix for this item, if estimated with the NR-LCDM, can be found in the first four columns of Table 2. For example, consider Alternative B where only Attribute 1 is measured. An examinee who selects Alternative B incorrectly converts feet to inches (has not mastered Attribute 2), but demonstrates an ability to find the area of a rectangle by applying the operation of multiplication to the dimensions given (has mastered Attribute 1). Thus, a response of *B* indicates the absence of Attribute 2, yet the presence of Attribute 1. Conversely, an examinee who selects Alternative C correctly converted feet to inches (mastered Attribute 2), but found the perimeter of the rectangle instead of the area (has not mastered Attribute 1). Thus, a response of *C* indicates the absence of Attribute 1, yet the presence of Attribute 2. Under the LCDM, a response of *B* is not unique from a response of *C*; both responses are evidence for a lack of mastery of Attribute 1 and 2.

CHAPTER 3

THE SCALING INDIVIDUALS AND CLASSIFYING MISCONCEPTIONS MODEL

The previous chapter presented a selection of models from the IRT and DCM psychometric frameworks. The chapter drew parallels and highlighted distinctions between the models to provide the background requisite for understanding components of the new model introduced in the present chapter. This chapter will begin by providing examples of previous assessments that acknowledged errors or misconceptions in the test development process. It will conclude by introducing the Scaling Individuals and Classifying Misconceptions (SICM) model, a new model that provides a means to measure misconceptions. The goal of this chapter is to illustrate the potential advantages and to delineate the structure of the SICM model by juxtaposing its nature and statistical properties with other existing latent variable models that are commonly applied to estimate latent traits in educational and psychological realms.

Previous Assessments Explicitly Mindful of Misconceptions

As previewed in Chapter 1, the SICM model seeks to provide more detailed, multidimensional feedback to stakeholders by identifying students' misconceptions. The SICM model identifies misconceptions by students' incorrect responses. Therefore, the SICM model hinges on a key feature of a test: the incorrect alternatives for items must reflect common misconceptions students have or typical errors students may make. The following sections highlight six assessments – four from science and two from statistics – that have this feature required by the SICM model. These sections demonstrate that empirical theories exist about students' misconceptions in various content domains, that the desire to capture them through an assessment is not uncommon, and that a psychometric model tailored to these types of empirical theories is needed. Each section will briefly describe the format and purpose of the assessment and also highlight the sub-optimal alignment of the test design and psychometric method used.

Force Concepts Inventory

The Force Concepts Inventory (FCI; Hestenes, Wells, & Swackhamer, 1992) is an assessment of the Newtonian concept of force, which is an important concept taught in introductory physics. The FCI is a revision of the Mechanics Diagnostic, an assessment developed by Halloun & Hestenes (1985). The assessment consists of 30 multiple-choice items. The authors define 30 misconceptions that describe students' beliefs about the concept of force that theoretically exist before a student has been exposed to instruction on the concept. For example, a misconception regarding gravity is the belief that heavier objects fall faster. Incorrect alternatives for the items are written to reflect these misconceptions, and on average an individual misconception appears in approximately 2.7 items and 3.7 alternatives. Researchers interviewed students about their responses to ensure that the selection of an incorrect response was an indicator of the presence of the misconception.

Hestenes et al. (1992) suggest use of the FCI as a diagnostic tool to "identify and classify misconceptions" (p.13); however, the results from the assessment focus on the total percentage of items answered correctly on the assessment, a score that reflects what the authors referred to as a student's "Newtonian understanding" (p. 11). Although the Hestenes et al. (1992) note that, "as a rule, 'errors' on the FCI are more informative than 'correct' choices" (p. 2), a summary or measure of the presence or absence of misconceptions was not discussed. Inconsistencies exist with respect to the type of information the authors are seeking to collect with their test design (i.e., the errors are students making), and the type of information the scoring of the assessment is providing (i.e., a composite score representing percentage correct).

Astronomy Concept Inventory

The Astronomy Concept Inventory (ACI; Sadler (1998)) is an assessment of introductory astronomical concepts. Referred to as a *distractor driven assessment*, like the FCI, the ACI is a multiple-choice assessment in which the incorrect alternatives reflect common incorrect conceptions that students may have. Researchers identified these misconceptions through qualitative research that studied developing or incorrect conceptions students hold.

The ACI is a 47-item inventory created to assess curriculum materials that explicitly target popular student misconceptions by measuring the students' growth in their understanding of astronomical concepts between 8th and 12th grade. The item responses were modeled by the multiple-choice model (MC; Thissen & Steinburg, 1984). The MC model provides the probability an examinee selects each alternative (correct and incorrect alternatives), and also the probability that an examinee "does not know" the answer and simply provided a guess by randomly selecting amongst the provided alternatives. Sadler (1998) interpreted trace lines for incorrect alternatives that do not monotonically decrease as a function of ability as indications that misconceptions may be "markers of progress toward scientific understanding and are not impediments to learning" (p. 265). This conclusion is made because the probability an alternative that measures a misconception is selected is not necessarily higher for examinees with lower abilities, suggesting that misconceptions are sometimes manifested in students who are progressing towards higher level of understandings.

Sadler (1998) suggests the use of this type of assessment to “aid teachers in diagnosing student conceptions and to easily measure conceptual change” (p. 290). Although changes in overall ability can be quantified using the MC model, the only way to assess or “diagnose” the misconceptions is to study the trace lines corresponding to measured misconceptions on an individual item and student basis, or to tally the number of times an alternative that measures a given misconception is selected by the student. However, this practice of post-hoc analysis is time consuming and tedious, particularly is if it delegated to teachers who may teach large numbers of students over a set of classes. As for the FCI, the researcher expended effort to carefully create alternatives that aligned to misconceptions and highlighted providing information about student’s misconceptions an important utility of the assessment. However, the psychometric method used did not measure these misconceptions; instead, the method focused on measuring an overall continuous ability.

Astronomy and Space Science Concepts Inventory

Building upon the ACI, Sadler, Coyle, Miller, Cook-Smith, Dussault, and Gould (2010) created the Astronomy and Space Science Concepts Inventory (ASSCI). A bank of 211 items were written to measure astronomical concepts for three grade bands spanning from kindergarten through 12th grade (K-4, 5-8, 9-12). For these grade bands, respectively, sets of items measure 4, 9, and 7 standards and address 9, 8, and 7 misconceptions through incorrect alternatives. From these items, shorter final versions of an assessment to measure astronomical concepts were created for each grade band; amongst selection criteria for items on the final form were the misconception strengths in the incorrect alternatives for the items. Misconception strength of an incorrect alternative was defined as the proportion of examinees selecting the incorrect alternative. Higher misconception strengths were preferred because the researchers sought items that strongly elicited students' misconceptions.

Measures of ability were given by overall percentage correct and measures of growth were given by increases in scores using the metric of the standard deviation. As in the ACI, Sadler et al. (2010) used distractor driven multiple-choice items to compose a test that has "the capability to identify both examples of student misconceptions as well as their prevalence/frequency within a population" (p.3). As with the FCI and ACI, the number of times a student selects an alternative consistent with a given misconception is the measure that the student holds that misconception. This is a very coarse and unreliable measure considering as few as one item may be indicating whether or not a student has a misconception.

Ordered Multiple-choice Item Format

Briggs, Alonzo, Schwab, & Wilson (2006) developed an item format called the ordered multiple-choice format. Student understanding of the science content area "Earth in the Solar System" was described using five sequential levels of understanding. Key understandings indicative of each level were delineated to create what was referred to as a concept map. A concept map defines the distinct levels that a given unidimensional construct is assumed to have.

Misconceptions, or common errors, were not the focus of the assessment, but played a unique role in the test development process. Determining a student's level of understanding for the Earth and Science concept was the focus of the assessment, but misconceptions were used as key indicators of different levels of understanding. The items were written in a somewhat similar fashion as items on the FCI, ACI, and ASSCI. Each possible alternative was linked to a student level of understanding by reflecting a misconception that corresponded to that level.

Interestingly, this practice explicitly defines a linear relationship between misconceptions and overall ability for this concept. Pinpointing where misconceptions appear along this continuum implies a hierarchical relationship among the misconceptions, meaning some misconceptions do not exist simultaneously with others.

The goal of this assessment was not to measure misconceptions, but Briggs et al. (2006) have difficulties in capitalizing on the information embedded in their items, similar to the researchers mentioned above. The number of items on the assessment was small, so a total score from Classical Test Theory was provided to measure an overall ability. Given the availability of a larger sample, Briggs et al. (2006) suggest using an IRT model called the Ordered Partial Credit model (Wilson, 1992). In this model, the information about the levels of the alternatives would contribute to estimating a continuous overall ability. With respect to information regarding the level at which a student performs, test results can be analyzed using a CTT subscore approach to provide numerical summaries of the proportion of items that an examinee answers correctly at each level. However, neither the CTT total score nor the IRT score describes the level along the continuum at which a student is performing. This is noted here to reiterate that CTT and IRT methods are ubiquitously used, even when the design of the assessment includes intricate, unique features and the purpose of the assessment has other goals besides measuring an overall ability. Furthermore, in a variety of situations where an assessment has a purpose other than measuring an overall ability, researchers are left to count, to revert back to the early psychometric methods of CTT.

Statistical Reasoning Assessment

The Statistical Reasoning Assessment (SRA) is a 20 item test with alternatives that measure both correct and incorrect reasoning (Garfield, 1998). Unique to this test is the feature that each item does not have a single correct answer, rather multiple correct (and incorrect) answers exist. Every correct answer aligns to a correct line of reasoning and every incorrect answer aligns to an incorrect line of reasoning. Eight types of correct reasoning were defined and measured, and eight categories of incorrect reasoning were defined. Total scores on the assessment were not the focus of the analysis, but rather subscores for correct and incorrect types of reasoning were emphasized. Garfield and Chance (2000) noted that "although individual items could be scored as correct or incorrect and total correct scores could be obtained, this single numerical summary seemed uninformative and did not adequately identify students' reasoning abilities" (p. 117). This observation reflects an emphasis on the specific components of correct and incorrect reasoning. CTT subscores reflected the number of times a line of correct or incorrect reasoning was supplied for the 16 categories of reasoning. The range of these raw scores was from two to eight. Thus, even though the focus of the assessment was on the individual components of reasoning, a statistical means other than CTT for determining whether or not a student possesses the ability to reason correctly or possesses an understanding that reflects incorrect reasoning was not used.

Probability Reasoning Questionnaire

Khazanov (2009) developed an assessment to measure misconceptions that students display when reasoning about statistical probability. The assessment was a 16 item two-tiered multiple-choice assessment. The stem of the item was followed by the first tier question.

Alternatives in the first tier provided answers to the question. The second tier involved a second question probing for the reasoning behind why students selected the answer in the first tier. Combinations of answers in the first and second tier reflected various misconceptions about probability, and for each item, one answer combination was considered accepted or accurate probabilistic reasoning. The number of alternatives for a question in either tier ranged from three to five. Each alternative on the assessment reflected answers students gave to open-ended versions of the items or types of reasoning they provided during interviews.

This assessment had a greater focus on trying to diagnose whether the student had the misconception, evidenced by the large number of items and alternatives that targeted each misconception. Three misconceptions that are well-documented in literature about statistical reasoning were the focus of the assessment: representative bias, equiprobability bias, and outcome orientation misconception (e.g., Khazanov, 2008). Respectively, the misconceptions were measured by 11, 14, and 15 items and by 19, 14, and 26 combinations of responses to the first and second tiers.

Despite the careful crafting of an exam to measure the misconceptions, the analysis of the results was simply counting the number of times a student provided an answer combination that aligned with a given misconception, similar to previous assessments discussed that used CTT total or subscores. A cut-off of two was assigned to designate whether or not a student had a misconception and would benefit from learning activities created to replace the misconception with correct understanding. This practice is akin to classifying examinees into proficiency categories according to a continuous score. The score does not classify examinees; a non-statistically grounded cut-off score results in a subjective dichotomization.

Diagnostic Information versus a Diagnosis

Salient similarities in the six assessment projects presented in the previous sections are (a) the presence of highly developed and empirically supported understandings of misconceptions that students possess while learning about a given domain, (b) careful construction of items whose incorrect alternatives elicit these misconceptions, and (c) a lack of a psychometric model consistent with the design of the items.

Researchers developing these assessments have theorized that misconceptions exist, and that they manifest themselves in incorrect alternatives selected by students. Making this theory explicit is a significant contribution that researchers such as these have provided for assessment development. Additionally, establishing the validity of these items (and alternatives) through rigorous interviews and one-on-one interactions with students demonstrates model validity studies that should be included in the assessment development process. However, the careful attentiveness to the validity of the misconceptions being measured is undermined by the omission of the reliability of scores or subscores implicitly used as measures of these constructs. An incorrect response to an individual item or the percentage of incorrect responses to a very few number of items do not provide reliable or objective measures of a misconception. Although none of the researchers who developed the six assessments described above are denying this claim, they do suggest or demonstrate the use of this information to make subsequent, subjective decisions about whether or not an examinee has a misconception, implicitly putting forth a simplistic psychometric model for measuring misconceptions. This simple model relies on the assumption, which cannot be tested, that each alternative contributes equally to the measure of the misconception.

In the light of the gap in psychometric theory and practice for providing reliable measures of misconceptions, using a simple model seems intuitive and reasonable, although far from fulfilling when the intent of the assessment and effort of the researchers is considered. Often in addition to seeking information about the level or degree of a student's overall ability, these researchers sought information about the student's misconception and how prevalent a given misconception is among a sample or population of students. They sought a diagnosis (e.g., does this student have this misconception?), and although they have gathered some *diagnostic information* using the assessment, a *diagnosis* is not made from the analysis of the assessment.

A distinction Rupp, Templin and Henson (2010) make about the definition of a diagnosis is that a diagnosis *is* a decision. Given this denotation of diagnosis, assessments described above do not provide diagnoses with respect to possession of each misconception. However, they can be considered to provide diagnostic information, information that can contribute to a subsequent decision, albeit a non-statistical decision. For example, the assessments provide frequencies that teachers or other interested stakeholders may consider and, perhaps combined with their own experiences with students, use to decide whether or not the student has the misconception. The assessments also provide a total score that can be subjectively dichotomized through human judgment to decide how to classify students.

The SICM model is a psychometric model that seeks to statistically diagnose students' possessions of misconceptions. The SICM model reflects empirical theories of the assessment development projects previously described. The model acknowledges that an overall ability lies on a continuum, but assumes that misconceptions also exist and can be measured with assessments created in the same vein as these projects. Using the SICM model, misconceptions can be treated as variables to be measured and included in the model of the item response. By attending to statistical considerations of the model during the test construction process, the SICM model can measure misconceptions more reliably.

By statistically modeling these theories, the SICM model provides a means to quantitatively evaluate existing empirical theories about misconceptions and their effects in a testing situation. Testing these theories helps strengthen them and also improves the assessment. The SICM model does not assume that each alternative contributes equally to the measure of a misconception. Instead it quantifies how strongly an alternative is related to a misconception, permitting a statistical test of whether or not an alternative is, indeed, measuring a misconception. Results from the model may be used to (a) improve the design of the item and/or alternative, or (b) shape theories. For example, if analyses indicate an alternative is not measuring a misconception, then test developers can revise the item to more validly measure a misconception, or researchers can view this result as evidence that questions the theory that the misconception exists.

Theories about the relationships among misconceptions can also be molded using results of analyses from the SICM model. The model estimates how correlated misconceptions are, contributing to theories by statistically describing how strongly related pairs of misconceptions are. The model can provide insights for other relationships among misconceptions by determining within a student which combinations of misconceptions are possible for a student to have. Some misconceptions may not simultaneously exist with or without other misconceptions.

The previous section explained a gap in psychometric theory that the SICM model seeks to fill by measuring an overall ability along with misconceptions. By doing so, the SICM model also offers a means to provide multidimensional feedback within the framework of prevailing IRT methods. The next sections provide the statistical details of the SICM model.

The Scaling Individual and Classifying Misconceptions Model

The SICM model is a new psychometric model unique from other psychometric models in two ways. First, the SICM model treats a misconception as a latent variable or construct of interest to be measured and estimated. The SICM model also makes assumptions about the nature of a misconception's existence. The model assumes that misconceptions are categorical latent variables that are dichotomous. The two levels are characterized by the presence or absence of a misconception. Presence of a misconception will also sometimes be referred to as possession of a misconception. Second, following empirical theories driving the assessments discussed in the previous section, the SICM model acknowledges that there is a larger construct to be measured on a continuum that exists in addition to misconceptions that are present or not. Thus, the model uniquely estimates a continuous ability *in addition to* a set of categorical misconceptions.

Like the NR LCDM, the SICM model is a nominal response model that capitalizes on the diagnostic information found in the incorrect alternatives on a multiple-choice test. However, the SICM model not only measures a set of categorical variables as the NR LCDM does, but also estimates a single continuous variable that represents an overall ability as a unidimensional IRT model does. Instead of modeling attributes that represent skills or abilities as DCMs commonly do, attributes in the SICM model represent misconceptions. To model an item response, these misconceptions are specified as latent predictors for the incorrect alternatives, and a continuous overall ability is specified as the latent predictor for the correct alternative. Before proceeding with the explanation of the SICM model, a brief discussion of the use of the term *misconception* to describe categorical latent traits for the SICM model will be provided.

Meaning of Misconception

The term misconception will be used to describe the latent traits estimated in the SICM model, even though the nature of the misconception may not always be conceptual and the negative connotation implied by the prefix *mis*- may not always accurately describe the trait. Depending upon the context, the misconceptions may be defined as procedural or conceptual errors, or as inaccuracies in thinking, reasoning, or problem solving. These inaccuracies may reflect commonsense or innovative lines of reasoning or partial understandings that are positive developmental stages of understanding through which students may progress. Although the reasoning manifested due to the misconception may or may not be more advanced in comparison to other conceptions, in a testing scenario, understanding reflective of a misconception is unilaterally inferior to understanding that yields a correct answer.

Misconceptions are ultimately defined by content experts or cognitive scientists and are founded upon strong empirical research; the SICM model is simply a statistical tool to model these traits and help form falsifiable statistical hypotheses regarding the nature of the misconceptions (i.e., how they are measured by an item, how they interact, how they are correlated). Although the nature of misconceptions may vary from one domain to another, or even within a domain, requisite for a misconception to be measured by the SICM model is that a misconception must be a trait of the examinee. A trait of an examinee is a characteristic of the examinee that is exhibited with a degree of stability. For a misconception to be a trait, possession of a misconception must result in systematic responses to an item measuring the misconception throughout the entire assessment. A careless error or an aberrant mistake is not the product of a misconception; misconceptions produce systematic incorrect answers.

Misconceptions are often targeted by instruction for the purpose of eliminating them, so they are not expected to be fixed traits over time. They are, however, expected to be stable enough to produce consistent responses at a given testing occasion. The SICM model rests heavily on the assumption of stable misconceptions; it cannot reliably measure unsystematic errors. The more the misconception represents a firmly cemented belief held by the examinee, the more consistently the misconception will produce the expected incorrect response and, in turn, the more reliably that misconception can be measured.

Formulation of the SICM Model

Given a set of J_i response categories or possible alternatives for an item i, the SICM model utilizes a nominal response mixture item response model that defines the probability of observing an examinee’s nominal response pattern to I items $(\boldsymbol{x_e})$ as

$$P(\boldsymbol{X_e} = \boldsymbol{x_e}) = \int_{-\infty}^{\infty} \sum_{p=1}^{P} \upsilon_p \prod_{i=1}^{I} \prod_{j=1}^{J_i} \pi_{n_{ij}|\boldsymbol{\alpha}_{e,}\theta_e}^{[x_{ei}=n_{ij}]} P(\theta)\, \partial\, \theta. \tag{16}$$

The terms υ_p and $P(\theta)$ are the structural components of the model, describing the distributions of and relationships among of the latent variables in the model, with θ and $\boldsymbol{\alpha}$ held independent. The term υ_p describes the proportion of examinees that have misconception/attribute pattern p, parameterized as a function of the individual attributes by a log-linear model as described in Chapter 2 for DCMs (e.g., Henson & Templin, 2005; Rupp, Templin, & Henson, 2010). The term $P(\theta)$ is the density function of ability, with $\theta_e \sim N(0,1)$ for identifiability.

The parameter $\pi_{n_{ij}|\boldsymbol{\alpha}_e,\theta_e}$ denotes the conditional probability that an examinee's response to item i will be the selection of alternative j from the set of J_i alternatives for item i (i.e., $x_{ei} = n_{ij}$), given examinee *e*'s attribute pattern ($\boldsymbol{\alpha}_e$) *and* continuous ability (θ_e). The brackets [·] are Iverson brackets indicating that if $x_{ei} = n_{ij}$, $[x_{ei} = n_{ij}]$ =1; otherwise $[x_{ei} = n_{ij}]$ =0. The parameterization of $\pi_{n_{ij}|\boldsymbol{\alpha}_e,\theta_e}$ represents the measurement component of the SICM model in that it quantifies how the latent variables (misconceptions and ability) are related to the observed item responses. For the SICM model, ability is measured by the correct alternative on each item, and the misconceptions are measured by the incorrect alternatives. Not every incorrect alternative measures each misconception, so an indicator variable is used to specify when a misconception is measured by an item alternative. Mimicking DCM practices, specifications are set *a priori* and are described in an item-by-alternative-by-misconception Q-matrix. The entries in the Q-matrix are indicators denoted by $q_{n_{ij}a}$, where $q_{n_{ij}a} = 1$ if alternative *j* for item i measures misconception a and $q_{n_{ij}a} = 0$ otherwise.

The SICM model parameterizes $\pi_{n_{ij}|\alpha_e,\theta_e}$ in Equation (16) by utilizing a multicategory logistic regression model (e.g., Agresti, 2002) that models the $J_i - 1$ non-redundant logits with the Jth alternative as the baseline category as:

$$\log\left(\frac{P(X_{ei}=n_{ij}|\alpha_e,\theta_e)}{P(X_{ei}=n_{iJ}|\alpha_e,\theta_e)}\right)= \lambda_{n_{ij},0} + \lambda^{*}_{n_{ij},\theta}(\theta_e)(q_{n_{ij}}) + \boldsymbol{\lambda}^{T*}_{n_{ij}}\boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_{n_{ij}}) \qquad (17)$$

for every n_{ij} such that $j \neq J$, where

$$\lambda^{*}_{n_{ij},\theta}(\theta_e)(q_{n_{ij}}) =\lambda_{n_{ij},\theta}(\theta_e)(q_{n_{ij}}) - \lambda_{n_{iJ},\theta}(\theta_e)(q_{n_{iJ}}) \qquad (18)$$

and

$$\boldsymbol{\lambda}_{n_{ij}}^{T*}\,\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}) = \boldsymbol{\lambda}_{n_{ij}}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}) - \boldsymbol{\lambda}_{n_{iJ}}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{iJ}}). \tag{19}$$

The correct alternative is specified as the baseline category, denoted n_{iJ}, to simplify Equation (17). In Equation (18), $q_{n_{ij}}$ equals zero for every alternative n_{ij} where $j \neq J$ because the incorrect alternatives do not measure θ. In Equation (19), $q_{n_{iJ}}$ always equals zero because the correct alternative does not measure any misconceptions. Therefore the $J_i - 1$ equations specifying the log-odds of selecting an incorrect alternative over the correct alternative in the SICM model can be equivalently formulated as:

$$\log\left(\frac{P(X_{ei}=n_{ij}|\alpha_e,\theta_e)}{P(X_{ei}=n_{iJ}|\alpha_e,\theta_e)}\right)= \lambda_{n_{ij},0} - \lambda_{n_{iJ},\theta}(\theta_e) + \boldsymbol{\lambda}_{n_{ij}}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}) \tag{20}$$

for every n_{ij} such that $j \neq J$, and the conditional probability that n_{ij} will be selected on item i can be reexpressed as:

$$P\big(X_{ei} = n_{ij}\big|\alpha_e,\theta_e\big) = \frac{\exp(\lambda_{n_{ij},0} - \lambda_{n_{iJ},\theta}(\theta_e) + \boldsymbol{\lambda}_{n_{ij}}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}))}{\sum_{j=1}^{J}\exp(\,\lambda_{n_{ij},0} - \lambda_{n_{iJ},\theta}(\theta_e) + \boldsymbol{\lambda}_{n_{ij}}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}))} \tag{21}$$

The intercept $\lambda_{n_{ij},0}$ is the logit of selecting an incorrect alternative n_{ij} over the correct alternative n_{iJ} for an examinee with an ability of zero who possesses *none* of the misconceptions measured by alternative n_{ij}. The more difficult the alternative is, the larger the intercept will be as it reflects the likelihood of choosing an incorrect alternative over the correct one.

The term $\lambda_{n_{iJ},\theta}$ is the loading for continuous ability. In an IRT context, this term is the discrimination parameter for ability and is interpreted similarly: the higher the value of $\lambda_{n_{iJ},\theta}$, the more different the probability of choosing the correct alternative is for examinees at different locations on the ability continuum. The probability the item is answered correctly should increase as ability increases, so the value of this parameter should be positive. The loading $\lambda_{n_{iJ},\theta}$ represents the increase in the logit of a correct response for every unit increase in ability.

Using notation consistent with the LCDM and the NR LCDM, the term $\boldsymbol{\lambda}_{n_{ij}}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}})$ is a linear combination of main and interaction effects of the model. The vector $\boldsymbol{q}_{n_{ij}} = [q_{n_{ij}1}, q_{n_{ij}2}, \dots, q_{n_{ij}A}]^T$ and denotes the A Q-matrix entries for alternative *j* of item i, and $\boldsymbol{\alpha}_e$ is the misconception (attribute) pattern for examinee *e*. The term $\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_i)$ is a column vector of indicators with elements that equal one if and only if (a) the item measures the misconception or set of misconceptions corresponding to the parameter $(q_{n_{ij}} = 1)$ *and* (b) the examinee possesses the misconception or set of misconceptions corresponding to the parameter$(\alpha_{ea} = 1)$. Specifically, $\boldsymbol{\lambda}_{n_{ij}}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}})$ equals

$$\sum_{a=1}^{A}\lambda_{n_{ij},1(a)}\,(\alpha_{ea}q_{n_{ij}a}) + \sum_{a=1}^{A-1}\sum_{b=a+1}^{A}\lambda_{n_{ij},2(ab)}\,(\alpha_{ea}\alpha_{eb}q_{n_{ij}a}q_{n_{ij}b})+\dots \qquad (22)$$

where $\lambda_{n_{ij},1(a)}$ is the main effect for misconception a for the j^{th} alternative of item i, $\lambda_{n_{ij},2(ab)}$ is the interaction effect between attributes a and b $(\neq a)$ for the j^{th} alternative of item i (if alternative j of item i measures two or more misconceptions), and the ellipses denote the third through A^{th}higher-order interactions for options on items that measure more than two misconceptions, where $\lambda_{n_{ij},A(1,2,\dots,A)}$ is the A-way interaction effect between all attributes.

Main effects and interactions are also discrimination parameters but with respect to misconceptions, so more discriminating items have larger main effects and interactions.

To identify the model, as is usual for identification of a baseline-category logit model, an arbitrary category must be treated as the baseline category, and all parameters for the baseline category are set equal to zero. Additionally, the main effect parameters are constrained to ensure monotonicity for attributes and for ability, meaning that (a) the possession of a misconception related to an alternative never leads to a decrease in the probability of selecting that alternative over a different alternative, and (b) an increase in ability never results in a decrease in the probability of answering the item correctly.

The SICM Model Illustrated as a Combination of an IRT model and a DCM

The SICM model posits that there is a dominant continuous trait being measured by the assessment that largely explains the covariance among the selection of the correct alternatives for a set of items. It additionally hypothesizes that there exists a set of categorical misconceptions, each of which a student does or does not possess, that systematically account for the variations in the selections amongst the incorrect alternatives for a set of items. It models that hypothesis by making alternative-specific manipulations of the presence of discrete latent variables in the model. The hypotheses of the SICM model can be verified by testing the significance of the main effects for ability and for misconceptions. In this way, the SICM model can help test and advance empirical theories about why an examinee may provide an incorrect response to an item.

To further illustrate the differences among the NR IRT, the NR LCDM model, and the SICM model, consider again the item in Figure 4. From an NR IRT perspective, the level of a person’s overall “math” ability explains the variations in the item responses and is the only latent variable being measured by the alternatives. Using the SICM model in Equation (17), if $\lambda_{n_{ij},\theta}$ is always equal to 1 (i.e., θ is measured by every alternative), and $\boldsymbol{q}_{n_{ij}}$ is fixed to be a 1 x *A* vector of zeros (i.e., no misconceptions are measured), then the item response function is specified as Bock's (1972) NR IRT model.

From a NR DCM perspective, using the NR LCDM as described in Chapter 2, two discrete abilities are needed to answer this item correctly: the ability to find the area of a rectangle (Attribute 1; α_1) and the ability to make conversions among units within a metric system (Attribute 2; α_2). Using the SICM model in Equation (17), if $\lambda_{n_{ij},\theta}$ is always equal to 0 (i.e., θ is measured by no alternative), then the item response function is specified as the NR LCDM where $\boldsymbol{\alpha}_e$ is defined as a pattern of attributes or skills, instead of a pattern of misconceptions as in the SICM model.

To model the example item from the SICM model perspective, the attributes are defined as misconceptions or errors. Attribute 1 (α_1) will be redefined as the *in*ability to find the area of a rectangle and Attribute 2 (α_2) as the *in*ability to make conversions among units within a metric system. An examinee is expected to answer the item correctly (to select Alternative A) if he or she possess neither of these attributes and has a modest level of overall ability. The Q-matrix for the SICM model for the example item is given the last five columns of Table 2. The difference in the Q-matrices for the NR LCDM and the SICM model is that the entries corresponding to the attributes are exactly opposite of one another (because attributes have opposite meanings in the two models), and the SICM model measures a continuous ability (θ) by the correct answer.

Figure 5 illustrates the SICM model for this item with a path diagram. This figure depicts the measurement components of the model, where the continuous ability is measured by the correct answer, Alternative A, and the two misconceptions (attributes) are measured by the incorrect alternatives (*B*, *C*, and *D*). The directional arrows from the misconception to an alternative indicate that the presence of the misconception influences the probability that that alternative is selected. The figure shows the structural components of the model by (a) the arrows between the two misconceptions, indicating correlations among attributes are modeled, and (b) the absence of arrows between the misconceptions and ability, indicating correlations between ability and misconceptions are not modeled, as discussed subsequently.

Figure 6 provides hypothetical item response probabilities for the NR IRT model, the NR LCDM, and the SICM model to compare the type of information each model would provide about the item in Figure 4. The legend of the first graph corresponds to each model containing the Q-matrix entries for the corresponding model. For the NR IRT model, in the top left graph, the item response probability is solely a function of ability (θ). The NR LCDM, shown in the next 4 graphs, provides the item response probability by the group that an examinee is in. Groups are defined by attribute patterns ($\boldsymbol{\alpha}$). When an attribute pattern an examinee has corresponds to the attribute pattern measured by an alternative, the examinee is most likely to select that alternative. The SICM model, shown in the last 4 graphs, provides the item response probability not only as a function of ability as the NR IRT model does, but also as a function of the group an examinee is in (the misconception pattern that an examinee has) as the NR LCDM does. For the SICM model, the set of trace lines differ for each group. Examinees in a group that have a given misconception are more likely to select the incorrect alternative that corresponds to that misconception, regardless of their ability level.

Figure 6 illustrates that the SICM model provides two types of examinee parameters that each contributes to the item response probability. An examinee in any group (with any misconception pattern) can have an ability anywhere along the scale. This result means examinees in different groups may have the same ability estimate, or examinees in the same group may have different ability estimates. If two examinees have the same scored response pattern, they will have approximately the same ability estimate. However, even though two examinees may miss the exact same set of items, classification according to the misconception pattern is dependent on *why* they are missing the item (which incorrect alternative they are selecting). Thus, if two examinees' nominal response patterns differ, even if they have the same scored response and ability level, they can be classified into different groups. The classification of examinees according to misconceptions is the multidimensional feedback the SICM model provides beyond the ability estimate offer by IRT methods.

Contrasting and Altering Specifications of the SICM Model

The previous sections have demonstrated that the purpose of developing the SICM model was to combine a common measure of ability as a unidimensional construct with actionable, multidimensional feedback in the form of diagnoses of misconceptions. How to best model this phenomena requires considerable statistical considerations that will be discussed in this section.

The SICM model is neither the first nor the only model including both a continuous trait and a set of dichotomous attributes within the same model; however, it does do so in a unique manner and with a unique purpose. In the DCM literature, continuous traits have been incorporated in both the structural and measurement components of various models. As the SICM model does, several models have estimated a continuous trait simultaneously with a set of dichotomous attributes in the measurement component of the models. The Reparameterized Unified Model (RUM) incorporates a continuous completeness parameter that is hypothesized to capture abilities not specified by the Q-matrix (Hartz, 2002; DiBello & Stout, 1995). In the diagnostic classification mixture Rasch model (DCmixRM), dichotomous attributes are specified as latent covariates in a mixture Rasch model to aid in understanding the nature of latent groups detected through exploratory mixture IRT modeling (Choi, 2009). More generally, Henson, Templin, Willse and Irwin (2009) extended the LCDM to accommodate multiple continuous and discrete factors in the measurement portion of the model, which is also a feature of the General Diagnostic Model (von Davier, 2005). Other DCMs have been specified to include continuous higher-order factors in the structural component of the model. Unidimensional or multidimensional higher-order models may be of interest when modeling continuous factor(s) that may explain how attributes are related to each other (e.g., Templin & Henson, 2006).

By contrasting features of the SICM model with other relevant models, this section will establish the rational for the specifications made in the SICM model. Specifically, this section will discuss three questions of model specifications that bear significant statistical implications:

1. Should ability contribute to the probability of selecting incorrect responses?
2. Should misconceptions and overall ability be correlated?
3. Should the model account for the phenomena of guessing on multiple-choice tests?

Relationships among Ability and Misconceptions

As mentioned previously, continuous factors have been incorporated into both structural and measurement components of several psychometric models. Although Questions 1 and 2 may seem similar, they pertain to distinct components of the model. Question 1 is concerned with how the continuous ability functions in the measurement component of the model, whereas Question 2 regards the behavior of the continuous ability in the structural component of the model. Each will be discussed in the following sections.

Ability in the Measurement Component of the Model

The SICM model is more closely related to a bifactor model than to a higher-order model. As a bifactor model (Holzinger & Swineford, 1937) does, the SICM model includes the continuous factor in the measurement portion of the model through the parameterization of the item response. In contrast, higher-order models include a continuous factor in the structural portion of the model through correlations. A bifactor model estimates a general, continuous construct hypothesized to be made up of several distinct yet highly related sub-domains. Bifactor models are applicable when the interpretation of the sub-domains is of interest in addition to the general construct (Chen, West, & Sousa, 2006).

As described previously, the continuous trait, θ, is the only predictor in the SICM model for selecting the correct answer, and the set of misconceptions ($\boldsymbol{\alpha}$) are the only predictors in the model for selecting an incorrect alternative. It is debatable whether θ should also be a predictor for the probability of selecting an incorrect alternative, as it is possible that the level of an examinee's ability would also influence which incorrect alternative is chosen. Thus, the SICM model could be specified in this alternative manner.

If the SICM model did allow $\lambda_{n_{ij},\theta}$ to be present on every alternative, it could be viewed as a nominal response version of the categorical bifactor model that Henson, Templin, Willse, and Irwin (2009) specified by extending the LCDM to include a single continuous predictor. The idea behind the bifactor model is that separate abilities that comprise an overall ability exist even after the general ability has been partialled out. Henson et al. (2009) refer to the continuous ability in the categorical bifactor model as “ancillary” and suggest that it captures any ability needed to correctly respond to an item that was not specified by the Q-matrix. Because the Q-matrix for the SICM model specifies attributes as misconceptions that are only present in the incorrect alternatives, none of the ability needed to correctly respond to the item is specified by the Q-matrix. Thus, the continuous ability modeled by the SICM model may more closely resemble the general ability with respect to a unidimensional construct being measured by IRT methods; however, interpreting the continuous ability and understanding the interplay of misconceptions and a continuous trait needs further investigation.

The present specification of the SICM model is motivated by the existing literature surrounding nominal response IRT models. Formulating the SICM model where the continuous ability is measured *only* with the correct answer results in plausible statistical properties. Like Thissen and Steinberg’s (1984) Multiple-choice (MC) model, the probability of selecting an incorrect alternative is non-zero for all incorrect alternatives. In contrast, for models like Bock’s (1972) NR IRT model, the probability of selecting a single incorrect alternative approaches one as ability decreases, meaning the probability of selecting any other alternative approaches zero. The present specification of the SICM model reflects the logical notion that as ability decreases, the examinee is expected to miss the item, with some non-zero probability of selecting each of the provided incorrect alternatives. For the SICM model, unlike the MC model, the probability of selecting the correct response is always monotonically increasing with respect to ability. This feature was also desired as a monotonic relationship between ability and answering the item correctly is more reasonable. In this way, the SICM model addresses the problematic feature of the NR IRT model without resulting in the undesired feature of the MC model.

Consider again Figure 6. In the NR IRT model and the SICM model, unlike the MC model which is not pictured in Figure 6, the trace lines for the correct alternative are monotonically increasing as ability increases. Also for the SICM model, unlike the NR IRT model or the MC model, incorrect alternatives are monotonically decreasing with an upper asymptote and thus never intersect. For the NR IRT model, the most likely incorrect answer changes as a function of ability. For the SICM model, within a misconception pattern, the *order* of the probability of selecting a given incorrect alternative is dependent upon the misconceptions and invariant with respect to ability. The upper asymptote for an incorrect alternative (n_{ij}) is

$$\frac{P\left(X_{ei} = n_{ij} \middle| \boldsymbol{\alpha}_e, \theta_e\right)}{P(X_{ei} = 0|\boldsymbol{\alpha}_e, \theta_e)} \tag{23}$$

where $P(X_{ei} = 0)$ is the probability of an incorrect answer. Across misconception patterns, asymptotes for the incorrect alternatives differ, yet within a pattern, the ratio in Equation (23) remains constant at all ability levels.

If specified as a nominal response categorical bifactor model, the SICM model would behave more like the NR IRT model in these graphs. One incorrect alternative would have a unit probability of being selected as ability approached negative infinity, and trace lines for incorrect alternatives would intersect. If trace lines intersect, the model-predicted incorrect alternative that an examinee is mostly likely to select can be different among examinees with the same misconception pattern. The present specification of the SICM model is more desirable because it fixes the predicted order of incorrect alternatives an examinee will pick based upon their misconception pattern. Therefore, the SICM model was specified to have ability load only onto the correct alternative instead of onto all alternatives.

Ability in the Structural Component of the Model

In the structural component of the SICM model, the misconceptions are correlated with each other, but uncorrelated with overall ability. An alternate specification of the SICM model could allow for ability to be negatively correlated with the attributes. In practice, reasons for either specification can be provided. If misconceptions and ability are restricted to have a zero correlation, having a higher ability does not necessarily imply that an examinee does not have certain misconceptions. That is to say, examinees with a high ability may have anywhere from all to no misconceptions. Intuitively, one might expect for an examinee with a high ability to be more likely to have no misconceptions than to have all misconceptions. However, relationships of misconceptions and abilities are domain-specific, and it is plausible to think an examinee with a high ability may still have misconceptions. To provide some counterexamples to this notion, imagine the continuous construct of interest is language fluency, and a misconception is the inability to conjugate verbs. One can imagine a person who is very fluent in speaking, reading, and interpreting a language, yet consistently exhibits errors in conjugating verbs.

Perhaps more concrete is an example of physical, not mental, ability. In the following two scenarios, regard a “misconception” to be a lack of a skill. A basketball player who gets over ten rebounds and points per game, yet cannot dribble or accurately shoot three pointers, may still be a great basketball player (presence of misconceptions does not preclude the athlete from having a high overall ability). Conversely, a golfer who can drive the ball over 300 yards and putt accurately, yet cannot chip is likely not a good golfer (presence of a misconception precludes the athlete from having a high overall ability).

Estimating the correlations of ability with each misconception would allow flexibility in the model but would increase the complexity of the structural component of the model. As an entry point to vet the performance of a model of this type, the more complex model will not be estimated. If the simpler model is found to be estimable under reasonable testing circumstances, future studies can examine more complex variations of the model.

A correlation among ability and misconceptions certainly would be needed if misconceptions were defined as the opposite of a complete set of skills needed to perform highly on a test, but misconceptions should not be defined in this way. Specifying misconceptions for the SICM model differs from specifying attributes for a DCM. For DCMs, attributes specified in the Q-matrix should fully account for the underlying skills needed to answer the items on the test correctly. Otherwise, traits that are unaccounted for would be contributing to the item response. The Q-matrix for the SICM model includes ability that functions as an overarching trait needed to answer the items correctly, so misconceptions need not represent the lack of all skills required to answer the items on a test correctly. In fact, misconceptions should not represent the lack of all skills needed to answer items correctly, or they will be redundant with the continuous ability. Rather, misconceptions should represent specific errors or beliefs that cause a student to miss the item, regardless of their ability level.

For example, if attributes were delineated for the NR LCDM and then their definitions reversed to be estimated along with a continuous ability using the SICM model, it is likely the lack of skills would be highly correlated with the overall ability. This is exactly what the example item in Figure 4 did. That main purpose of using that item was to simply illustrate how to statistically specify the different models; its purpose was not to be an exemplar for defining misconceptions. For examples of items that measure true misconceptions, and for further direction on developing those kinds of items, the assessment development projects described at the beginning of this chapter are good references.

Guessing Parameter for SICM

One theoretically undesirable feature of the SICM model as expressed in Equation (20) is that the probability of selecting the correct response approaches the limit of zero as ability decreases. Because multiple-choice items provide alternatives from which examinees choose an answer, examinees who do not know the answer can still guess to get the item correct. Guessing occurs more frequently in low-ability examinees. These examinees unexpectedly answer difficult items correctly, producing aberrant responses that result in misfit for a model that does not account for guessing.

The 2-PL IRT model and the NR IRT model are two commonly used IRT models that do not account for guessing. Versions of these models have been created that estimate an additional parameter for each item to quantify the affect guessing has on the item. The 2-PL IRT model was extended to the three-parameter logistic (3-PL) IRT model (Birnbaum, 1968). The 3-PL IRT model adds an additional item parameter to provide a lower-asymptote for the probability of a correct response for an item. Although the asymptote is desired theoretically because it reflects an empirical theory widely agreed upon —that examinees can guess on multiple-choice items, it is difficult to estimate in practice. A widely-used IRT program, BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), puts a prior on the estimator for the guessing parameter in the 3-PL model by default, indicating difficulty in practice of estimating the parameter. Anytime additional parameters are added to a model, estimation difficulty increases, but in this case, estimation is also impacted because the form of the model is changed. Although named the three-parameter *logistic* model, this model no longer has a logistic form. Similar to the 3-PL extension of the 2-PL IRT model, Thissen and Steinberg's (1984) MC model extended Bock's NR-IRT model by including a parameter in the model that represents the probability that an examinee does not know the answer and supplies a guess. The 2-PL IRT and the NR IRT model are used more commonly than either of their extensions, likely because they are easier to estimate.

An alternative formulation of the SICM model was developed and will be used to provide a lower-asymptote for the probability of a correct response *without* adding an additional parameter to the model. To facilitate a visual comparison of the original and new form of the SICM model, recall the original formulation of the SICM model in Equation (20) is

$$\log(\frac{P(X_{ie}=n_{ij}|\alpha_e,\theta_e)}{P(X_{ie}=n_{iJ}|\alpha_e,\theta_e)})= \lambda_{n_{ij},0} - \lambda_{n_{iJ},\theta}(\theta_e) + \boldsymbol{\lambda}_{n_{ij}}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}). \tag{24}$$

The new formulation of the SICM model that accounts for examinees' propensities to guess in the event they do not know the answer and do not have a line of reasoning that corresponds to a misconception measured by one of the possible alternatives is

$$\log(\frac{P(X_{ie}=n_{ij}|\alpha_e,\theta_e)}{P(X_{ie}=n_{iJ}|\alpha_e,\theta_e)})= \lambda_{n_{ij},0} - \exp(\lambda_{n_{iJ},\theta}(\theta_e)) + \boldsymbol{\lambda}_{n_{ij}}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}). \tag{25}$$

The difference in Equations (24) and (25) is the ability portion of the model is now exponentiated. The intercept of the model is now interpreted as the logit that an examinee with an extremely low ability who possesses no misconceptions will choose alternative n_{ij}. Holding other parameters constant, as ability decreases, the value of $\exp(\lambda_{n_{iJ},\theta}(\theta_e))$ decreases, and the logit of selecting the correct answer decreases, satisfying the monotonicity assumption for the model. As ability approaches negative infinity, $\exp(\lambda_{n_{ij},\theta}(\theta_e))$ approaches 0, meaning the logit approaches $\lambda_{n_{ij},0}+\boldsymbol{\lambda}_{n_{ij}}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}})$, yielding a lower asymptote of the probability of selecting the correct response where

$$\lim_{\theta\to\infty}(\pi_{n_{iJ}|\theta_e,\alpha_e})=\frac{\exp(-1)}{\sum_{j=1}^{J}\exp(\,\lambda_{n_{ij},0}+\boldsymbol{\lambda}_{n_{ij}}^{T}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}))}. \qquad (26)$$

This correction results in a more realistic model of the item response, without also resulting in an increased difficulty in estimation due to an increased number of item parameters to be estimated, as is commonly encountered when using the 3-PL IRT model. Analogous adjustments can also be made in the 2-PL IRT model and Bock's NR-IRT model to provide a pseudo-lower asymptote. These models will be the foci of future research.

Figure 7 illustrates the lower asymptote versions of the SICM model and the 2-PL IRT model, denoted as the SICM* model and the 2-PL IRT* model. Throughout the remainder of this dissertation, the name of a model specified with a lower asymptote as the SICM model is in Equation (25) will be followed by an asterisks (e.g., SICM*) to distinguish between the two versions of the models. First, consider the SICM models on the left hand side of the figure. The trace lines are for a hypothetical item that has three alternatives. The same misconception is measured in each of the two incorrect alternatives. The trace line for the correct answer is denoted with + lines. The trace line for the misconception with the higher and lower intercept are denoted with upward and downward facing triangles, respectively. These trace lines were for the group of examinees who did not have the misconception measured by the alternatives. For the SICM model without the lower asymptote, the loading for ability was 1.1 and the intercepts for the incorrect alternatives were -0.5 and 0. For the SICM* model, the loading for ability was 0.55 and the intercepts were -0.5 and 0.50. Although the parameters from the two models are not directly comparable due to the exponentiation, an item for which an examinee with an average ability has about a .50 probability of answering the item correctly is illustrated for each model.

Although for each version of the SICM model an examinee at the upper end of the ability scale ($\theta_e = 3$) has almost unit probabilities of responding correctly, the probability an examinee at the lower end of the ability scale ($\theta_e = -3$) answers the item correctly are very different for the two models. For the SICM model, an examinee with $\theta_e = -3$ has almost a zero probability of answering the item correctly, but for the SICM* model, the examinee has approximately .35 probability of answering the item correctly. Note in the SICM* model the upper asymptotes for the incorrect alternatives also remain intact; the exponentiation did not impact other features of the SICM model previously described.

The two right hand graphs are of the 2-PL IRT and 2-PL IRT* models to further illustrate the use of this type of lower asymptote. The trace lines for the incorrect responses in the 2-PL IRT models are denoted with x lines. As in the SICM model, the correct alternative has a non-zero probability of being selected by examinees with low abilities, resulting in the response probabilities for the lower ability extremes being very different in the two versions of the model.

SICM Framework: Concluding Remarks

It is important to acknowledge that the SICM model exists within a specific mathematical framework, not as a fixed mathematical equation. Although the framework has a very distinct purpose, the mathematical equation may be altered to more closely adapt to the context within which the model may be applied. In applications of the SICM model, specific empirical theories may drive the statistical specifications of the model. For example, there may be documented literature or empirical evidence that supports the notion that misconceptions are correlated to overall ability in a given domain, and the specification of the SICM model may change to allow the continuous ability and the categorical misconceptions be correlated in the structural component of the model. The discussion in the previous section sought to highlight considerations made in developing the model in order to bring those considerations to the forefront of any subsequent practical application of the model.

Estimation of the SICM Model

The SICM* model was estimated using a Markov Chain Monte Carlo (MCMC) algorithm written in Fortran. Using the specification of the model by Equation (20), the SICM model is estimable using Mplus Version 6.1 (Muthén & Muthén, 1998-2010). However, Mplus cannot estimate the SICM* model with the exponentiation, so writing a unique estimation algorithm using Fortran was necessary. Other advantages to using MCMC methods in Fortran include providing a more efficient way to estimate the model parameters in comparison to Marginal Maximum Likelihood estimators from Mplus. The following sections detail the specific steps of the algorithm used to estimate the SICM* model.

MCMC Estimation Algorithm

Using the $E\ x\ I$ matrix of E examinees responses to I observed items ($\mathbf{X}$), the goal is to estimate the parameters that define the SICM* model (i.e., a set of item ($\boldsymbol{\Lambda}$) and a set of structural ($\boldsymbol{\Gamma}$) parameters) and the parameters that describe latent traits of the examinees (i.e., the misconception patterns ($\boldsymbol{\alpha}$) and examinee abilities ($\boldsymbol{\theta}$)). To estimate these parameters, a Bayesian technique that employees an MCMC estimation algorithm called Gibbs sampling was implemented to sample from the posterior distributions for the parameters. The posterior distribution is the conditional probability distribution of all unobserved variables, given the observed variables. From the posterior distributions, quantities of interest for describing parameter estimates, such as means and variances, can be determined. Using Bayes Theorem, the posterior distribution of the parameters of the SICM* model, given the observed data, are defined as

$$\pi(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta}) \;=\; P(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta} \,|\boldsymbol{X}) \;=\; \frac{P(\boldsymbol{X}|\, \boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\theta})\; P(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\alpha},\boldsymbol{\theta})}{P(\boldsymbol{X})} \tag{27}$$

where $P(\boldsymbol{X}|\, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is the probability of the data, given the model parameters, and defined by the likelihood function for the SICM* model, and $P(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is the joint probability of the model parameters. The term $P(\boldsymbol{X})$ is the marginal distribution of the observed data, often referred to as the normalizing constant. In theory, estimates for each individual parameter can be determined by sampling directly from this posterior distribution. The posterior distribution of the SICM* model is made complex by its multivariate nature and its composition of both continuous and discrete variables. Direct sampling from the SICM* model's non-standard posterior distribution would be difficult, if not impossible.

MCMC simulation methods can be used when it is not possible or is inefficient to sample directly from the posterior distribution, as is the case for the multivariate posterior distribution in Equation (27). MCMC estimation methods repeatedly draw a random value from an approximate posterior distribution. Each draw is evaluated by considering the likelihood that the draw came from the target posterior distribution. The evaluations allow for adjustments to be made for the approximate distribution until it converges to a stable target distribution. For MCMC estimation, the draws that are accepted or retained at each of the T stages of sampling represent entries in a Markov chain. A Markov chain is a sequence of T random variables, $\tau^1, \tau^2, \ldots, \tau^T$, such that the distribution for any variable of the sequence at stage t (τ^t) is solely dependent upon the variable in the previous step (τ^{t-1}).

Metropolis-Hastings Algorithm

The Metropolis–Hastings algorithm is an MCMC method for governing the values that are accepted for the Markov Chain by defining a transitional kernel that will direct the chain to converge to the target distribution for a large T. The algorithm defines the probability the candidate parameter proposed at iteration t (τ^*) is accepted (τ^t will be set to equal τ^*) over the previous value of the parameter at iteration $t-1$ (τ^t will be set to equal τ^{t-1}) as

$$min(1, r) \tag{28}$$

where r equals

$$\frac{\pi^*(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta})\ \mathrm{Q}(\ {\tau_t}^{t-1} |\ {\tau_t}^*)}{\pi^{t-1}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta})\ \mathrm{Q}(\ {\tau_t}^* |\ {\tau_t}^{t-1})} \tag{29}$$

where $\pi(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is the joint posterior distribution as defined in Equation (27). The density $\pi^*(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is the distribution given τ_t equals ${\tau_t}^*$ and all other parameters are held constant; similarly, $\pi^{t-1}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is the distribution given τ_t equals ${\tau_t}^{t-1}$ and all other parameters are held constant. Because the normalizing constant for $\pi(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is a function of the observed data, and not a function of the estimated parameters, it is the same for $\pi^*(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ or $\pi^{t-1}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta})$. Therefore, the normalizing constant can be factored from this equation and does not need to be known. To estimate the model parameters, using this algorithm, it is sufficient to know that the posterior distribution is proportional to the product of the conditional probability of the data given the model parameters and the joint probability of the parameters.

This proportional relationship is expressed as

$$\pi(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta}) \; = \; P(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \theta \; | X) \; \propto P(X | \; \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \theta) \; x \; P(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \theta). \qquad (30)$$

Therefore, Equation (29) simplifies to

$$\frac{P^{*}(\boldsymbol{X} | \; \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \theta) \; P^{*}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \theta) \mathrm{Q}(\tau^{t-1} | \tau^{*})}{P^{t-1}(\boldsymbol{X} | \; \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \theta) \;\; P^{t-1}(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \theta) \mathrm{Q}(\tau^{*} | \tau^{t-1})} \qquad (31)$$

where $P(\boldsymbol{X} | \; \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \theta)$ is the probability of the data, given the model parameters, and $P(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \theta)$ is the joint probability of the model parameters. P^{*} $(\cdot)$ indicates the probability given τ^{t} equals τ^{*} and all other parameters are held constant; similarly P^{t-1} $(\cdot)$ is the probability given τ^{t} equals τ^{t-1} and all other parameters are held constant.

The term $\mathrm{Q}(\tau^{*} | \tau^{t-1})$ is the candidate generating density, the density of the distribution from which the candidate parameter (τ^{*}) is drawn given the previous value of the parameter (τ^{t-1}). The term $\mathrm{Q}(\tau^{t-1} | \tau^{*})$ is the probability density of the previous parameter given the value of the candidate parameter. For some parameters in the SICM* model (i.e., item parameters), non-symmetric proposal distributions are used. These non-symmetric distributions are used due to boundaries placed on the parameter space. For example, main effects for misconceptions cannot be negative. Experience with MCMC algorithms suggest that parameter boundaries make estimation difficult due to symmetric proposal distributions having a large probability mass occur at the boundary. The large mass causes more proposed values to be at the boundary than a non-symmetric distribution, slowing convergence of the algorithm. For estimation of the SICM* model, when the proposal distribution was not symmetric, a moving window proposal distribution was used (Henson & Templin, 2003). A moving window proposal distribution draws τ^* from a uniform distribution with bounds UB and LB ($U(LB, UB)$) where

$$LB = max\left(\tau^{t-1} - \frac{w}{2}, a\right); UB = min\left(\tau^{t-1} - \frac{w}{2}, b\right). \tag{32}$$

The parameter w controls the width of the sampling interval. The parameters a and b constrain the sampling intervals to lower and upper boundaries, respectively. The value of $\mathrm{Q}(\tau^*|\tau^{t-1})$ is calculated as the height of the density of the uniform distribution $U(LB, UB)$:

$$\mathrm{Q}(\tau^*|\tau^{t-1}) = \frac{1}{UB - LB}. \tag{33}$$

The value of $\mathrm{Q}(\tau^{t-1}|\tau^*)$ is calculated as in Equation (33) where LB and UB instead equal

$$LB = max\left(\tau^* - \frac{w}{2}, a\right); UB = min\left(\tau^* - \frac{w}{2}, b\right). \tag{34}$$

The values of τ that are accepted at each of the T stages, $\tau^1, \tau^2, \dots, \tau^{\mathrm{T}}$, comprise the Markov chain. Because the chain may require initial stages to find the target posterior distribution, the first R entries of the Markov chain are discarded. Stages 1 thought R are regarded as the burn-in period, where the value of R is large enough for the chain to reach stationarity. Stationarity is reached when the chain has converged to the target posterior distribution. Stages $R + 1$ through T provide samples from the target posterior distribution to accurately describe its shape and moments.

The Metropolis-Hastings algorithm eliminates the need to know the complete posterior distribution, which significantly reduces computation. Gibbs sampling can further reduce computational demands of complex integration. Gibbs sampling is a special case of the Metropolis-Hastings algorithm that generates random values by sampling from the distribution of a single parameter conditional on the values of all other parameters. This sampling method eliminates the need to integrate over the joint posterior distribution to yield marginal distributions for each parameter of interest.

Using Gibbs sampling, each parameter is updated individually, meaning the marginal distribution $P(\tau^*)$ of a parameter replaces the joint posterior distribution of all model parameters $P(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ in Equation (31). When updating the k^{th} parameter in a set of Π model parameters, the value of *r* expressed in Equation (29) simplifies to

$$\frac{L(\Pi_{p=1,k-1}^{t}, \tau_k = {\tau_k}^{*}, \Pi_{p=k+1,P}^{t-1}) P({\tau_k}^{*}) \mathrm{Q}({\tau_k}^{t-1} | {\tau_k}^{*})}{L(\Pi_{p=1,k-1}^{t}, \tau_k = {\tau_k}^{\mathrm{t-1}}, \Pi_{p=k+1,P}^{t-1}) P({\tau_k}^{t-1}) \mathrm{Q}({\tau_k}^{*} | {\tau_k}^{t-1})} \tag{35}$$

where $L(\cdot)$ is the likelihood of the model, where the first $k - 1$ parameters have already been updated during the t^{th} iteration ($\Pi_{p=1,k-1}^{t}$), and the $k + 1$ through P^{th} item parameters have not yet been updated and retain their values from the $(t - 1)^{th}$ iteration ($\Pi_{p=k+1,P}^{t-1}$)). The term $P({\tau_k}^{*})$ is the marginal distribution, or the prior distribution, of the proposed parameter. To begin the estimation process, starting values are assigned to every parameter and prior distributions of each parameter in the model are specified so that all conditional distributions are known. For estimating multivariate models, the posterior distribution of each parameter is updated at each stage and each parameter has its own Markov chain.

To illustrate this procedure, consider the p^{th} item parameter, λ_p, for item i. The probability a newly proposed parameter at stage t (${\lambda_p}^{*}$) is accepted over the previous value of the parameter at stage $t - 1$ (${\lambda_p}^{t-1}$) is defined in Equation (28) where r equals

$$\frac{\prod_{e=1}^{E}(\prod_{j=1}^{J}\pi_{m_{ij}|\boldsymbol{\alpha}_{e,}\theta_e}^{[x_{ei}=m_{ij}]}\,(\Lambda_{l=1,p-1}^{t},\lambda_p=\lambda_p{}^{*},\Lambda_{p=p+1,L}^{t-1})\,P(\lambda_p{}^{*})\mathrm{Q}(\lambda_p{}^{t-1}|\lambda_p{}^{*})}{\prod_{e=1}^{E}(\prod_{j=1}^{J}\pi_{m_{ij}|\boldsymbol{\alpha}_{e,}\theta_e}^{[x_{ei}=m_{ij}]}\,(\Lambda_{l=1,p-1}^{t},\lambda_p=\lambda_p{}^{\mathrm{t-1}},\Lambda_{l=p+1,L}^{t-1})\,P(\lambda_p{}^{t-1})\mathrm{Q}(\lambda_p{}^{*}|\lambda_p{}^{t-1})}. \tag{36}$$

The term $\prod_{e=1}^{E}(\prod_{j=1}^{J}\pi_{m_{ij}|\boldsymbol{\alpha}_{e,}\theta_e}^{[x_{ei}=m_{ij}]}\,(\Lambda_{l=1,p-1}^{t},\lambda_p=\lambda_p{}^{*},\Lambda_{p=p+1,L}^{t-1})$ defines the likelihood of observing all E examinees' responses to item I given each examinee's attribute pattern, ability, and the set of item parameters. Calculating the full likelihood of the model, a process that would use a product across all items, is redundant. The terms in the likelihood that correspond to items other than the current item will be equal and hence will not influence the value of the expression due to factorization (i.e., they cancel in the Metropolis-Hastings ratio). The item response probability ($\pi_{n_{ij}|\boldsymbol{\alpha}_{e,}\theta_e}$) is a function of the set of item parameters ($\boldsymbol{\Lambda}$). When updating the p^{th} item parameter, $p-1$ parameters have already been updated during the t^{th} iteration ($\boldsymbol{\Lambda}_{l=1,p-1}^{t}$). In contrast, the $p+1$ through L^{th} item parameters have not yet been updated and retain their values from the $(t-1)^{th}$ iteration ($\boldsymbol{\Lambda}_{l=p+1,L}^{t-1}$). The term $P\left(\lambda_p{}^{*}\right)$ is the prior probability for the proposed value of λ_p for the t^{th} iteration; similarly, $P(\lambda_p{}^{t-1})$ is the prior probability for the value of λ_p at the $(t-1)^{th}$ iteration. If the proposal value is accepted, $\lambda_p{}^{*}=\lambda_p{}^{t}$; if it is rejected, $\lambda_p{}^{\mathrm{t-1}}=\lambda_p{}^{t}$. Accepted values, $\lambda_p{}^{\mathrm{R+1}},\lambda_p{}^{R+2},\ldots,\lambda_p{}^{\mathrm{T}}$, comprise the Markov chain.

Gibbs Sampling Steps for Individual Parameters in the Model

To begin the estimation algorithm, starting values for each parameter were chosen at random. The choice of starting value was arbitrary and did not impact the final destination of the chain. In each stage, the algorithm first updated all item parameters one-by-one, then updated the structural parameters individually, and finally updated the examinee parameters. For each type of parameter, the following sections detail the steps of the algorithm by defining the value of r from

Equation (28), providing the distributions from which candidate parameters were drawn $\mathrm{Q}(\tau^{*}|\tau^{t-1})$, and specifying the prior distribution $P(\tau^{*})$ of the parameter used for estimation.

Item Parameters

Item parameters included the intercepts ($\lambda_{n_{ij},0}$), misconception main effects for an attribute a ($\lambda_{n_{ij},1,(a)}$), the interactions between pairs of misconceptions a and b present on the same alternative ($\lambda_{n_{ij},2(a,b)}$), and the main effects for ability ($\lambda_{n_{iJ},\theta}$). For item parameters the value of *r* can be found in Equation (36).

Intercepts

For the intercept corresponding to each incorrect alternative, $\lambda^{*}_{0,n_{ij}}$ was drawn from $N(\lambda^{t-1}_{0,n_{ij}}, .1)$. An (improper) uniform prior across the space of real numbers was used. An uninformative prior was used to have little impact on the estimation process. Uninformative priors also make the MCMC estimator parallel to marginal maximum likelihood estimators. Thus, the choice of uninformative priors for the item intercepts, and other item parameters, provides a bridge between the two estimation techniques.

Main Effects for Misconceptions

For the main effect for a misconception a present in an alternative, $\lambda^{*}_{n_{ij},1(a)}$ was drawn from the moving window proposal distribution ($\lambda^{*}_{n_{ij},1(a)} \sim U(LB, UB)$), with UB and LB as defined in Equation (34). For this equation, $\tau^{t-1} = \lambda^{t-1}_{n_{ij},1(a)}$, w = .1, a = 0 (defining a lower bound), and *b* = 10,000 (an ill-defined lack of an upper bound). An (improper) uniform prior across the space of real numbers was used.

Interactions of Misconceptions

For interactions of misconceptions that occur when an alternative measures two misconceptions (a and b), the proposal was drawn from $\lambda^{*}_{n_{ij},2,(a,b)} \sim U(LB, UB)$ as defined in Equation (34). For this equation, τ^{t-1}, w = .1, $a = -\lambda^{min}_{n_{ij},1(a)}$, and b = 10,000 (an ill-defined lack of an upper bound). Here, $\lambda^{min}_{n_{ij},1(a)}$ denotes the smallest misconception main effect for alternative n_{ij}. This specification restricted the absolute value of the lower bound of the interaction to be less than the main effect for any misconception present in that alternative. This restriction ensured the monotonicity assumptions for misconception main effects were satisfied. An (improper) uniform prior across the space of real numbers was used.

Main Effects for Ability

For the loading for ability for each item, the proposal was drawn from $\lambda^{*}_{n_{iJ},\theta} \sim U(LB, UB)$ as defined in Equation (34). For this equation, τ^{t-1} = $\lambda^{t-1}_{n_{iJ},\theta}$, w = .1, a = 0 (defining a lower bound), and b = 10,000 (an ill-defined lack of an upper bound). An (improper) uniform prior across the space of real numbers was used.

Structural Parameters

The full structural model includes 2^{A} structural parameters, which model the probability any profile of misconceptions are present in the sample. Using a log-linear model, these parameters include an intercept, A main effects, and $\binom{A}{x}$ x-way interactions, where $2 \leq x \leq A$. Consider a given structural parameter, γ_p, in the set of all G structural parameters ($\boldsymbol{\Gamma}$). The probability ${\gamma_p}^{*}$ was accepted over ${\gamma_p}^{t-1}$ is defined in Equation (28) where the value of

r is now

$$r = \frac{\prod_{e=1}^{E}(\upsilon_p(\Gamma_{g=1,p-1}^{t}, \gamma_p = {\gamma_p}^{*}, \Gamma_{g=p+1,G}^{t-1}))\ P({\gamma_p}^{*})\mathrm{Q}({\gamma_p}^{t-1}|{\gamma_p}^{*})}{\prod_{e=1}^{E}(\upsilon_p(\Gamma_{g=1,p-1}^{t}, \gamma_p = {\gamma_p}^{\mathrm{t-1}}, \Gamma_{g=p+1,G}^{t-1}))\ P({\gamma_p}^{t-1})\mathrm{Q}({\gamma_p}^{*}|{\gamma_p}^{t-1})}. \tag{37}$$

The term υ_p is the marginal probability that examinee e has attribute pattern p. The proportion of examinees with pattern p, υ_p , is a function of the set of structural parameters ($\boldsymbol{\Gamma}$). The set of parameters $\Gamma_{g=1,p-1}^{t}$ have already been updated in this stage, and $\Gamma_{g=p+1,G}^{t-1}$ have not.

For each of the 2^Astructural parameters γ^* was drawn from $N(\gamma^{t-1}, .1)$. An (improper) uniform prior across the space of real numbers was used.

Examinee Parameters

The SICM model has two kinds of examinee parameters, the examinee attribute pattern ($\boldsymbol{\alpha}_e$) and the examinee ability estimate(θ_e).

Examinee Attribute Pattern

The acceptance probability for a proposed value for an examinee's ability, $\boldsymbol{\alpha}_e^*$ is a function of r defined as

$$r = \frac{\prod_{i=1}^{I}\prod_{j=1}^{J}(\pi_{n_{ij}|\boldsymbol{\alpha}_{e,}\theta_e}^{[x_{ei}=n_{ij}]}\,(\Lambda^t))\,P(\alpha_e^*)\mathrm{Q}(\alpha_e^{t-1}|\alpha_e^*)}{\prod_{i=1}^{I}\prod_{j=1}^{J}(\pi_{n_{ij}|\boldsymbol{\alpha}_{e,}\theta_e}^{[x_{ei}=n_{ij}]}(\Lambda^t))\,P(\alpha_e^{t-1})\mathrm{Q}(\alpha_e^*|\alpha_e^{t-1})} \tag{38}$$

The term $\prod_{i=1}^{I}\prod_{j=1}^{J}(\pi_{n_{ij}|\boldsymbol{\alpha}_{e,}\theta_e}^{[x_{ei}=n_{ij}]}\,(\Lambda^t))$ is the likelihood of observing examinee e's responses to all I items, given examinee e's attribute pattern, ability, and the set of current item parameters(Λ^t), which have already been updated in stage t.

The proposed attribute pattern for each examinee $\boldsymbol{\alpha}_e^*$ was drawn from a multivariate Bernoulli distribution defined by the proportion of examinees having each attribute pattern at the previous step, $MVB\left(\boldsymbol{v}_p^t\right)$, where $\boldsymbol{v}_p^t$ was set by the parameters of the structural model earlier in stage t. This distribution also served as an empirical prior distribution for $\boldsymbol{\alpha}_e^*$, mirroring common practice in factor analytic, mixed effects, and item response models estimated with MCMC. Because $\boldsymbol{\alpha}_e^*$ was drawn from the prior distribution, $P(\alpha_e^*)$ cancelled with $\mathrm{Q}(\alpha_e^*|\alpha_e^{t-1})$, and $\mathrm{Q}(\alpha_e^{t-1}|\alpha_e^*)$ cancels with $P(\alpha_e^{t-1})$, simplifying Equation (38) to a function of only the item response likelihood of an examinee.

Examine Ability Estimate

Similarly, the acceptance probability for a proposed value for an examinee's ability, θ_e^* was a function of r defined as

$$r = \frac{\prod_{i=1}^{I}\prod_{j=1}^{J}(\pi_{n_{ij}|\boldsymbol{\alpha}_{e,}\theta_e}^{[x_{ei}=n_{ij}]}(\Lambda^t))\,P(\theta_e^*)\mathrm{Q}(\theta_e^{t-1}|\theta_e^*)}{\prod_{i=1}^{I}\prod_{j=1}^{J}(\pi_{n_{ij}|\boldsymbol{\alpha}_{e,}\theta_e}^{[x_{ei}=n_{ij}]}(\Lambda^t))\,P(\theta_e^{t-1})\mathrm{Q}(\theta_e^*|\theta_e^{t-1})} \tag{39}$$

The ability for each examinee θ_e^* was drawn from $N(\theta_e^{t-1},.1)$. A standard normal, $N(0,1)$, prior was used (i.e., $P(\theta_e^*) = \frac{1}{\sqrt{2\pi}}exp(\frac{-(\theta_e^*)^2}{2})$ to set the scale for ability.

Conclusions

In this chapter, a new psychometric model was presented that provides multidimensional feedback to students in addition to an overall ability estimate. This multidimensional feedback is in the form of classifications of examinees according to the misconceptions they have. A practical need for the model was demonstrated, the statistical specifications of the model were delineated, considerations when applying the model were discussed, and the estimation procedure for the model was presented in step-by-step detail. The next chapter examines whether the estimation algorithm accurately estimates the parameters of the model in a simulated context. The final chapter will analyze an existing data set to demonstrate an application of the model in practice.

CHAPTER 4

SIMULATION STUDY

Chapter 2 provided a review of some existing psychometric models from item response theory (i.e., Rasch, 2-PL, NR IRT, and MIRT models) and diagnostic classification modeling (i.e., LCDM, C-RUM, and NR-LCDM). The delineation of these models was followed by Chapter 3 which specified a new psychometric model: the Scaling Individuals and Classifying Misconceptions (SICM) model, which combines features from a unidimensional NR IRT model and the NR LCDM. The chapter illustrated the dual-purposes of the model: scaling examinees and diagnosing misconceptions. Chapter 3 also detailed the estimation procedure for the model.

The potential usefulness of the SICM model depends on its statistical properties and estimation capabilities, which have not yet been examined. The present chapter outlines a simulation study to assess the estimation algorithm for the SICM model. The estimation accuracy of its item, examinee, and structural parameters and the reliability of examinee parameter estimates will be evaluated.

Simulation Study Purpose

The SICM model is a complex model. The complexity is due to the large number of parameters, and also to the different types of parameters (i.e., continuous and categorical), that are estimated for the model. A simulation study is a study in which the researcher specifies the model parameters and examinee parameters so their true values are known. Using these known values, a sample of data is generated, and then the model is estimated using the data set. The study can evaluate how well the model can estimate the parameters because the known true values of each parameter can be compared to the estimated values. The goal of this simulation study is to answer open questions about the SICM model that must be addressed before the model can be used in practice. For example, information needs to be gathered about the performance of the model under typical testing situations. To develop an assessment to be analyzed by the SICM model, practical limitations of the model must be known. Researchers need to know, for example, how many items or examinees are needed to estimate a given number of misconceptions along with an overall ability. Also, the interplay of a continuous ability and a set of categorical misconceptions within a single model must be better understood. Specifying continuous and categorical variables in the measurement part of a nominal response psychometric model has never been tried in this manner before. We need to understand how one type of variable will affect the estimation of the other within the model (e.g., whether the effect of one type of variable will dominant or mask the other's effect).

Simulation Study Design

The simulation study was designed to mirror both ideal and realistic situations under which the SICM model might be used. In ideal situations, the model has an abundance of information with which to estimate the model (e.g., many items, many examinees, strong effects of parameters). Under these conditions, the SICM model should readily recover the true values of the parameters. These conditions afford the opportunity to determine if the estimation algorithm was accurate and to glean information about the model when used on carefully constructed tests with large samples. Realistic conditions tried to emulate a variety of testing scenarios that may be found in practice (e.g., limited number of items and test-takers, range of weak to strong effects). These conditions provided information about settings that may be familiar to researchers seeking to apply the model on a smaller scale.

Table 3 provides an outline of the design of the simulation study. The study had five manipulated factors: sample size, test length, number of misconceptions, magnitude of main effects, and correlation among attributes. The design fully crossed the 2 attribute, 2 test length, 2 sample size, 4 main effect, and 2 correlation conditions. Fifty replications were made for each of the 64 conditions. The following sections provide further details of the simulation conditions.

Test Length

Test lengths of 30 and 60 items were used. The test length of 30 items was considered to be an average test length and was a test length of interest to current research projects in this field (e.g., the Diagnosing Teacher's Multiplicative Reasoning Project; Izsak et al., 2010). The long test length of 60 items was used in order to investigate a test length that provided ample data for the estimation of the attribute profiles and examinee ability. This length is comparable to an end-of-course test (EOCT) that is federally mandated for states to administer (e.g., state of Georgia Mathematics I and II EOCT).

The length of the test influences the number of times an attribute can be measured on a given test. For this study, each simulated item had four alternatives. Each alternative was specified to measure one or two attributes (or misconceptions); therefore, no three-way or other higher-order interactions were modeled. A balanced Q-matrix was used, meaning individual attributes were measured with the same number of alternatives and items. For the 3-misconception, 30-item conditions, each attribute was measured by 34 alternatives in 21 items. For this condition, on average, 2.1 attributes were measured per item and 1.13 attributes were measured per alternative. Table 4 contains the Q-matrix that was used for this condition. For the 3-misconception, 60-item conditions, every item in the Q-matrix in Table 4 appeared twice, meaning attributes were measured by twice as many alternatives and items, but the average attribute measured per item and per alternative remained the same.

For the 6-misconception, 30-item conditions, each attribute was measured by 17 alternatives in 11 items. For this condition, on average, 2.2 attributes were measured per item and 1.13 attributes were measured per alternative. Table 5 contains the Q-matrix that was used for this condition, although the entries for Alternative D are not displayed as they were in Table 4. For every item, *D* was the correct answer and was only measured by the continuous ability (i.e., the entry for the θ column was the only non-zero entry in each row corresponding to the correct answer). As before with the previous pair of test length conditions, when the test length was doubled (in the 6-misconception, 60-item condition) every item in the Q-matrix in Table 5 appeared twice. Again, this specification doubled the number of times each misconception was measured, but maintained an equal complexity of the items for the 6-misconception conditions. Specifying the Q-matrices in this way prevented confounding the effects of longer tests with the effects of more complex Q-matrices.

Sample Size

Templin and Bradshaw (under review) recommended that the LCDM with dichotomous data be used instead of the NR LCDM if samples of less than 1,000 examinees are available. Because the SICM model estimates a continuous ability in addition to the parameters estimated in the NR LCDM, this recommendation that the NR LCDM be used only for large-scale applications is likely to apply for the SICM model as well. For that reason, larger samples of 3,000 and 10,000 examinees were simulated.

Simulations from nominal response IRT models also provided some information for sample size considerations. DeMars (2003) found when estimating the NR IRT model with a sample of 2,400 examinees, the root mean squared error (RMSE) values were less than 0.10 for item parameters. These results came from items with three alternatives, and DeMars (2003) acknowledged that because many factors, like the number of alternatives per item, play into the specification of the model and test, it is difficult to provide general recommendations for sample size requirements. de Ayala and Sava-Bolesta (1999) suggest the ratio of the sample size to the total number of item parameters on the assessment should be at least 10:1 for the nominal response IRT model. Because there are a larger number of item parameters per item for the SICM model than for the NR IRT model, this ratio becomes difficult to reach. Even if alternatives have simple structure, meaning each incorrect alternative measures one misconception, the model has $2J_i - 1$ parameters per item and $I(2J_i - 1)$ total item parameters. Namely, each item has $(J_i - 1)$ intercepts, one ability main effect, and $(J_i - 1)$ misconception main effects. The item design for the test may also be more complex; the model allows an alternative to measure up to A misconceptions. The number of parameters for an item then must include up to $(J_i - 1)(A\ )$ misconception main effects, and $(J_i - 1)\binom{A}{2}$ misconception two-way interactions. Higher-order interactions can also be included, but were not for this simulation.

For the assessments in this simulation, each incorrect alternative measured either one or two misconceptions and J_i = 4 (for every item). The number of item parameters for the 30-item conditions were 240 and 256, for the 3- and 6- attribute conditions, respectively. The number of item parameters for the 60-item conditions were 474 and 490, for the 3- and 6- attribute conditions, respectively. The 3,000 sample supplied at least a 10:1 item parameter-to-examinee ratio for the 30 item test, and the 10,000 sample did so for the 60 item test.

In practice, for a national company seeking to diagnose misconceptions a sample size of 3,000 examinees seemed reasonable. The sample of 3,000 also approximates the number of students a medium-sized county may have per grade level for implementing county-wide benchmark testing. Although the larger sample of 10,000 was much lower than the number of students per grade level per state, it was a larger sample that would offer results about the utility of the SICM for large-scale assessment, such as a state-wide type of assessment.

Number of Attributes

Two attribute-number conditions were used: a three attribute and a six attribute condition. Three and six attributes are within a usual range of attributes found in current applications of IRT-based DCMs (Rupp & Templin, 2007).

Absolute and Relative Magnitude of Main Effects

Of particular interest in estimating the SICM model was the interplay between the overall ability and the misconceptions. In the simulation study, the strengths of the continuous trait and categorical traits were varied systematically. The relative and absolute magnitudes of the main effects for these latent traits were manipulated to create four conditions. The values chosen were selected to mirror what might be found in practical applications, based upon previous work with the NR LCDM (Templin & Bradshaw, under review). For two conditions, the relative magnitudes were different. The first condition was one in which ability had a large effect (λ_θ drawn from a uniform distribution, $U(.6,.8)$) and misconceptions had a small effect

($\lambda_{1\,(a)}$ drawn from a uniform distribution, $U(.75, 1.25)$). In a sense, a more extreme specification in this manner may suggest a scenario where the nominal response IRT model would be more appropriate. In the second condition, misconceptions had a large effect ($\lambda_{1\,(a)}$ drawn from $U(1.75, 2.25)$) and ability had a small effect (λ_{θ} drawn from $U(.3, .6)$). A more extreme specification of this sort may suggest the NR-LCDM would be more appropriate. For the remaining two conditions, the relative magnitudes were comparable, but two different absolute magnitudes were tested. For the third condition, both ability and misconceptions had small effects, providing a condition that should to be difficult to estimate. For the last condition, both had large effects, providing a condition that should be more easily estimated.

Figures 8 and 9 illustrate the behavior for a sample item from all four of these conditions. In this example item, the same misconception is measured by incorrect Alternative A (dark gray trace line) and Alternative B (light gray trace line). The main effect for the misconception is always larger for *A* than *B*. The trace lines under these conditions are displayed for examinees who do not and do possess the measured misconception in Figures 8 and 9, respectively. In Figure 8, which depicts a situation in which examinees do not possess the misconception, conditions do not vary as the absolute magnitude of the main effects for the misconception varies.

Correlations among Attributes

Generally, we expect latent traits to be correlated. The degree to which misconceptions are correlated is an empirical question and may vary in different contexts. For the simulation study, structural parameters were modeled with a log-linear model in which parameters were set to values that yielded two different tetrachoric correlations between all pairs of misconceptions.

For one condition, the correlation was .50, reflecting a reasonable hypothesis for the level of correlation we may expect to find empirically. The second condition used a lower correlation of .25. Considering the extreme case where attributes were nearly perfectly correlated may aptly illustrate the impact of correlated attributes. If the attributes were almost perfectly correlated, then all examinees would be expected to be members of the class that possesses no misconceptions or the class that possesses all misconceptions, as the misconceptions would not be distinguishable from one another. Conversely, the less correlated misconceptions are, the more likely it is for all attribute patterns to be plausible. Estimation is expected to improve when all attribute patterns that are posited by the model exist and have substantial membership. A correlation of .25 was hypothesized to provide a more favorable scenario under which to estimate the model.

Data Generation

Data were generated with a program written in Fortran. For all conditions, item intercept parameters were randomly drawn from a uniform distribution, $U(-1,1)$, and item two-way interactions were randomly drawn from a uniform distribution, $U(0.5, 1)$. Main effects and structural parameters were specified as described above. Ability parameters were sampled from a standard normal distribution (i.e., $N(0,1)$).

Estimation

Identification of the model was described in Chapter 3, as were the specific steps of the estimation algorithm. The estimation procedure was set to iterate for 10,000 stages, with a burn-in period of 5,000 stages. The starting value for each parameter was set to its true value. By allowing the chain to start at the mean of the target posterior distribution, the estimation algorithm reached stationarity more quickly. The purpose for choosing these values was to reduce the time each algorithm took to run and thus allow for more replications in the study.

Evaluation of Simulation Study Results

The performance of the SICM model was evaluated with respect to convergence, the accuracy of the recovery of parameters, and the reliability of the examinee parameters. Convergence was assessed by visually inspecting plots and by calculating one-chain Gelman and Rubin (1992) diagnostics using the CODA package (Plummer, Best, Cowles, & Vines, 2006) in R (R Development Core Team, 2011). The accuracy of the model parameter estimates was evaluated with three measures: bias, root mean squared error (RMSE), and Pearson correlations. These indices provided different, yet complementary, information about the accuracy of the estimates. The standard errors of the estimates were also examined and are provided for the reader, but to avoid repetitiveness, results will not be discussed with respect to the standard errors because the RMSE also captures the variability of the estimate. The classification accuracy produced by each condition was examined with the overall correct classification rate (CCR) and with Cohen’s kappa. These measures were used to examine classification accuracy with respect to individual misconceptions and with respect to the whole pattern of misconceptions. The reliabilities of the examinee parameters (i.e., examinee ability and classifications) were calculated using a test-retest type of reliability coefficient. The following sections define and explain these evaluation criteria in further detail.

Convergence

To assess the convergence of each MCMC chain, time series plots and density plots of the posterior distribution of the estimated parameters provided visual indicators of whether or not the chain had converged. If the chain converged properly, time series plots showed the parameter estimate narrowing in on a certain value and not transitioning to values far from the estimated value in the later part of the chain. Densities of the posterior distribution were plotted to evaluate the symmetry of the distribution. If the estimate for the parameter is being determined by the mean of the posterior distribution (as is the case for this study, which will be described in further detail in the next section), density plots of the posterior distribution should be symmetric. A skewed posterior distribution indicated the mean of the distribution was biased. To statistically assess convergence, a variation of Gelman and Rubin's (1992) test statistic ($\hat{R}$) was used. When multiple MCMC chains are run, this statistic compares the within and between variances for separate Markov chains to determine if the chains are mixing well, a feature requisite for the separate chains to converge to the same value. Separate chains were not run for this simulation study; however, to evaluate whether the first and second halves of the single chain were reaching the same value, $\hat{R}$ was calculated as if each half was a separate chain. The value of this statistic can be interpreted as the potential scale reduction factor of the confidence interval for an individual parameter if the chains are never stopped; thus an $\hat{R}$ value of 1 indicates running the chain longer will not decrease the width of the confidence interval. Gelman and Hill (2007) suggested a value of $\hat{R} \leq 1.1$ as a cut-off to determine if the chains have run long enough, and a value of $\hat{R} \leq 1.5$ as a more liberal cut-off for determining convergence (p. 358).

Accuracy of Parameter Estimates

The accuracy of the parameter estimates was examined using the bias of the parameter estimate, the root mean squared error of the estimate, and the correlation between the estimated and true values of the parameter.

Bias

Bias is a measure that reflects the signed difference in the estimated value of the parameter and true value of the parameter. The bias of a given parameter estimate, $\hat{\tau}$, is

$$B(\hat{\tau}) = E(\hat{\tau}) - \tau \tag{40}$$

where $E(\hat{\tau})$ is the expected value of the parameter estimate and τ is the value of the true parameter. The larger the magnitude of the bias, the further away the estimated value of the parameter is from the true value. The sign of the bias reflects whether the parameter is being over- or under-estimated by the model. Specifically, a negative bias indicates the parameter is being under-estimated.

The value of $E(\hat{\tau})$ can be determined is different ways. For this study, the Expected A Posteriori (EAP) estimate of $\hat{\tau}$ was used. The EAP estimate is the mean of the posterior distribution, which was approximated by taking the mean of the values of the Markov chain after K burn-in stages:

$$E(\hat{\tau}|\boldsymbol{X}) = \int P(\hat{\tau}|\boldsymbol{X})\, d\tau \;\approx\; \sum_{s=K+1}^{T} \frac{\hat{\tau}_s}{T-K} \tag{41}$$

Although not used in this study, an alternative to the EAP estimate is the Maximum A Posterior (MAP) estimate, which uses the mode of the posterior distribution instead of the mean to determine the estimate. Both EAP and MAP estimates will converge to Maximum Likelihood estimates if non-informative priors are used in the MCMC algorithm.

Root Mean Squared Error

The root mean squared error (RMSE) is a measure of estimation accuracy influenced by both the bias and the variance of the estimate. The mean squared error (MSE) is the expected value of difference in the estimate and true parameter values, which can be decomposed into the sum of the variance of the estimate and the squared bias of the estimate:

$$MSE\ (\hat{\tau}) = E((\hat{\tau} - \tau)^2)\ = V(\hat{\tau}) + B\ (\hat{\tau})^2 \qquad (42)$$

The RMSE is the square root of the MSE, for which the expected value of the squared differences in the true and estimated parameter is approximated by taking the square root of the average of the square distances across replications:

$$RMSE\ (\hat{\tau}) = \sqrt{MSE\ (\hat{\tau})} = \sqrt{E((\hat{\tau} - \tau)^2)} \approx \sqrt{\frac{\sum_{r=1}^{R}(E(\hat{\tau}) - \tau)^2}{R}} \qquad (43)$$

where R equals the total number of replications of the estimation. In this study, R = 50.

The RMSE does not offer information about the direction of the estimation inaccuracy as the bias does, but instead offers information about the variability and inaccuracy of the estimated parameter in an absolute sense. An increase in RMSE reflects an increase in the expected value of the difference in the true and estimated parameter values, which indicates a decrease in estimation accuracy.

Pearson Correlation

The Pearson product moment correlation coefficient was used to describe direction and strength of the linear relationship between the estimated and true values of the parameters. Specifically, the correlation is calculated as:

$$r(\hat{\tau}, \tau) = \frac{\sum_{s=K+1}^{T}(\hat{\tau}_s - \overline{\hat{\tau}_s})\ (\tau_s - \overline{\tau_s})}{\sqrt{\sum_{s=K+1}^{T}(\hat{\tau}_s - \overline{\hat{\tau}_s})^2}\sqrt{\sum_{s=K+1}^{T}(\tau_s - \overline{\tau_s})^2}} \qquad (44)$$

A strong positive correlation between the estimated and true parameter indicates that the estimation process is able to accurately recover the true value of the parameter.

Standard Error

The standard error of the estimate is the standard deviation of the posterior sampling distribution. This value is estimated by taking the standard deviation of the values of the Markov chain after K burn-in stages:

$$SE(\hat{\tau}|\boldsymbol{X}) = \sqrt{E(\tau - E(\tau))^2} \approx \sqrt{\sum_{s=K+1}^{T} \frac{\hat{\tau}_s{}^2}{T-K} - (\sum_{s=K+1}^{T} \frac{\hat{\tau}_s}{T-K})^2} \tag{45}$$

The standard error provides a measure of how variable the estimate is. Large standard deviations indicate highly variable estimates. Using the standard error, credible intervals can be calculated to determine the posterior probability the true value of the parameter lies within a given range of values. For example, the following equation approximately describes the interval within which the true value of τ lies, with a 95% probability:

$$(\hat{\tau} - SE(\hat{\tau}|\boldsymbol{X})(1.96), \hat{\tau} + SE(\hat{\tau}|\boldsymbol{X})(1.96)). \tag{46}$$

As the standard error decreases, this credible interval decreases and the location of the true value of the parameters becomes more precise.

Accuracy of Examinee Classifications

The classification accuracy was examined with the correct classification rate (CCR) and with Cohen’s kappa. For each examinee, the attribute pattern classification was assigned by determining the most likely class, given the EAP estimate for classification. As with item parameter estimates, the EAP estimates for classifications were approximated by the mean of the values of the Markov Chain. The attribute pattern is a multivariate Bernoulli variable such that the mean of its posterior distribution is a vector of 2^A probabilities ($P\left(\widehat{\boldsymbol{\alpha}}_{ep} = \boldsymbol{\alpha}_p\right)$ for

$p = 1,2, \ldots, 2^A$), providing the probability that an examinee possesses each possible attribute pattern. Each examinee was assigned to have the overall attribute pattern p^*, where

$$max(P(\widehat{\boldsymbol{\alpha}}_{ep} = \boldsymbol{\alpha}_p | p = 1, 2, \ldots, 2^A)) = P(\widehat{\boldsymbol{\alpha}}_{ep} = \widehat{\boldsymbol{\alpha}}_{p^*}). \quad (47)$$

Examinees were classified according to each individual attribute. That is, an examinee either possessed ($\hat{\alpha}_{ea} = 0$) or did not possess ($\hat{\alpha}_{ea} = 1$) each misconception. Taken individually, misconceptions are Bernoulli variables with a mean that represents the probability that an examinee possesses that misconception. If $P(\hat{\alpha}_{ea} = 1) > .50$, an examinee was classified as having that misconception. The classification of individual attributes is parallel to the attribute pattern classification of selecting the category (i.e., pattern) for which $P\left(\widehat{\boldsymbol{\alpha}}_{ep} = \boldsymbol{\alpha}_p\right)$. Here only two categories exist, so $P(\hat{\alpha}_e = 1) > .50$ if and only if $P(\hat{\alpha}_e = 1) > P(\hat{\alpha}_e = 0)$.

Correct Classification Rate

The CCR is the proportion of estimated classifications that are accurate (i.e., the estimated classification equals the true classification). The CCR for the attribute pattern is the number of examinees who were assigned the correct attribute pattern divided by the total number of examinees. For each individual attribute, the CCR was also calculated by dividing the number of examinees who were correctly classified according to that attribute (i.e., $\hat{\alpha}_{ea} = \alpha_{ea}$) by the total number of examinees.

Cohen's Kappa

Cohen's kappa is an index used to describe the level of agreement between two classifications relative to the level of agreement that would be reached by chance. In this setting, the two classifications are the estimated and true classification. The level of agreement that would be reached by chance is the calculated as if the estimated and true classification were statistically independent.

For the attribute pattern, statistical independence means that the joint probability an examinee has a matching estimated attribute pattern ($\hat{\boldsymbol{\alpha}}_{ep}$) and true attribute pattern ($\boldsymbol{\alpha}_{ep}$) equals the product of the estimated ($\hat{\upsilon}_p$) and true (υ_p) marginal probabilities of attribute pattern possession. For determining Cohen's kappa for attribute pattern agreement, a $2^A\ x\ 2^A$ matrix is formed. In this matrix, the rows represent estimated attribute patterns, columns represent true attribute patterns, and the cell entries are joint probabilities of $\boldsymbol{\alpha}_p$ and $\hat{\boldsymbol{\alpha}}_p$. Cohen's kappa is typically denoted as (e.g., Agresti, 2007)

$$\kappa = \frac{\sum \pi_{ii} - \sum \pi_{i+}\pi_{+i}}{1 - \sum \pi_{i+}\pi_{+i}}. \tag{48}$$

The term $\sum \pi_{ii}$ is the sum of the diagonal entries of the matrix. In terms of the attribute pattern,

$$\sum \pi_{ii} = \sum_{p=1}^{2^A} P(\hat{\upsilon}_p = \upsilon_p). \tag{49}$$

The value of this term increases as the accuracy of classification increases, which in turn increases the value of Cohen's kappa. Thus, larger values of Cohen's kappa reflect greater classification accuracy. The term $\sum \pi_{i+}\pi_{+i}$ is the sum of the expected value of each cell in the diagonal of the matrix under the assumption of independence. In terms of the attribute pattern,

$$\sum \pi_{i+}\pi_{+i} = \sum_{p=1}^{2^A} (\hat{\upsilon}_p)(\upsilon_p). \tag{50}$$

Using this notation to contrast Cohen's kappa with the CCR, the average correct classification rate across all attribute patterns can be expressed as

$$CCR(\boldsymbol{\alpha}_.) = \frac{\sum_{p=1}^{2^A} P(\hat{\upsilon}_p = \upsilon_p)}{2^A}. \tag{51}$$

Interpreting the meaning of 0 for each of these indices facilitates contrasting the two measures. If the CCR equals 0, no examinee was correctly classified. A zero value for Cohen's kappa would reflect an estimated classification what would be expected merely by chance. Both the CRR and Cohen's kappa have an upper bound of 1, which indicates perfect agreement of classification for both measures.

Cohen's kappa was calculated similarly for marginal attribute possession agreement. In this case, the matrix was a $2\ x\ 2$ matrix, where rows were estimated attribute possession ($\hat{\alpha}_{ea} = 0$ or $\hat{\alpha}_{ea} = 1$), columns were true attribute possession ($\alpha_{ea} = 0$ or $\alpha_{ea} = 1$), and cell entries were again joint probabilities.

Reliability of Examinee Estimates

In testing situations, the reliability of an estimate can be conceptualized as the consistency of an examinee's estimate over repeated testing occasions using parallel tests. In this study, two types of examinee estimates were considered: a continuous ability and a set of categorical misconceptions. As in Templin & Bradshaw (in press), reliability for the examinee estimates was conceptualized as a test-retest reliability influenced by the variability of the posterior distributions. Calculating reliability in this way placed reliability on the same metric for both types of examinee estimates and allowed for comparisons among the reliabilities for the continuous and categorical estimates.

Examinee Ability Estimates

To simulate a test-retest situation, two independent draws were made from the posterior ability distribution for each examinee. For an examinee's ability estimate, the mean of the posterior distribution $\hat{\theta}_e$ and the standard deviation of the posterior distribution $SE(\hat{\theta}_e)$ specified the parameters of the normal distribution from which two draws were made. From Templin and Bradshaw (in press), the steps for calculating the reliability coefficient for the examinee ability parameter were

1) Randomly draw a value θ_{e1} from $N(\hat{\theta}_e, SE(\hat{\theta}_e))$ for each examinee.
2) Record these values in vector T_1, which represents the examinee estimates for Test 1.
3) Randomly draw a second value θ_{e2} from $N(\hat{\theta}_e, SE(\hat{\theta}_e))$ for each examinee.
4) Record these values in vector T_2, which represents the examinee estimates for Test 2.
5) Calculate the Pearson correlation of T_1 and T_2: $r(T_1, T_2)$.

Although $r(\theta_{e1}, \theta_{e2}) = 0$ for an individual examinee, the test-retest correlation $r(T_1, T_2)$, becomes the reliability estimate across examinees. As is the case with other measures of reliability, the measure of the reliability is inversely related to the standard error of the estimate. If the standard error of the parameter estimate is large, then draws θ_{e1} and θ_{e2} are more likely to be distant from each other. The practical interpretation of the simulated draws being distant from each other is that the examinees' estimates from one testing occasion to the next are not consistent. In other words, the examinees' estimates of abilities are not reliable. This notion is reflected in the measure of reliability as defined here. The more different the two draws are for examinees, the lower the correlation (i.e., reliability) will be among the estimates from Test 1 and Test 2 across all examinees.

Examinee Classification

For examinee classifications in DCMs, an analogous form of the test-retest reliability described above for IRT models is calculated. As described before, the salient difference between IRT model and DCM estimates is the assumed distributions of the examinee parameters. For DCMs, the distribution of an attribute *a* is Bernoulli with probability of mastery, p_a. In the case of the SICM model, the probability is the probability a misconception is present. For an examinee *e*, the mean of the estimated distribution is $\hat{p}_{ea}$, the estimated marginal probability that examinee *e* has misconception a. Similarly, the variance of the estimate is the resulting variance of a Bernoulli variable, or $(1-\hat{p}_{ea})$ $(\hat{p}_{ea})$. Unlike the variability of examinee estimates in IRT models, the variance of examinee estimates in DCMs are not conditional upon the level or value of the estimate (on whether $\hat{\alpha}_{ea}=0$ or $\hat{\alpha}_{ea}=1$).

For DCMs, and thus the SICM model, the inconsistency of the measure does not have to be simulated because the measures can only differ in two ways: either an examinee has a misconception ($\alpha_{ea_1}=1$) on the first assessment and does not have the misconception ($\alpha_{ea_2}=0$) on the second, or vice versa ($\alpha_{ea_1}=0, \alpha_{ea_2}=1$). The probability of observing any combination of misconception values is found by the product of the estimated probabilities for each. This product is the analog to drawing two independent values for the continuous ability from the same distribution.

For a binary misconception, there are four possible outcomes for an examinee *e*'s estimates with respect to misconception *a* on the two tests (as the misconception can take on one of two values on each of the tests). Because of the assumption of parallel tests, the marginal probability of a misconception being present on either test is equal, or $P(\alpha_{ea_1}=1)=P(\alpha_{ea_2}=$

$1) = \hat{p}_{ea}$. Similarly, the probability of observing any combination of misconceptions for the pair of tests is found by the product of the marginal probabilities. Specifically, $P(\alpha_{ea_1} = 1;\ \alpha_{ea_2} = 1) = \hat{p}_{ea}\hat{p}_{ea}$, $P(\alpha_{ea_1} = 1;\ \alpha_{ea_2} = 0) = \hat{p}_{ea}(1 - \hat{p}_{ea})$, $P(\alpha_{ea_1} = 0;\ \alpha_{ea_2} = 1) = (1 - \hat{p}_{ea})\hat{p}_{ea}$, and $P(\alpha_{ea_1} = 0;\ \alpha_{ea_2} = 0) = (1 - \hat{p}_{ea})(1 - \hat{p}_{ea})$. By summing these possible joint probabilities for each examinee, a $2\ x\ 2$ contingency table can be formed whose entries represent the number of examinees expected to have each combination of classification. For instance, the first cell of the table describes the number of examinees expected to be a classified as having the misconception on each assessment: $P(\alpha_{.a_1} = 1;\ \alpha_{.a_2} = 1) = \sum_{e=1}^{E} \hat{p}_{ea}\hat{p}_{ea}$. Dividing this value by the number of examinees (E) yields the probability a given examinee will be classified as having the misconception on both tests. Therefore, sampling to simulate $\hat{\alpha}_{ea}$for a pair of parallel tests is not necessary; the probability can be fully determined given the mean of the estimated distribution for an examinee.

As with the test-retest reliability metric defined for IRT models, for any given examinee, the correlation between α_{ea_1} and α_{ea_2} is zero. Upon aggregating across examinees in a sample, however, the correlation becomes non-zero, provided estimated misconception probabilities for examinees are different from .5 – a case which would indicate zero reliability from the zero correlation between α_{ea_1} and α_{ea_2}. Because the Pearson correlation coefficient is bounded above by -1 and below by 1 for $2\ x\ 2$ contingency tables where the marginal proportions are not both .5 (a scenario that is likely for most applications of DCMs), the tetrachoric correlation coefficient is used as the metric of test-retest reliability.

Simulation Results

The results of the simulation study are reported and evaluated using the measures detailed in the previous section. For this study, each of the 64 conditions was estimated 50 times. Replicating the estimation a large number of times allowed for the variability in the parameter estimates across separate estimations to be captured. Thus, the reported value of each measure being used for evaluation (e.g., bias) is an average value across replications. Along with the average value of the measure, the standard deviation of this measure for the 50 replication is provided in parentheses following the mean value to quantify the variability across replications.

First, successful convergence of the estimation algorithm upon the target distribution will be assessed. Then, estimates for parameters will be evaluated in the following order: (a) item parameters, (b) structural parameters, (c) examinee ability estimates, and (d) examinee classifications.

Generally, results will be provided in tables where values were (a) averaged *across* the magnitude of main effects factor and/or (b) averaged across all other factors and given *by* the magnitude of main effects factor. Results aggregated across the magnitude of main effects factor represent the average value irrespective of the magnitude of the main effects, and results aggregated by the magnitude of main effects factors represent the average value irrespective of all other factors manipulated in the simulation study. Recall there were four variations of the magnitude of main effects factor, which were (a) Low Misconception Effect, Low Ability Effect; (b) Low Misconception Effect, High Ability Effect; (c) High Misconception Effect, Low Ability Effect; and (d) High Misconception Effect, High Ability Effect.

Convergence

Convergence was assessed for item, structural and examinee parameters. Time series plots and density plots of the posterior distribution of the estimated parameters were visually inspected for a replication within each condition to informally assess whether or not the chain had converged. Generally, time series plots showed the parameter estimate narrowing in on a certain value in the later part of the chain, and densities of the posterior distribution were reasonably symmetric; however, these desirable results did not always occur. As an illustration, a time series and density plot is provided in Figure 10 for two examinee ability estimates that converged (Examinee 8 and 123), although indicators for convergence were more certain for Examinee 8 than 123. Examinee 99's ability estimate is also shown in Figure 10. The ability estimate showed a significant lack of convergence, indicating a need for a longer chain. Results similar to Examinee 99 were very atypical. The true and estimated parameters, along with their standard errors, for these examinees are: $\theta_8 = 1.469, \hat{\theta}_8 = 1.723, SE(\hat{\theta}_8) = .178$; $\theta_{123} = 1.034, \hat{\theta}_{123} = 0.868, SE(\hat{\theta}_{123}) = .340$; $\theta_{99} = -0.850, \hat{\theta}_{99} = -.097, SE(\hat{\theta}_{99}) = .964$.

Results from a variation of Gelman and Rubin's (1992) $\hat{R}$ test statistic, which was used to statistically assess convergence, are given in Tables 6a and 6b. Table 6a contains the results for the conditions where pairs of attributes each had a .50 correlation, and Table 6b contains the results for conditions where pairs of attributes each had a .25 correlation. Results in Table 6a and 6b were aggregated across magnitude of main effect conditions. In each table, the mean and median values of $\hat{R}$ across replications are given, and the average percentage of item parameters that were deemed to have converged according to typical ($\hat{R} \leq 1.1$) and more liberal ($\hat{R} \leq 1.5$) criteria are provided.

A comparison of Tables 6a and 6b shows that for item and examinee parameters, the results do not consistently differ according to the correlation among attribute pairs. The conditions for which the least number of item parameters converged were the 3,000 examinee/30 item/6 attribute conditions (43.25 % <1.1 and 78.85% <1.5, when ρ_{tet}= .50). For all other conditions, at least 85% of item parameters converged according to the more liberal criteria. The greatest number of item parameters converged in the 10,000 examinee/60 item/3 misconception conditions (72.25 % <1.1 and 96.75% <1.5, when ρ_{tet}= .50). For all conditions, between 85.75% and 93.25% of examinee parameters converged according to the more liberal criteria.

Tables 6a and 6b do differ by the percentage of structural parameters that converged according to the $\hat{R}$ statistic. When the correlation between attribute pairs was .50, the percentage of structural parameters that converged according to the more liberal criteria ranged from 55 to 96.25%; when the correlation between attribute pairs equaled 0.25, this corresponding range was from 64.5 to 95.75%. Particularly for the six attribute conditions, a considerable difference was seen between these two conditions. The estimation of structural parameters converged more frequently with the lower correlations when a larger number of attributes were measured. Because the correlation among attributes pertains to the structural component of the model, this result was expected.

A high percentage of parameters were expected to converge for all conditions. To improve convergence, the number of iterations and/or the length of the burn-in period in the algorithm can be increased. To ensure the algorithm converges, Gelman and Hill (2007) suggested checking the value of $\hat{R}$ during steps of the estimation process and continuing the estimation algorithm until $\hat{R} \leq 1.1$ for all parameters. Given the number of parameters being estimated in each condition (between 240 and 490, as shown in Table 7), the number of conditions (64), and the total number of replications (50) being assessed through the simulation study, a compromise between complete convergence and estimation time was considered and a large, yet fixed, number of iterations was used.

Accuracy of Parameter Estimation

Generally, item and examinee parameter estimates were most accurately estimated for conditions with more examinees, more items, and fewer misconceptions to measure. These trends are consistent with the psychometric literature at large; estimation is improved when there are fewer parameters to estimate and when the model has more information with which to determine the parameters. More specifically, the results reported in this section are compatible other simulation studies in the DCM literature (e.g., Choi, 2010; Henson, Templin & Willse, 2009). Results from varying the magnitude of the main effects conditions uncovered no barriers to estimating both the categorical and continuous latent predictors and also shed some light on which conditions yielded more accurately estimated parameters. The only surprise in the results was that the correlation among misconceptions did not always impact estimation as initially expected. That is, estimation was not always better for the lower correlation conditions (.25). This surprise has a reasonable explanation and was found to be a useful phenomenon to demonstrate considering .50 correlations are closer to the value of the correlations among misconception we may expect to find in practice. Results from item, structural and examinee parameters will be discussed in turn.

Item Parameters

Four types of item parameters will be discussed: (a) intercepts ($\lambda_{n_{ij},0}$), (b) main effects for misconceptions ($\lambda_{n_{ij},1(\alpha)}$), (c) two-way interactions for pairs of misconceptions ($\lambda_{n_{ij},2(\alpha,b)}$), and (d) main effects for ability estimates ($\lambda_{n_{ij},\theta}$). Tables 8a and 8b provide the bias, RMSE, and correlation for each of these types of parameters averaged across the magnitude of main effects factor and given by four of the manipulated factors in the simulation design: number of examinees, test length, and number of misconceptions measured. Table 8a contains results for the 32 conditions for which a tetrachoric correlation of .50 was specified among every pair of misconceptions. Table 8b contains analogous results for the other half of the conditions for which a tetrachoric correlation of .25 among the attributes was specified. The results did not vary greatly nor consistently between these two tables, indicating that a decrease in correlation among the attributes did not improve estimation accuracy for item parameters as expected. Discussion in later parts of the results section addresses why this may have been the case. Because results for conditions with different tetrachoric correlations among attribute are so similar, detailed interpretations of results from each table would be repetitive. Given that a correlation of .50 is more realistic, results will be discussed with respect to Table 8a.

First, consider the item intercepts. The average bias and RMSE across all item intercepts was greatest for the 3,000 examinee/30 item/6 attribute condition. Consistent with these indices, the average Pearson correlation among the estimated and true parameter values was less in this condition, indicating estimation inaccuracy was greatest for this condition. However, for all conditions, the absolute value of the bias was less than 0.012, the RMSE was less than 0.015, and the correlation was greater than .981, indicating accurate estimation of the item intercepts.

Similar trends exist for the main effects for the misconceptions and the interactions of the misconceptions: conditions with fewer examinees, fewer items, and more misconceptions were estimated less accurately. For the main effects, the bias was slightly less than the bias for the intercepts, but the RMSE was greater. For the intercepts, bias was consistently negative, indicating the intercept was under-estimated. An explanation for the near zero bias for the main effects in each of the conditions, yet greater RMSE for the main effects than for the intercepts is that the bias for the main effects did not tend toward over- or under- estimation. Thus, the over- and under- estimation effectively canceled each other out in the bias measure, but was captured in the RMSE, illustrating why the RMSE is a useful measure to use in addition to the bias.

In an absolute sense, the main effects were estimated accurately. The RMSE values for the main effects were all under 0.015 for conditions with 10,000 examinees and under 0.050 for conditions with 3,000 examinees. For estimation of the main effects of the misconceptions, the number of examinees seemed to have an effect on estimation accuracy, and this trend was also reflected by Pearson correlations which were greater for the 10,000 examinee conditions (ranging from .767 to .879) than for the 3,000 examinee conditions (ranging from .561 to .707).

For the two-way interaction parameters, the number of examinees did not impact estimation accuracy as much. Holding other factors constant, conditions with more examinees did consistently result in more accurate estimates; however, all conditions with more examinees did not have more accurate estimates than all conditions with fewer examinees. The RMSE value for the 3,000 examinee/60 item/3 misconceptions conditions (0.055) indicated only slightly greater estimation accuracy in comparison to the 10,000 examinee/30 item/6 misconceptions conditions (0.078), although Pearson’s correlation was stronger (.479) for the later condition than for the former condition (.313). In an absolute sense, the RMSE values were all under 0.10, except for the two 6 attribute conditions measured with 3,000 examinees which still had low RMSE values of 0.228 and 0.110 for the 30 and 60 item assessments, respectively. The Pearson correlations (ranging from .313 to .750) indicated less accurate estimation than for other item parameters.

The last section of Table 8a contains results for the main effects for ability. Because the absolute magnitude was smaller for ability main effects than for the other parameters (e.g., high main effects for misconceptions ranged from 1.75 to 2.25, while high main effects for ability ranged from 0.60 to 0.80) , information for a relative comparison of accuracy was not available from the bias or RMSE. According to Pearson correlations, however, the estimated parameter was more strongly related to the true value of the parameter for ability main effects than for misconception main effects or interactions, but not as strongly related as the intercepts.

The bias and RMSE did indicate that ability main effects were accurately estimated. The less favorable conditions were again conditions that had fewer examinees and items, although the RMSE values were very close for the different conditions. The RMSEs for the 3,000 examinee/30 item conditions were 0.042 and 0.044 for the 3 misconception and 6 misconception conditions, respectively. Corresponding correlations were .810 and .811. Even for these conditions, a 95% credibility interval suggested that on average the true value of the loading for ability was within approximately 0.08 of the estimated value. As for the main effects for the misconceptions, all 10,000 examinee conditions had slightly greater estimation accuracy than the 3,000 examinee conditions, regardless of the test length or the number of misconceptions.

Table 9 provides the estimation accuracy results with respect to all four types of item parameters by the magnitude of the main effects. For this table, results were aggregated across the all other factors (including the tetrachoric correlation factor) to examine the effect that the absolute and relative magnitude of the main effects had on estimation accuracy.

Results from this table showed that strength of the main effects for misconceptions most significantly impacted estimation. Low main effects for misconceptions led to decreases in estimation accuracy, regardless of ability main effects. Low main effects for ability seemed to have little impact on estimation. Overall estimation was reasonably accurate for all conditions and parameters, as indicated by the largest bias of 0.011 and the largest RMSE of 0.117. However, the correlations for the interactions were somewhat low, ranging from .437 to .603. The correlation was still positive, but less strong than the correlations for the intercepts (ranged from .986 to .991), main effects for misconceptions (ranged from .703 to .776), or main effects for ability (ranged from .853 to .933).

Structural Parameters

Three types of structural parameters will be discussed: (a) intercepts, (b) main effects for misconceptions, and (c) interactions for pairs of misconceptions. Table 10a and 10b provide the bias and RMSE for the .50 and .25 tetrachoric correlation conditions, respectively. The Pearson correlation was not broken down by structural parameter type, as there were low numbers of each type (and only one intercept), but was used as an overall measure for all structural parameters. As for the item parameters, results in these tables were aggregated across all four variations of the magnitude of main effects factor.

For each type of structural parameter, the indices showed agreement on the conditions for which estimation was the most and least accurate. The 10,000 examinee/60 item/3 misconception condition had the most accurately estimated parameters, and the 3,000 examinee/30 item/6 misconception condition had the least accurately estimated parameters.

The bias and RMSE measures indicated the intercept was estimated slightly better than the main effects, and the main effects were estimated slightly better than the interactions. Referring to Table 10a (the ρ_{tet}=.50 conditions), estimation for the 3,000 examinee/30 item/6 misconception condition was significantly less accurate than the other conditions. The number of misconceptions seemed to most significantly impact estimation. Holding the number of items and examinees constant, the RMSE values increased significantly when 6 misconceptions were measured instead of 3. However, accuracy did improve as the number of items and examinees increased. When six misconceptions were measured by 60 items using a sample of 10,000 examinees, the RMSE values remained under 0.10 for all types of structural parameters. The effect of the increased number of misconceptions was less pronounced for the ρ_{tet}=.25 conditions shown in Table 10b. Estimation accuracy was slightly better for the main effects and interactions in the 6 attribute conditions.

The Pearson correlation between the true and estimated values of the parameters is given in the last section of Tables 10a and 10b. These correlations ranged from .972 to .999, indicating a near perfect linear relationship between the true and estimated parameter values for all of these conditions.

Table 11 provides the evaluation criteria corresponding to the magnitude of the main effects conditions. For all three types of parameters, the same trend in estimation accuracy according to the absolute and relative magnitudes of the main effects for the misconceptions and ability was seen. This trend was the same as for the item parameters discussed previously. Increasing misconception main effects from low to high led to an improvement in the bias, RMSE, and correlation. The strength of the ability estimate had little impact on estimation, although, for the conditions with high misconception main effects, decreasing ability main effects from high to low resulted in an additional slight improvement in estimation. Although the absolute value of the bias remained under 0.027 for all structural parameters, the main effects were consistently under-estimated and the interactions were consistently over-estimated.

Examinee Ability Estimates

For each examinee, an overall measure of ability was estimated. Table 12 displays the bias, RMSE, and correlation for these estimates according to four factors that were manipulated in the simulation study: tetrachoric correlation among pairs of attribute, number of examinees taking the assessment, number of items on the assessment, and the number of misconceptions being measured by the assessment.

Table 12 shows that the number of items strongly influences the estimation accuracy. For both the .25 and .50 correlation conditions, the measures showed greater estimation accuracy for the 60 item conditions compared to the 30 items conditions, holding other factors constant. Results varied little with respect the tetrachoric correlation conditions, and thus will be discussed with references to the ρ_{tet}=.50 conditions. As the number of examinees increased from 3,000 to 10,000, estimation was not significantly improved. For example, for the 60-item conditions, only the 3,000 examinee conditions had only a slightly better RMSE (0.592 for three attributes and 0.599 for 6 misconceptions) than the 10,000 examinee conditions (0.589 for 3 misconceptions and 0.596 for 6 misconceptions). The RMSE was also not significantly impacted as the number of misconceptions increased from three to six. The correlations were also varied with respect to the number of items measured. For the 30 item conditions, the correlations ranged from .675 to .697, while for the 60 item conditions, the range was from .795 to .803. The bias measure varied little. Most conditions under-estimated examinee ability, albeit with very small bias (bias ranging from -0.003 to -0.015).

Table 13 provides the results for the examinee ability estimates by the magnitude of main effects factor. The indices indicated that as ability main effects increased, the estimation accuracy increased. The strength of the misconception main effects did not significantly impact accuracy, although low misconception main effects resulted in slightly better accuracy than high misconception main effects. The impact of the magnitude of main effects for ability on examinee ability was exactly the opposite of the impact for item and structural parameters.

In Table 14, the reliability for the examinee ability estimate is given in the column labeled θ. Results for the .25 condition will not be discussed, but are provided for the reader to see that they are similar to results from the .50 condition. In the .50 condition, reliability ranged from .523 to .675. Although there is not a minimum reliability that is universally accepted, these measures fall short of reaching values of .70 or above that are often reported in achievement testing (Crocker & Algina, 1986). Reliability most significantly increased as the number of items increased, decreased very slightly as the number of misconceptions increased, and showed little change when the number of examinees increased.

Table 15 shows the reliability for examinee ability as a function of the varying magnitudes of the main effects. Reliability was greatest (.709) when the main effect for ability was high and the main effects for misconceptions were low and decreased slightly when both main effects are high (.687). Reliability was least (.491) when the main effect for ability was low and the main effects for misconceptions were high and increased when both main effects were low (.522). Trends seen for reliability of ability in Tables 14 and 15 were the same as trends seen for estimation accuracy of ability seen in Tables 12 and 13.

Examinee Classifications

The correct classification rate was calculated for each individual misconception and for the misconception pattern as a whole to examine how often examinees were correctly classified. Table 16 contains the CCR across the different magnitude of the main effects conditions. Table 17 contains Cohen's Kappa and is organized in the same way as Table 16. When results found in the .25 and .50 conditions were similar, only results from the .50 conditions will be discussed.

Classification accuracy was high, even for the most disadvantageous conditions. The number of items and the number of misconceptions most strongly influenced the classification accuracy of individual misconceptions, as indicated by the CCR and Cohen's kappa. The CCR and Cohen's kappa, respectively, ranged from .863 and .728 found in the 3,000 examinee/30 item/6 misconception condition to .958 and .917, respectively, for the 3,000 examinee/60 item/3 misconception condition. These ranges were similar for the 10,000 examinee conditions.

Similar results were found for the classification accuracy with respect to the whole misconception pattern. Overall, the CCR and Cohen's kappa were lower for the whole pattern classification than for the marginal classification. The values were lower because the misclassification with respect to any single misconception marginally results in the misclassification of that examinee's whole misconception pattern. The CCR and Cohen's kappa, respectively, ranged from .573 and .505 found in the 3,000 examinee/30 item/6 misconception condition to .894 and .875 found in the 3,000 examinee/60 item/3 misconception condition. These ranges were similar for the 10,000 examinee conditions. This result indicates that increasing the number of examinees had less of an increase in estimation accuracy than increasing the number of items on the test or decreasing the number of misconceptions measured. Similar results were seen for examinee ability.

Although results from the .25 and .50 conditions were very similar for the 3 attribute conditions, they varied for the 6 attribute conditions. Marginally, the lowest observed attribute classification rate and Cohen’s kappa were .842 and .686 for the .25 conditions, which was less than the corresponding values of .864 and .730 for the .50 condition. For classification with respect to the whole pattern of misconceptions, for the 6 attribute conditions, the CCRs ranged from .481 to .682 for the .25 conditions and from .573 to .735 for the .50 conditions; and, Cohen’s kappa ranged from .449 to .665 for the .25 conditions and from .595 to .700 for the .50 conditions. These results indicate that when a higher number of attributes are measured, a decrease in the correlation of attributes does not improve classification accuracy. This result can be explained on a conceptual level by understanding that if misconceptions are correlated, then the model can use information about one misconception to classify examinees according to a related misconception. Thus, having correlated attributes results in more accurate classification because the misconceptions provide information about each other, resulting in the model having a greater amount of information for classifying examinees.

Table 18 displays the CCR and Cohen’s kappa by the magnitude of main effects factor. A clear and consistent trend was seen for the accuracy of marginal and whole pattern classification: classification was most accurate when main effects for misconceptions were high in an absolute sense. The main effects for ability did not significantly impact classification accuracy; accuracy decreased only slightly when main effects for ability increased. These were the same trends seen for item and structural parameters.

In Table 14, the reliability for the each individual misconception ($\alpha_1, \alpha_2, \dots, \alpha_6$) is given in addition to the average reliability across the total number of misconceptions measured in that condition ($\alpha_.$). Reliabilities in this table were averaged across all magnitude of main effects conditions. For the .50 conditions, the average marginal reliability ranged from .853 to .988. These measures were high for reliabilities commonly reported in achievement testing (Crocker & Algina, 1986). The reliabilities increased as the number of misconceptions decreased from six to three and when the number of items increased from 30 to 60, but there was virtually no change in reliability when the number of examinees was increased from 3,000 to 10,000.

In addition to the reliability being impacted by the number of items on the test and the number of misconceptions being measured on the test, the tetrachoric correlation also seemed to have an impact on reliability. Reliabilities for the .25 conditions were very similar to those for the .50 conditions when three attributes were measured. However, for the 6 attribute conditions, the reliabilities in the .50 conditions were higher than the .25 conditions. More specifically, when the tetrachoric correlation decreased from .50 to .25, reliabilities of .853, .945, .853, and .946 decreased to .806, .923, .807, and .924 for the 3,000 examinee/30 item/6 attribute; 3,000 examinee/60 item/6 attribute; 10,000 examinee/30 item/6 attribute; and 10,000 examinee/60 item/6 attribute conditions, respectively. The higher correlation of .50 resulted in greater reliabilities when six attributes were measured. This result may reflect information about possession of a misconception being leveraged from other correlated misconceptions, improving the reliability of the classification of examinees with respect to those misconceptions.

Table 15 shows the average reliability for individual misconceptions as a function of the magnitudes of the main effects. Reliability was extremely high (.998) when the main effect for ability was low and the main effects for misconceptions were high. Reliability was the lowest (.820) when the main effect for ability was high and the main effects for misconceptions were low. The reliability increased only slightly when the main effect for both types of parameters were low (.850); however, it increased drastically when the main effect for both types of parameters were high (.991). These trends are the same those seen for the classification accuracy of misconceptions.

Simulation Study Conclusions

These results provide information about how the SICM model performs under various testing conditions. As is the case with most psychometric models, this model performed better when the sample size was larger, the assessment was longer, and the number of parameters to be estimated by the model was fewer. On average, all item parameters had a RMSE less than .1 under all conditions where 3 misconceptions were being measured or when 10,000 examinees were being assessed. Structural parameters were more difficult to estimate than item parameters and had a RMSE less than .1 when 3,000 examinees were taking the assessment only when 3 misconceptions are measured by 60 items. When 10,000 examinees were used for estimation, the 30 item/6 misconception conditions were the only conditions that yielded RMSE values greater than .10. Given the complexity of the model, these data demands are quite reasonable, potentially placing the SICM model as a viable option amongst other large-scale psychometric models to be used in practice.

Similarly consistent with psychometric model research, examinee parameters (ability estimates and classifications) were less affected by the number of examinees responding to the assessment and more affected by the length of the test and number of misconceptions being measured. Holding other factors constant, the results for the accuracy of the examinee estimates and classifications show greater estimation accuracy for the 60 item conditions (RMSE for ability estimates ranged from .588 to .599 and the CCR for individual attributes ranged from .922 to .958) compared to the 30 items conditions (RMSE ranged from .708 to .725 and CCR ranged from .863 to .918), while the RMSE or classification accuracy was not significantly improved as the number of examinees increased from 3,000 to 10,000.

The examinee estimates were also impacted by the magnitude of main effects factor, which offered some insights into the interplay of continuous and categorical variables being estimated within the same model. The accuracy and reliability of the estimated abilities were greatest when ability had a high main effect in an absolute sense; estimation improved only slightly when ability also had a high main effect in a relative sense (i.e., when misconceptions had a low main effect). Similarly, the accuracy and reliability of the classifications were greatest when misconceptions had a high main effect in an absolute sense, and estimation only improved slightly when the main effect was higher than the main effect for ability in a relative sense. These results indicate that strong main effects for ability improve estimation for ability without significantly hurting estimation of the misconceptions, and strong main effects for misconceptions improve estimation for misconceptions without significantly hurting estimation of ability. Thus, when estimating the SICM* model in practice, the larger concern for estimation regarding main effects is the strength of the main effect in an absolute sense. Given strong main effects for each type of variable, the different types of variables can co-exist within the same model without one dominating the other.

For the SICM model, the reliabilities for classification were uniformly much greater than reliabilities for examinees *regardless* of the characteristics of the conditions under which the estimates were obtained. This finding echoes the results in Templin & Bradshaw (in press) that found across a set of models, DCM classifications (with 2, 3, 4 and 5 categories) were consistently more reliable than ability estimates using IRT.

Lastly, the effect of the correlation amongst the attributes was somewhat different than expected. The notion that the lower correlation, although unrealistic, would provide a more ideal scenario under which to estimate the parameters was untrue. Only small differences seemed to be attributed to the correlation amongst the attributes, with the strongest differences occurring for the six misconception conditions. These differences were not always improvements of estimation or classification for the .25 correlation conditions. Although the estimation accuracy for the structural parameter estimates was more accurate for the .25 correlation/6 misconception conditions, the .50 correlation/6 attribute conditions actually classified examinees more accurately and reliably.

These results indicate that the SICM* model can be estimated with MCMC algorithm used here. This study shows that the lower asymptote can be implemented in the SICM model to provide a more realistic model that accommodates a guessing assumption. Because the SICM model is a combination of other models commonly used to analyze multiple-choice tests, these results provide reason to believe that lower asymptotes specified in this way could also be used in other common models, like the NR IRT and 2-PL IRT models.

Limitations of Simulation Study/Directions for Future Research

Although the results of the simulation study provide some insights for using the SICM model, it did so under ideal conditions where the estimation model was correct. From the results of this simulation study, complete recommendations about theoretical and practical conditions that are required for accurate estimation and classification using the SICM model cannot be made. The many conditions under which this model may be applied make it difficult to delineate all plausible conditions to study how the model is expected to perform under various conditions. This study does provide some general guidelines according to five different factors: sample size, test length, number of misconceptions, correlation among misconceptions, and magnitudes of main effects. Different values of these factors, or different factors all together, may also be encountered when the SICM model is applied in practice. For example, Q-matrices may have different levels of complexity, or Q-matrices may have different levels of accuracy. Fairly complex Q-matrices were used for this simulation study, but perfect accuracy was assumed such that model misspecification was not examined. Model misspecification is an important topic in psychometrics because misspecification of the model has expected negative consequences. Other situations in practice may offer a different number of alternatives or items, and main effects for misconceptions and ability may be mixed within a test instead of having designated absolute and relative magnitudes across the test. Additionally, as discussed in Chapter 3, alternative specifications of the model may be required to align the psychometric theory reflected by the model to mirror researchers’ domain-specific theories. For instance, in Garfield (1992) more than one correct alternative exists for each item, and the SICM model could be adjusted to model this phenomenon by adding an effect for ability on these alternatives.

Future studies surrounding the SICM model can investigate estimation characteristics of this model given different conditions. Also of interest may be to estimate the SICM model without a lower asymptote under similar conditions to answer questions about how estimation accuracy is impacted if a practical lower-asymptote is not a feature of the model. Presently, the estimation effects of the lower asymptote are confounded with the estimation properties of the new model; therefore, it is unknown how the SICM model would fare without the lower asymptote. This information would disentangle these effects to understand the psychometric properties of each and may be relevant to researchers who do not theorize guessing to be present in examinee responses.

CHAPTER 5

EMPIRICAL DATA ANALYSIS

In Chapter 4, the results of a simulation study indicated that the Scaling Individuals and Classifying Misconceptions (SICM) model is able to estimate both a continuous trait representing an overall ability and a set of categorical traits defined as misconceptions. The previous chapter served to demonstrate the conditions under which the model can recover parameters with a reasonable degree of accuracy. The present chapter presents an empirical study that illustrates an application of the model in practice, first using the SICM* model and then using other similar or common psychometric models. Results from the SICM* model are described and then results from the set of models used for estimation are compared. This application is of an educational assessment, although the SICM* model may be applicable in other areas, such as clinical psychology and epidemiology.

Analysis of Empirical Data

The SICM* model tries to provide a psychometric solution to a realistic need in educational assessment. To demonstrate the SICM* model's use in a practical setting, data from a reading comprehension test constructed and administered by a large scale testing company was analyzed. The goal of the reading comprehension test is to measure an overall literacy level to determine whether or not an examinee would benefit from additional instruction via instructional modules, in addition to determining what weakness should be targeted within the modules. Thus, the SICM* model was aligned with the purpose of the assessment. These data are presently modeled with CTT total scores for ability and subscores for misconceptions. Through a partnership with the company, the SICM* model was applied with the goal of providing more reliable estimates of ability and misconceptions, in addition to providing information about the design of the items and test.

Data

The data were the nominal responses of 1097 examinees to the reading comprehension test. The test was a Level B pre-test for the Foundations module. This module is part of the company's literacy assessment and intervention program called the Literacy Navigator. Level B assessments are typically administered to students in the sixth grade. In this 28-item multiple-choice test, short passages were provided for the student to read and were followed by a set of items based on that passage for students to answer. Each item had four possible alternatives. One alternative was the correct answer, and each of the incorrect alternatives corresponded to a type of error that students make when responding to reading comprehension items. Content experts and item writers pre-determined and specified the errors. Information gathered from this pre-test was used to determine what types of instructional modules were appropriate to offer students. The modules target students' weak areas to in turn improve their reading comprehension ability.

Definition of Misconceptions Measured by the Level B Foundations of Literacy Pre-Test

The three errors that content experts defined reflected the types of errors students make on reading comprehension tests of this format. A *non-text based response* was the first type of error. An alternative that measured a non-text based response provided a response that was not based upon or derived from the text in the passage corresponding to the item. This response may, or may not, have been a logical response to the question posed in the stem of the item, but was not found anywhere or was not based on anything in the passage text. A *text-based* *misinterpretation of the passage* was the second type of error. An alternative that measured a text-based misinterpretation of the passage provided a response that was based upon information from the text of the passage; however, the interpretation of or conclusion drawn from the content in the text was incorrect. A *text-based misinterpretation of the question* was the third type of error. An alternative that measured a text-based misinterpretation of the question provided a response that reflected accurate interpretation of the text in the passage, but an inaccurate response to the question. In other words, the student was not able to correctly reason about the response to the item, but they were able to interpret the text meaning from the passage. Names of these three errors were based upon descriptions of errors provided by the test developers, but were not given by the content experts or test-developers themselves.

These types of errors were different from the types of misconceptions measured in assessments that were described at the beginning of Chapter 3. The root of misconception, “concept” implies reference to an idea or thought. These types of errors did not reflect reasoning that resulted from a given understanding or belief that was nascent, still developing, or incorrect. Although the errors did not define certain misconceptions examinees had about reading or text interpretation, these error types did describe where the examinee was making the mistake. Determining where an examinee is making a mistake (i.e., was the passage read, the passage misread/misinterpreted, or the question misread/misinterpreted) does provide additional information about the student beyond an overall measure of ability. If theory exists that examinees consistently make a certain type of error, then the SICM* model could be applied to measure the type of error being made. Labeling these types of errors as misconceptions may be a misnomer, so the term *error* will be used to describe the categorical latent variables being measured by the incorrect alternatives in the SICM* model for this test.

Q-matrix for the Level B Foundations of Literacy Pre-Test

The Q-matrix describing which latent variable was measured by each alternative for the Level B Foundations of Literacy Pre-Test is given in Table 20. On average, each item measured 1.93 errors. Six items measured all three types of errors. Every incorrect alternative measured exactly one error, meaning the test had simple structure. The first error (α_1, the non-text based response) was measured by 30 alternatives in 21 items. The second error (α_2, the text-based misinterpretation of the passage) was measured by 32 alternatives in 19 items. The third error (α_3, text based misinterpretation of the question) was measured by 22 alternatives in 14 items.

Estimation of the Model

This pre-test is a multiple-choice assessment, so the probability that a student will answer the item correctly may be inflated by guessing, particularly for students with a low ability level who may frequently not know the answer to a question. For this reason, the SICM* model will be used to estimate the data.

Estimation of the SICM* model utilized a Markov Chain Monte Carlo algorithm written in Fortran. The same Gibbs sampling steps that were used for the estimation in the simulation study and were outlined in Chapter 3 were used, with one difference: a prior distribution was used to estimate the main effect for ability in the model. This prior distribution will be discussed in the next session. As real data typically take a long chain to converge, the SICM* algorithm was run for 100,000 steps. The first 50,000 steps were designated as the burn-in stage such that values of estimates at these stages did not contribute to the posterior distribution mean that determined values of parameter estimates.

Selection of a Prior Distribution for the Main Effect(s) for Ability

A prior distribution was specified for the main effects for ability. This main effect is exponentiated in the SICM* model. The value of this parameter needs to be held within certain bounds, but how to hold it within those bounds is unknown because this type of exponentiated parameter has not been estimated in a psychometric model before. The parameter is constrained to a lower boundary of zero in order to satisfy the monotonicity assumption that as ability increases the probability of responding correctly to the item increases. Practically, the highest value this parameter should take is unknown and is an empirical question. However, mathematically if the value is much greater than one, an increase in one unit of ability results in a drastic and unrealistic increase in the probability of answering the item correctly.

Four different prior distributions were chosen to estimate the SICM* model in practice. These priors were chosen because of their large probability masses between 0 and 1, a range within which I expected the parameter would typically fall in practice. The priors used to estimate the model were: (a) a lognormal prior with a mean of 0 and standard deviation of 1.5, (b) a lognormal prior with a mean of 0 and a standard deviation of 0.5, (c) a normal prior with a mean of 0.60 and a standard deviation of 0.25, and (d) a normal prior with a mean of 0.60 and a standard deviation of 0.15. The top four plots in Figure 11 provide a histogram illustrating the distribution of the main effect for ability when estimated with each of these priors. The density curves for each of the prior distributions are depicted over the histograms to visually assess the effect the prior distribution had on the estimated distribution of the main effects for ability.

Estimation with either lognormal prior yielded a higher concentration of items with main effects ranging from 0.25 to 0.60 than lie in this range for the prior distribution. Similar observations can be made for the normal prior in the 0 to 0.50 range. This deviation from the prior distribution is common and illustrates that specification of the prior distribution does not overly impact values of the estimated parameters. However, when comparing the lognormal (0, 0.50) prior to the lognormal (0, 1.5) prior, there is considerable difference in the range of values from 0 to approximately 0.25. For the lognormal (0, 0.50) prior, hardly any many effects lie in this range.

Convergence of the model parameters using the different priors informed the choice of the prior to be used. Although primary importance was given to the convergence of the SICM* model, results from two of the other models that will be subsequently used to analyze this data set for comparison with the SICM* model were also investigated. Bock's NR IRT model with a lower asymptote (NR IRT* model) and the 2-PL IRT model with a lower asymptote (2-PL IRT*) were estimated using an analogous MCMC algorithm with (a) a lognormal prior with a mean of 0 and a standard deviation of 0.5 and (b) a normal prior with a mean of 0.60 and a standard deviation of 0.15. Figure 11 also illustrates a comparison of the estimated and prior distributions of the main effects for ability for these two models.

Convergence for all three models was assessed using a variation of Gelman and Rubin's $\hat{R}$, as described in Chapter 4. Results are presented in Table 21. Convergence for nearly all of the parameters for the NR IRT* and 2-PL IRT* models was perfect, even according to the more conservative of the two criterion for convergence (i.e., $\hat{R} < 1.1$). Thus, the prior was picked solely due to the difference in convergence for the SICM* model parameters. The most problematic parameters with respect to convergence were the structural parameters. Although according the $\hat{R} < 1.5$ criterion, the $N(0.6, 0.15)$ prior yielded the largest number of converged structural parameters, according to the $\hat{R} < 1.1$ criterion, the $logN(0, 0.5)$ criterion yielded the largest number of converged structural parameters. The $logN(0, 0.5)$ prior also yielded a considerably higher proportion of main effect for ability parameters that converged according to the more stringent criteria. For these two reasons, the $logN(0, 0.5)$ was used as the prior distribution for the main effects for ability in all models with a lower asymptote for this application. Using this prior, all main effect for ability parameters and virtually all (99.9%) of the examinee ability parameters converged. For the item parameters, 95.2 % converged, and for the structural parameters, only 42.9% converged. The convergence rate for the structural parameters was unacceptably low. Although convergence rates can typically be increased by lengthening the chain, an examination of the chain plots for the structural parameters indicated chain length was not the problem. Possible reasons for poor structural parameter convergence will be explained in a later section of this chapter.

Results from Empirical Data Analysis with the SICM* Model

The results from the Foundations of Literacy Pre-Test will be provided to illustrate the estimation of the SICM* model and the types of information that can be obtained from the model. Specifically, model-data fit will be assessed and then parameter estimates will be summarized and discussed.

Model Fit

Prior to an evaluation of the estimated model and examinee parameters, the degree to which the model fit the data was assessed. Attractive properties of IRT models (see Green, Yen, & Burkett, 1989 for a list of properties) and DCMs are enabled by item and examinee parameter invariance, which is present only when the model fits the data. To assess the results of the SICM* model in an absolute sense, limited-information goodness-of-fit statistics were used. More specifically, bivariate information statistics were used to compare the observed responses to the model-predicted (or expected) responses for every pair of items. Model fit is good when the model's predictions are close to the actual data that was observed. For a pair of multiple-choice items, a $J_{i_a}\ x\ J_{i_b}$ contingency table can be formed, where J_{i_a} and J_{i_b} are the number of alternatives present for items a and b, respectively. A chi-squared statistic with $\left(J_{i_a}-1\right)(J_{i_b}-1)$ degrees of freedom can be used to test whether the observed and expected responses are statistically different. This statistics is calculated as:

$$\chi^2{}_{(J_{i_a}-1)(J_{i_b}-1)} = \sum_{j=1}^{J_{i_a}} \sum_{j'=1}^{J_{i_b}} \frac{(O_{jj'}-E_{jj'})^2}{E_{jj'}} \tag{52}$$

where $O_{jj'}$ is the observed frequency in cell jj' and $E_{jj'}$ is the frequency is cell jj' predicted by the model. Using bivariate-information goodness-of-fit statistics is a recommended and practical way to assess absolute fit for DCMs (Rupp, Templin, & Henson, 2010; Templin & Bradshaw, under review).

For each item pair on the Foundations of Literacy Pre-Test, a $4x4$ contingency table was formed and the chi-squared statistic in Equation (52) was calculated. For each of the 378 pairs of items, the hypothesis test suggested a *lack* of model-data fit. However, the χ^2distribution utilized for this hypothesis testing may inadequately approximate the test statistic distribution if there are not a sufficient number (greater than or equal to five) of expected observations in each cell of the contingency table (Agresti, 2007). For a given sample size E, the expected value of each cell decreases as the number of alternatives for an item increases; J_i alternatives per item results in

2^{J_i} cells of the table. For this assessment, four alternatives are present for each item and 1097 examinees responded to the items. For 25 item pairs, cells had expected values less than five, which renders the hypothesis test incorrect. However, for all other item pairs this test hypothesis can be used to conclude that poor model-data fit exists. To illustrate the SICM* model for didactic purposes, the analysis will proceed in spite of a lack of fit.

In a test-construction scenario where a test is being developed to be estimated with the SICM* model, a test can be piloted to determine which items exhibit bivariate misfit with many of the other items. These items can be flagged to be reviewed. Reviews can determine whether the item needs to be revised or culled. The model-data fit for the test as a whole can be improved by revising or eliminating items that demonstrate misfit. However, items should not be deleted, irrespective of content considerations, for the sole reason that they do not fit the model. Such deletions may conflict with other test construction principles and compromise the construct validity of the test (Cohen, Templin, & Bradshaw, 2009). The process of revising items is not possible in a scenario, like this one, where the test is being retrofitted to the model.

Model Parameter Interpretation

Model-data misfit precludes any valid interpretations of the model and examinee parameters estimated using the model. However, to illustrate how the SICM* model functions, sample results will be provided in-depth for a single item and a pair of examinees and followed by summaries for model and examinee parameter estimates. The examinee estimates shed light on why model-data misfit was present and why the structural parameters were unable to converge. Reasons for each of these results will be discussed after the results are provided.

Parameter Estimates for an Item

The item parameters from a single item are discussed to demonstrate how the model parameters impact the nominal response probabilities. In the SICM* model, the item response probabilities are dependent upon the error pattern an examinee has and also upon his or her continuous ability. The Q-matrix and estimated model parameters for Item 26 are given in Table 22. The specifications of the three logits modeled by the SICM* model for this item are given by error pattern in Table 23.

As the Q-matrix details, this item measured all three types of errors. Each error was measured by a single incorrect alternative. Nominal response probabilities are given by ability and by error pattern in Figure 12. Each graph in Figure 12 shows the response probabilities by ability, and there is a separate graph for each error pattern that an examinee may possess according to the model. Combinations of three errors yield a total of eight possible patterns.

The main effect for ability was below the average main effect for ability of 0.507 across all items on the test. According to SICM* Lognormal (0, 0.5) graph in Figure 11, a main effect of 0.228 for ability was one of the weaker effects of that kind. In Figure 12, this is reflected by the gradual curve of the trace line for the correct alternative (Alternative A, the dotted curve in Figure 12). As demonstrated in Figure 12, for the SICM* model, regardless of the error pattern an examinee has, as θ increases, the probability of selecting the correct answer (A) increases. For examinees with no errors (Pattern [000] in Figure 12), the probability of a correct response to this item only ranges from approximately 0.35 to 0.75 for abilities within three standard deviations of the mean ability, and for examinees with all errors (Pattern [111]) this probability only ranges from approximately 0.22 to 0.55. A higher main effect for ability would provide more discrimination amongst the probability of a correct response for examinees with low and high ability. That is, these probability ranges would have been greater had the item been more discriminating.

The order of the intercepts for incorrect Alternatives B, C, and D can be deduced from the response probabilities of the first of the eight graphs in Figure 12. This graph provides the response probabilities for examinees who have no misconceptions or errors (Pattern [000]), so the order of the preferred incorrect alternatives is determined solely by the intercept. The most likely incorrect alternative is *D*, with the largest intercept of .666, and the least likely is *B*, with the smallest intercept of -1.154.

Interpretations of trace lines for the next six graphs are difficult as *no examinees actually have the error patterns*. This undesirable result will be discussed in the following sections. Alternative D is actually the most likely incorrect alternative that examinees with the next six patterns will select. This result is a reflection of the model-data misfit. In theory, for a situation where each incorrect alternative measures a different item, the probability of an item response for an examinee who possesses a single error or a misconception should be greatest for the alternative that measures that error. For example, the probability of selecting *B* that measures the third error (the black curves in Figure 12) should be greater than the probability of selecting *C* or *D* for an examinee who has error pattern [001]. However, a glance at the second graph on the top shows this is not the case for this item. This main effect has the largest value of the three main effects for errors, but the intercept for this alternative has the smallest value, which is why this alternative is not preferred by examinees the model would predict it to be preferred by.

In the last graph, trace lines of incorrect alternatives suggest similar probabilities exist of making any one of these errors when examinees have all of the misconceptions (Pattern[111]). In other words, if examinees are making all three types of these errors, when an item is presented where each incorrect alternative measures one of those errors, the examinee has a similar probability for making each of the errors.

The main effect for the third error was large (1.705). The main effect for the first error is smaller (.681), and the main effect of the second error was near zero. The second error had very little impact on the response probabilities. In a test-construction scenario, when a main effect for a misconception is near zero for an alternative, this alternative could be rewritten to reflect a response that more strongly measures this, or another, error.

Examinee Results

In this section, example examinee parameters for two examinees will be provided first to illustrate the type of information the SICM* model provides about examinees. Then the aggregated classifications will be provided for the sample of examinees and followed by a discussion about the misfit of the model.

Examinee Parameters for Two Example Examinees

Table 24 will be used to illustrate the two distinct types of information gained from the SICM model with respect to examinee parameters. This table contains the responses to this test for two examinees, Examinee 403 and Examinee 199. These two examinees have similar scored response patterns; they answered the first 22 items correctly. Each examinee answered two of the last six items correctly, giving both a total correct score of 24. However, the two final items they answered correctly were different items, resulting in their ability estimates to be slightly different

($\theta_{403} = 2.149, \theta_{199} = 1.729$). This piece of information about examinees shows what IRT ability scores offer beyond CTT total scores; which items an examinee answers correctly impacts the score. Furthermore, for the items these examinees answered incorrectly, they selected different incorrect answers. Thus, their different incorrect answers on the last six items led to the examinees being classified as having drastically different error patterns ($\boldsymbol{\alpha}_{403} = [111]$, $\boldsymbol{\alpha}_{199} = [000]$). This piece of information about examinees illustrates what the DCM portion of the SICM* model offers beyond IRT ability scores; which incorrect alternatives an examinee selects impacts the classification. From the SICM* model estimates, we can conclude that both students have an above average ability, yet Examinee 403 needs instruction relevant to all three errors, while Examinee 199 does not.

Remember the misfit of the model that makes the use of these scores invalid, but these results still illustrate the utility that the SICM* model examinee estimates can add beyond IRT model estimates. For an IRT model, as also seen above in the SICM* model, it matters which items an examinee answers correctly. The same total score can yield different ability estimates because items are differentially related to the target ability being measured and thus count differentially towards the estimated ability. The SICM* model goes a step further and uses information not only from which items an examinee answers incorrectly, but also why the examinee answered the item incorrectly. As a result, two examinees can have the exact same scored response pattern and be classified as possessing a very different set of misconceptions. In practice, this did not occur for this assessment because for the three pairs of examinees that had the exact same scored response pattern, they also had similar nominal response patterns. However, Figure 13 illustrates that the ability distributions for examinees who have the estimated pattern of no errors (Pattern [000]) and for the examinees who have the estimated pattern of all errors (Pattern [111]) are very similar. In other words, a high ability estimate does not necessarily indicate a lack of misconceptions that are being measured. Conversely, a low ability does not necessarily indicate a presence of misconceptions that are being measured. Within a pattern, the ability estimates are spread out along the continuum.

Classification of Examinees

The results in Figure 13 are actually the ability distributions for all 1097 examinees, meaning that, as mentioned previously, the SICM* model estimates placed examinees into one of two patterns: all errors present or no errors present. The other six patterns were not observed. As noted in Templin and Henson (2006), observing a large number of examinees in patterns for which all or none of the attributes are possessed may indicate that the construct being measured is truly unidimensional. Both a theoretical and a practical explanation exist to explain this result. Theoretically, perhaps these errors did not exist; they were not latent traits of the examinees. Alternatively, perhaps multidimensionality that actually exists was not captured by the assessment, meaning in theory the errors were traits of the examinees, but in practice the test did not elicit these errors. Because the test data were retrofitted to the SICM* model, effects of theory and practice cannot be disentangled to determine which explains the results that were found. The following paragraphs will further discuss these two possible explanations.

The first reason given for examinees being classified into one of two attribute patterns was that the domain-specific theory that the SICM* model relies on may be incorrect. The theory is incompatible with the model if these errors are not traits of examinees that would manifest themselves in a systematic fashion such that a statistical model could capture and predict the response pattern. That is not to say that these are not types of errors that students sometimes make, but rather to say that the errors themselves are based more on the context of the item, question, or testing situation rather than the examinee. An analogous example may be a person who runs sometimes, but does not identify themselves as a *runner*. Classifying oneself as a runner implies certain traits that an individual has that, although could be defined in different ways, may include: a certain level of dedication to the sport, frequency of participation in the sport, or knowledge of the sport. A person may sometimes go for a run to get exercise or to enjoy the outdoors, but they do not consistently exhibit the hallmark characteristics of runner. To tie the analogy back to misconceptions and errors, just as one who runs is not necessarily a runner, an error a student makes is not necessarily a trait that can be measured with the SICM* model. The error or misconception must be a trait inherent in the individual to the point that their responses are systematically governed by the trait to a significant degree.

Alternatively, the domain-specific theory may be compatible with the model, but the test may have been constructed in a way that prohibits the errors manifesting themselves in a systematic way. Just as extensive theory must be developed to determine a set of misconceptions relevant to a given domain, extensive validity studies must be completed to verify that alternatives on an assessment are eliciting misconceptions that they purport to measure. The test development process must also attend to additional statistical considerations that may include investigating whether (a) the misconception is measured enough times (in enough alternatives and items) to yield a reliable classification, (b) enough examinees are selecting each alternative to provide enough information to yield accurate item parameter estimates for each alternative, and (c) the sample of examinees is large enough to yield accurate model parameters.

We expect for the model-data fit for an assessment designed from the onset to be measured with the SICM* model to be much better than the model-data fit observed for this assessment. Model-data fit should be improved by verifying the content validity of the alternatives (i.e., confirming the alternatives are measuring what they seek to measure) and by measuring misconceptions whose presence is illuminated and verified by extensive research.

The classification of examinees into only categories with all or no errors empirically suggests that the structural model of the SICM* model was incorrect. Because the errors were highly correlated, they were not practically distinct and could not be estimated as separate categorical variables. As a result, the structural parameters did not converge because there was no information about examinees in the six classes that were not observed. Essentially, the goal of this model was to model the variation of item responses according to predetermined patterns; however, there was no observed variation across these patterns. This also explains why, as shown in Figure 12, the nominal response probabilities for the patterns that no examinees are classified as possessing looked very similar and did not reflect expected response probabilities conditional upon the error pattern.

Comparison of Results from Empirical Data Analysis with Different Psychometric Models

In addition to the SICM* model, seven other psychometric models were used to provide examinee estimates from the Foundations of Literacy Pre-Test. The SICM model (without the lower asymptote) provided estimates of examinee ability and classifications of examinees according to errors. The NR IRT model, the NR IRT* model, the 2-PL IRT model, and the 2-PL IRT* model provided only estimates of examinee ability. This unidimensional trait was measured by each alternative in the NR IRT and NR IRT* models and by each item in the 2-PL

IRT and 2-PL IRT* models. The NR LCDM and LCDM provided only classifications of examinees according to the three types of errors on the assessment. To specify the NR LCDM, the correct alternative was taken as the baseline category and main effects for errors were constrained to be greater than zero. To specify the LCDM, main effects for errors were constrained to be less than one because the odds of answering the item correctly were modeled, as is usual for the LCDM. For the LCDM, errors were measured at the item level instead of the alternative level. Results with respect to convergence, parameter estimates, examinee parameters and relative model fit will be provided and discussed.

Convergence

Convergence results according to the variation of Gelman and Rubin's $\hat{R}$ used previously are provided for all eight models in Table 25. For each model, the percentage of parameters that converged according to the $\hat{R} < 1.1$ and $\hat{R} < 1.5$ criteria are provided, by the parameter type. All parameters for all four IRT models reached convergence. For the SICM and SICM* models, almost all main effects for ability and the ability parameters converged according to the more lenient criterion, and over 95% of the item parameters converged according the this criterion. Structural parameters had a higher convergence rate in the SICM* model as compared to the SICM model, although both rates were unacceptably low. Similarly, in the NR LCDM and LCDM models, item parameters converged at a higher rate than structural parameters, although for the NR LCDM convergence was much greater for the item parameters.

Method for Comparing Model-Data Fit

Generally, competing psychometric models can be examined to determine which model accurately represents the characteristics of the items and the test. To evaluate model fit in an absolute sense, correspondence between model predictions and observed data can be assessed at the item or test level, as explained previously for the SICM* model. To make a relative comparison of model-data fit, Akaike's information criteria (AIC; Akaike, 1974) and Schwarz's Bayesian information criteria (BIC; Schwarz, 1978) were used. These indices are not accompanied by a statistical test, but rather the respective values of each of these indices is compared for a set of models to determine which model is the better fit to the data. A model with a smaller AIC or BIC is considered to fit better. Both of these measures consider parsimony between model-data fit and model complexity. The indices incorporate a penalty for added model complexity, thereby preferring more parsimonious models. The AIC measure is not a function of sample size and, as a result, tends to favor complex models in large samples. The BIC is a function of sample size, as well as model complexity (number of estimated parameters), and tends to select simpler models. Rupp, Templin, and Henson (2010) caution against using relative fit statistics in isolation. They recommend using them subsequently to absolute model-data fit analyses for the purpose of selecting the most parsimonious model among a set models shown to have adequate absolute fit. However, the goal of the model comparisons is not to select a model to be used in practice and will have no impact upon an examinee or stakeholder, so relative model fit will be evaluated without respect to absolute model fit.

Both the AIC and BIC require the maximized model likelihood, which for MCMC estimation algorithms using uninformative prior distributions can be determined using parameter estimates defined as the mean of each parameter's respective stable posterior distribution. The AIC and BIC are defined as

$$AIC = 2m - 2(\mathrm{LL}) \tag{53}$$

$$BIC = m(\ln(E)) - 2(\mathrm{LL}) \tag{54}$$

where LL is the log likelihood of the model, m is the number of model parameters and E is the number of examinees.

For the SICM and SICM* models, the likelihood for an examinee is calculated as

$$P(\boldsymbol{X_e} = \boldsymbol{x_e}) = \int_{-\infty}^{\infty} \sum_{p=1}^{2^A} \upsilon_p \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{n_{ij}|\boldsymbol{\alpha}_{e,}\theta_e}^{[x_{ei}=n_{ij}]} P(\theta)\, \partial\, \theta \tag{55}$$

where $\boldsymbol{X_e}$ is an $1\ x\ I$ vector of nominal responses to I items on an assessment. The likelihood for the model is the product of the likelihoods for each examinee:

$$P(\boldsymbol{X} = \boldsymbol{x}) = \prod_{e=1}^{E} P(\boldsymbol{X_e} = \boldsymbol{x_e}) = \prod_{e=1}^{E} \int_{-\infty}^{\infty} \sum_{p=1}^{2^A} \upsilon_p \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{n_{ij}|\alpha_{e,}\theta_e}^{[x_{ei}=n_{ij}]} P(\theta)\, \partial\, \theta \tag{56}$$

where $\boldsymbol{X}$ is an $E\ x\ I$ matrix of E examinees' nominal responses to I items on an assessment. Because $P(\boldsymbol{X_e})$ is independent of $P(\boldsymbol{x_{e'}})$ for any two examinees (e and e'), the joint probability of observing a set of examinees' responses is the product of the marginal probability of observing each examinee's responses. The model likelihood was approximated to avoid complex numerical integration. Instead of integrating over the ability distribution with respect to ability, the ability distribution was divided into R different quadrature points (rectangular regions), the area of which represented the probability that a given examinee's ability was within that range of ability. By summing across the R quadrature points, an approximation of the integral reflecting the cumulative distribution function for ability was obtained. As R increases, the interval decreases, and as R approaches infinity, the sum of the quadrature points approach the integral of the probability density function of ability. The log likelihood for all examinees is the sum of the log likelihood for each examinee:

$$\mathrm{LL_{SICM}} = \log\left(\prod_{e=1}^{E}\int_{-\infty}^{\infty}\sum_{p=1}^{2^A}\upsilon_p\prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{n_{ij}|\alpha_{e,}\theta_e}^{[x_{ei}=n_{ij}]}P(\theta)\,\partial\,\theta\right) \approx \tag{57}$$

$$\sum_{\mathrm{e=1}}^{\mathrm{E}}\log(\sum_{r=1}^{R}\sum_{p=1}^{2^A}\upsilon_p(\boldsymbol{\alpha})P_r(\theta)\prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{n_{ij}|\alpha_{e,}\theta_e}^{[x_{ei}=n_{ij}]}).$$

For the SICM model, $R = 50$ was used for determining the model-log likelihood.

Similarly, the model log likelihoods for each model whose relative fit was compared were calculated. The log likelihoods for the NR LCDM and NR IRT model are expressed in Equations (58) and (59), respectively:

$$\mathrm{LL_{NRLCDM}} = \log\left(\prod_{e=1}^{E}\sum_{p=1}^{2^A}\upsilon_p\prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{n_{ij}|\alpha_e}^{[x_{ei}=n_{ij}]}\right) \tag{58}$$

$$\mathrm{LL_{NRIRT}} = \log\left(\prod_{e=1}^{E}\int_{-\infty}^{\infty}\prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{n_{ij}|\theta_e}^{[x_{ei}=n_{ij}]}P(\theta)\,\partial\,\theta\right) \approx \tag{59}$$

$$\sum_{\mathrm{e=1}}^{\mathrm{E}}\log(\sum_{r=1}^{R}P_r(\theta)\prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{n_{ij}|\theta_e}^{[x_{ei}=n_{ij}]}).$$

Model-data Fit Comparisons

Table 29 displays the number of parameters estimated by each of the eight models along with the log likelihood, AIC and BIC values. The relative rank of model-data fit according to each index is provided after the value of the index is given. For both measures, the 2-PL IRT model demonstrated the best fit, closely followed by the 2-PL IRT* model. These two models measured considerably fewer parameters than the other models. The AIC and BIC values do not agree on the next best-fitting model, although it is among the SICM* and NR IRT models. The AIC and BIC values for these models are close to the values for the other versions of these two models (SICM and NR IRT*). The models with the worst model-data fit were the NR LCDM and the LCDM.

These results were not surprising. Given that the results of the examinees' all-or-none error patterns found for the SICM* model, it is expected that models estimating high dimensionality with respect to the errors would not be preferred. Instead, models estimating a continuous unidimensional trait are expected to be preferred. The unidimensional models with the fewest number of estimated parameters, the 2-PL IRT and 2-PL IRT* models, were the most preferred models. However, the NR IRT models were not clearly preferred over the SICM models. When modeling the nominal responses, the errors did demonstrate some dimensionality, as examinees were in one of two classes, not just one class. However, the complete dimensionality that was modeled by the SICM models resulted in over-fitting the model. If no dimensionality with respect to the errors was preferred, the NR IRT models should uniformly be preferred to the SICM models because they estimate far fewer parameters. These results suggest the test may be measuring something more than a unidimensional trait, but it is not measuring the three distinct errors in the Q-matrix. Perhaps a single error that placed examinees into two groups would be preferred to three errors that failed to place examinees into eight groups.

The result that the LCDM was the least preferred model was also expected. The specification of the LCDM in this way was not theoretically sound, so model-data fit was not theoretically expected for this model. Other evidence suggested the LCDM model parameters did not converge (see Table 25), and the results in the next section will show that extreme parameters had some extreme values (as will be seen in Table 26) and incongruent classifications with other models (as will be seen in Figure 17). Because the types of errors appear in different incorrect alternatives, collapsing these incorrect alternatives into a single category of an incorrect response means the model does not have the ability to distinguish between the types of errors being made; that information is only provided with the nominal response data. To specify this model, any error present in one of the three incorrect alternatives was measured by the item. Therefore, if the item is missed, each present error has a main effect even though the reason for the item being missed is only due to one type of error—the type aligned to the error measured in the nominal response that was selected. The LCDM was estimated for this study to illustrate this point. Although an error or misconception is simply an attribute redefined as an inability instead of an ability, unless every incorrect alternative measures the same error within each item, the LCDM is ill-equipped to measure misconceptions or errors. Collapsing the alternatives loses all of the necessary information to measure the alternative-specific misconceptions.

A caveat to comparing these models with these indices is that some of these models have very different purposes (e.g., provide estimates of a continuous unidimensional ability versus classify examinees according to a set of categorical traits). Model selection should consider alignment of the type of information that is needed or desired from the assessment and the type of information that the psychometric model is engineered to provide. Statistical relative fit is not a sole reason upon which to base model selection.

Parameter Estimates and Standard Errors

The following sections will compare the examinee estimates for the different models. For future reference to discuss model results, the average estimated item and structural parameters and examinee abilities are given in Table 26, along with their standard errors (in parentheses). Blanks in the table indicate that no parameter of that type was estimated for that model.

Comparison of Ability Estimates

Figure 14 displays the distributions of estimated abilities for examinees for each of the models that estimated a continuous parameter that represented examinee ability. Comparing these histograms shows that the models with the lower asymptotes yielded a slightly more left-skewed distribution of ability parameters. However, in general the distributions had a slight left-skew, indicated by more examinees with abilities in the interval (1.5,3) than in (-3, -1.5). Average ability estimates for each model are given in the last column of Table 29.

Correlations of ability estimates from these models were also examined. The more highly correlated the estimates are, the more closely the different models ranked the examinees. Practically, this information becomes important in answering the question of whether adding misconceptions to a model will alter the rank of examinees. For example, if misconceptions are measured in addition to a continuous ability on an end-of-grade federally mandated test to provide multidimensional feedback, it is important to know how modeling misconceptions will impact the estimate of the examinees’ ability. Significantly altering the rank may alter recipients or winners of scholarships or awards. The effects on ability estimation could be examined with an with an additional simulation study. The results here show how correlated the estimates were for this test as a practical example of the effects. Comparisons are made under the caution of drastic misfit of the SICM* model.

Figure 15 shows a plot of SICM* ability estimates in comparison to, first itself (to facilitate relative comparisons to a perfect correlation) and then to each of the other five models that estimated a continuous ability for examinees. The SICM* model estimates are more highly correlated with the lower asymptote versions of the NR IRT and 2-PL IRT models, and most highly correlated with the SICM model.

Correlations for these estimates, along with correlations among all model pairs, are given in Table 27. The highest correlations were between the regular and lower asymptote versions of the SICM, NR-IRT and 2-PL IRT models. These high correlations (0.915, 0.988, and 0.994, respectively) are reflected by the plots of examinee ability estimates given in Figure 16 and indicate that the lower asymptote did not considerably alter the ranks of examinees' ability estimates. The next highest correlations were among the four IRT models. The lowest were correlations of the SICM models with the NR IRT models. SICM models may be more correlated with 2-PL IRT models than NR IRT models because NR IRT models have alternative-specific discrimination parameters for ability, while the SICM and 2-PL IRT models are constrained to a single discrimination parameter for ability per item.

Comparison of Classifications

Figure 17 displays the classifications of examinees for each of the models that measured a set of categorical variables that represented errors. As for the SICM* model, the SICM model places examinees into one of two error patterns. Examinees either possessed all errors (Pattern 8, [111]) or no errors (Pattern 1, [000]). The NR LCDM has similar classifications, with only six examinees classified as having Pattern 2 ([001]) and two examinees as having Pattern 3 ([010]). The classifications from the LCDM were very different; examinees were classified as possessing one of three patterns [001], [011], or [110]. Table 28 displays the percentage of individual misconception and whole pattern classifications that pairs of models agreed upon. The SICM and

SICM* models classified 90.79% of examinees to the same error pattern. Although how much the lower asymptote affects classification is an empirical question, classification was not expected to vary greatly between these two models. The SICM* and NR LCDM models were in agreement with respect to individual misconception and whole pattern classification for approximately 84% of examinees. In a practical sense, understanding how the presence of a continuous ability impacts classification is important, although this too is an empirical question that cannot be answered by an example application that demonstrates model-data misfit. Agreement with respect to the LCDM is drastically lower, explanations for which were provided in the previous discussion of why the data does not fit the LCDM.

The findings with the NR LCDM and SICM models reiterate what was found in the SICM* model: high multidimensionality in the assessment with respect to the errors does not exist. Again, this can either be due to an issue of these errors not being a stable trait or to an issue of test development. The SICM and NR LCDM also had low convergence rates of the structural parameters, as seen in Table 28, and extreme values of the structural parameter main effects and interactions, as seen in Table 29. As for the SICM* model, these results are reflective of the misspecification of the structural model due to the lack of multidimensionality of the assessment. The models could not differentiate the main effect and interactions for the errors because the data represented no systematic variation due to these traits.

Discussion for Empirical Data Analysis

Although model-data fit was not present for the SICM model and these data, steps were taken to highlight considerations that an analysis of a test with the SICM model should include and to demonstrate the types of information that the model can provide. Consideration was given to not presenting the analysis from these data and turning to another data set that might bear “better” results with the SICM model. However, there is value in sharing the story this analysis tells. This analysis shows that even in a scenario like this one where the purpose of the assessment was aligned with the purpose of the model, limitations exist and issues arise when retrofitting an assessment to a model. Model-data fit is expected to improve in a test-construction scenario where a test is being developed from the onset to be estimated with the SICM* model and is thus recommended.

Retrofitting the Foundations of Literacy Assessment to the SICM model was a limitation to this study. One criticism in applications of DCMs is that assessments to which the models have been fit were not created for the purpose of classifying individuals according to a set of dichotomous attributes. Although applications of DCMs are in progress (e.g., Izsak et al., 2010; Patterson 2011), DCMs meanwhile have been fitted to existing assessments to illustrate their characteristics and capabilities. Development of the psychometric theory of a model necessarily precedes its applications. However, several limitations arise when assessments that are constructed within one psychometric framework are modeled with another.

With respect to the Foundations of Literacy Pre-Test, the goals of the test was very well-aligned with the features of the SICM model; however statistical and theoretical considerations mentioned previously in this chapter should be emphasized from the beginning of the test development process. Retrofitting other tests, designed with a completely different purpose in mind, is heavily cautioned. For example, if the SICM model was fitted to data from an IRT-based assessment, the mission essentially would be to extract estimates on multiple dimensions for classification purposes from an assessment that purposefully attempted to eliminate such dimensional information via the test construction process. Templin and Henson (2006) note:

> Attempting to obtain classification results from a single latent continuum can result in estimates that a large number of individuals possess either all attributes or none of the attributes. In this case, all attribute correlations (from the structural model) would approach unity, an indication that a single continuum is truly underlying the data (pg. 302).

Assuming that one could educe commensurate dimensionality from the assessment, other issues of retrofitting the assessment are present. First, determining the set of misconceptions that are measured by the assessment is non-trivial with respect to the number, nature, and grain-size of the misconceptions being defined. The number of misconceptions that are measured is a compromise between statistical feasibility and domain-specific theory fidelity. For DCMs, incorrectly specifying the attributes required to answer an item correctly has negative consequences, whether the misspecification is due to including attributes when they are not required or to omitting attributes that are indeed required (Rupp and Templin, 2008). These consequences are expected for the DCM portion of the SICM models as well. Evidence must be provided that examinees do in fact elicit these misconceptions when selecting corresponding incorrect alternatives. The validity of the estimates, inferences and decisions that are made based upon the DCM is dependent upon the correct specification of which attribute or combination of attributes are required to answer each item (Rupp and Templin, 2007); this is true for the SICM model as well. Therefore, incorrect specification of attributes or entries in the Q-matrix will negatively impact the fit of the SICM model to the IRT-designed test.

Lastly, retrofitting the SICM model to a test will likely fail to capitalize on the salient, useful characteristics of the SICM model. The SICM model can readily handle complex structures and model more than one misconception being measured with a single alternative. If items were not written from the SICM framework, they may not be crafted to measure the misconceptions as many times as possible. The six assessments that were outlined in Chapter 3 provide examples of assessments whose items have alternatives that were consciously written to be aligned to a common misconception. However, none of these assessments have alternatives that measure more than one misconception. If alternatives or items did measure multiple misconceptions, equally reliable classifications may be obtained with shorter tests or more reliable classifications may be provided with equally long tests.

Concluding Remarks of Dissertation

The SICM model was presented as a psychometric solution to a realistic need in educational assessment—to gain more feedback from assessments about what students do not understand. As discussed previously, researchers have developed multiple-choice assessments with incorrect alternatives carefully constructed to determine which common misconceptions students have. Existing psychometric models do not quantify examinee misconceptions and commonly focus only on overall abilities. The SICM model provides a way for researchers who have these aims to capitalize statistically on their complex assessment designs. The model incorporates misconceptions as latent predictors of an item response and is able, in turn, to provide estimates of misconceptions for an examinee.

The efficacy of the SICM model under various testing conditions was demonstrated through a simulation study. Thus, the SICM model can enable diagnostic score reports that reflect statistical estimates of student misconceptions *in addition to* information about student ability that is typically provided to stakeholders by current modeling and testing procedures.

Whereas the simulation results provided guidelines for test and sampling conditions, they did not give guidance in terms of how to create the test itself. As seen in the empirical data analysis, the development of a test from the SICM framework *a priori* is very important.

Although some general test-development considerations can be applied in developing an assessment for the SICM model, open questions still exist as to how to create an assessment that can utilize the statistical features of the SICM model. Assessments exist that can provide insights for writing items to measure misconceptions with incorrect answers. When the SICM model is employed to model these types of items, however, unique statistical considerations arise. For unidimensional IRT models, items that exhibit multidimensionality are often screened and revised or deleted from the assessment. In contrast, the SICM model uses items that measure a single continuous trait and a set of multidimensional categorical traits. Thus, items are expected to show multidimensionality and have to be screened differently.

This dissertation explains how the SICM model can be estimated and applied in empirical testing situations. Future assessment development projects can hopefully build upon this information to leverage the SICM model in practical settings to provide actionable information about where students' misunderstandings lie.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22*(3), 37-53.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

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

Alexeev, N., Templin, J., & Cohen, A. S. (in press). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement*.

Bauer, D. J., & Curran, P.J. (2003). Distributional assumptions of growth mixture models: Implications for over-extraction of latent trajectory classes. *Psychological Methods, 8,* 338-364.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores.* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972) Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37,* 29-51.

Briggs, Alonzo, Schwab, & Wilson (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33-63.

Buros, O. K. (1977). Fifty years in testing: Some reminiscences, criticisms and suggestions. *Educational Researcher, 6*(7), 9-15.

Chen, F. F., West, Stephen G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41* (2), 189-225.

Choi, H.-J. (2009). *A diagnostic mixture classification model*. (Unpublished doctoral dissertation). University of Georgia, Athens, GA.

Cohen, A. S., Templin, J. L., & Bradshaw, L. P. (2009). *Beyond unidimensionality: Measuring all of achievement.* Paper presented at the annual National Council on Measurement in Education conference in San Diego, California.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Wadsworth Group/ Thomson Learning.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

de Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement, 23,* 3-19.

DeMars, C. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement, 27,* 275-288.

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* (p. 361-

389). Hillsdale, NJ: Lawrence Erlbaum Associates.

Garfield, J. (1998, April). Challenges in assessing statistical reasoning. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Garfield, J., & Chance, B. (2000). Assessment in statistics education: Issues and challenges. *Mathematical Thinking and Learning*, *2*, 99-125.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

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

Green, D.R., Yen, W.M., & Burkett, G.R. (1989). Experiences in the application of item response theory in test construction. *Applied Measurement in Education, 2*(4), 297-312.

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*, 79-95.

Haertel, E. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.

Halloun, A. & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics, 53* (11), 1043-1055.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

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

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

Henson, R. A., & Templin, J.L. (2003). The moving window family of proposal distributions. Educational Testing Service, External Diagnostic Research Group, Unpublished Technical Report.

Henson, R., & Templin, J. (2005). Hierarchical log-linear modeling of the joint skill distribution. External Diagnostic Research Group Technical Report.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. *Psychometrika, 74*, 191-210.

Henson, R., Templin, J., & Irwin, P. (2009, April). Ancillary random effects: A way to obtain diagnostic information from existing large scale tests. Paper presented at the annual meeting of the National Council on Measurement in Education in San Diego, California.

Hestenes, D.,Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*, 141–151.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*, 41-54.

Huff, K. & Goodman, D.P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). Cambridge University Press.

Izsák, A., Lobato, J., Druken, B., Orrill, C., Jacobson, E., & Bradshaw, L. (2010, July). *Applying cognitive diagnosis models to measure middle grades teachers' multiplicative reasoning.* Paper presented at the annual International Meeting of the Psychometric Society in Athens, GA.

Izsák, A., Lobato, J., Orrill, C. H., Cohen, A. S., & Templin, J. (2009, February). Psychometric Models and Assessments of Teacher Knowledge. *Proceedings for the Twelfth Special Interest Group of the Mathematical Association of America on Research in Undergraduate Mathematics Education*. Paper retrieved July 2, 2009 from http://rume.org/crume2009/proceedings.html.

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

Khazanov, L. (2008). Addressing students' misconceptions about probability during the first years in college. *Mathematics and Computer Education*, *42*(3), 180-192.

Khazanov, L. (2009, February). *A diagnostic assessment for misconceptions in probability*. Paper presented at the annual Georgia Perimeter College Mathematics Conference in Clarkston, GA.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Lazarsfeld, P. F. (1950). Chapters 10 and 11 in S.A. Stouffer et al. (Eds.), *Studies in social psychology in World War II: Vol 4. Measurement and Prediction.* Princeton, NJ: Princeton University Press.

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

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*(2), 3–16.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*(2), 117-138.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 197-212.

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

Muthén, L. K., & Muthén, B. O.(1998-2010). Mplus user's guide. (5th ed.). Los Angeles, CA: Muthén & Muthén.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Research Council. (2010). State Assessment Systems: Exploring Best Practices and Innovations: Summary of Two Workshops. Alexandra Beatty, Rapporteur. Committee on Best Practices for State Assessment Systems: Improving Assessment While Revisiting Standards. Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*(4), 575-603.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, 115 Stat/ 1449-1452 (2002).

Patterson, J. (2011, February). *Cognitive diagnosis modeling of introductory statistics.* Colloquium given at the University of Georgia Statistics Department in Athens, GA.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnostics and output analysis. *R News*, *6* (1), 7-11.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25-36.

Rupp, A. A. (2007). The answer is in the question: A guide for investigating the theoretical potentials and practical limitations of cognitive psychometric models. *International Journal of Testing*, *7*, 95-125.

Rupp, A. A., & Templin, J. (2007). Effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement, 68,* 78-98.

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

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

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching, 35*, 265.

Sadler, P.M., Coyle, H., Miller, J.L., Cook-Smith, N., Dussault, M., & Gould, R.R. (2010). The Astronomy and Space Science Concept Inventory: Development and validation of assessment instruments aligned with the K-12 National Science Standards. *Astronomy Education Review*, *8*.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*(4), 21-28.

Stouffer, S.A. (1950). The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star & J.A. Clausen (Eds.), *Measurement and prediction.* (pp. 3-45). Princeton, NJ: Princeton University Press.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345-354.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnoses. In N. Frederiksen, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition.* Hillsdale, NJ: Erlbaum.

Templin, J. (2004). *Generalized linear mixed proficiency models* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Champaign, IL.

Templin, J. (2006). *CDM user's guide.* Unpublished manuscript.

Templin, J., & Bradshaw, L. (under review). *Diagnostic models for nominal response data.* Manuscript under review.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11,* 287-305.

Templin, J., & Bradshaw, L. (in press). The comparative reliability of diagnostic model examinee estimates. *Journal of Classification.*

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika, 49,* 501-519.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distracters are also part of the item. *Journal of Educational Measurement, 26* (2), 161- 176.

van der Linden, W. J. & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern Item Response Theory*. New York: Springer.

von Davier, M. (2005). A General Diagnostic Model applied to language testing data. RR-05-16. Research Report. Educational Testing Service. Princeton.

Wilson, M. (1992) The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement, 16,* 309-325.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software International.

Table 1

Class-dependent Values of the Main Effect and Interactions in LCDM Kernel for Example Item

$\boldsymbol{\alpha}_c$	$\boldsymbol{\lambda}_3^T\boldsymbol{h}(\boldsymbol{\alpha}_c, \boldsymbol{q}_3)$
(1,1,0); (1,1,1)	$\lambda_{3,1(1)}+\lambda_{3,1(2)}+\lambda_{3,2(12)}$
(1,0,0); (1,0,1)	$\lambda_{3,1(1)}$
(0,1,0); (0,1,1)	$\lambda_{3,1(2)}$
(0,0,0); (0,0,1)	0

Table 2

Q-matrix for NR LCDM and SICM model for Example Item in Figure 4

NR LCDM				SICM				
Alternative	α_1	α_2	α_3	Alternative	α_1	α_2	α_3	θ
A	1	1	0	*A*	0	0	0	1
B	1	0	0	*B*	0	1	0	0
C	0	1	0	*C*	1	0	0	0
D	0	0	0	*D*	1	1	0	0

Table 3

Simulation Design and Conditions

Characteristics		**Value or Interval**	
Test	Number of Items	30,60	
Sample	Number of Examinees	3000; 10,000	
Measurement Model		Low	High
	Sampling interval for intercepts	(-1, 1)	
	Sampling interval for α main effects	(.75, 1.25)	(1.75, 2.25)
	Sampling interval for λ_{θ}	(.3, .5)	(.6, .8)
	Sampling interval for two-way interaction effects	(0.5, 1)	
	Higher-order interactions	0	
Structural Model	Number of Attributes (Misconceptions)	3,6	
	Tetrachoric Correlation among Attributes	0.25,0.50	
	Distribution of Continuous Trait	**N**(0,1)	

Table 4

Q-matrix for 3 Attribute, 30 Item Test

Item/						Item/						Item/					
Alternative		α_1	α_2	α_3	θ	Alternative		α_1	α_2	α_3	θ	Alternative		α_1	α_2	α_3	θ
1	a	1	0	0	0	11	a	1	0	0	0	21	a	1	0	0	0
1	b	1	0	0	0	11	b	0	1	0	0	21	b	1	0	0	0
1	c	1	0	0	0	11	c	1	1	0	0	21	c	1	1	0	0
1	d	0	0	0	1	11	d	0	0	0	1	21	d	0	0	0	1
2	a	0	1	0	0	12	a	1	0	0	0	22	a	0	1	0	0
2	b	0	1	0	0	12	b	0	0	1	0	22	b	0	1	0	0
2	c	0	1	0	0	12	c	1	0	1	0	22	c	0	1	1	0
2	d	0	0	0	1	12	d	0	0	0	1	22	d	0	0	0	1
3	a	0	0	1	0	13	a	0	1	0	0	23	a	0	0	1	0
3	b	0	0	1	0	13	b	0	0	1	0	23	b	0	0	1	0
3	c	0	0	1	0	13	c	0	1	1	0	23	c	1	0	1	0
3	d	0	0	0	1	13	d	0	0	0	1	23	d	0	0	0	1
4	a	1	0	0	0	14	a	1	0	0	0	24	a	1	0	0	0
4	b	1	0	0	0	14	b	1	0	0	0	24	b	1	0	0	0
4	c	0	1	0	0	14	c	1	1	0	0	24	c	0	1	0	0
4	d	0	0	0	1	14	d	0	0	0	1	24	d	0	0	0	1
5	a	1	0	0	0	15	a	0	1	0	0	25	a	1	0	0	0
5	b	1	0	0	0	15	b	0	1	0	0	25	b	1	0	0	0
5	c	0	0	1	0	15	c	0	1	1	0	25	c	0	0	1	0
5	d	0	0	0	1	15	d	0	0	0	1	25	d	0	0	0	1
6	a	0	1	0	0	16	a	0	0	1	0	26	a	0	1	0	0
6	b	0	1	0	0	16	b	0	0	1	0	26	b	0	1	0	0
6	c	1	0	0	0	16	c	1	0	1	0	26	c	1	0	0	0
6	d	0	0	0	1	16	d	0	0	0	1	26	d	0	0	0	1
7	a	0	1	0	0	17	a	1	0	0	0	27	a	0	1	0	0
7	b	0	1	0	0	17	b	1	0	0	0	27	b	0	1	0	0
7	c	0	0	1	0	17	c	0	1	1	0	27	c	0	0	1	0
7	d	0	0	0	1	17	d	0	0	0	1	27	d	0	0	0	1
8	a	0	0	1	0	18	a	0	1	0	0	28	a	0	0	1	0
8	b	0	0	1	0	18	b	0	1	0	0	28	b	0	0	1	0
8	c	1	0	0	0	18	c	1	0	1	0	28	c	1	0	0	0
8	d	0	0	0	1	18	d	0	0	0	1	28	d	0	0	0	1
9	a	0	0	1	0	19	a	0	0	1	0	29	a	0	0	1	0
9	b	0	0	1	0	19	b	0	0	1	0	29	b	0	0	1	0
9	c	0	1	0	0	19	c	1	1	0	0	29	c	0	1	0	0
9	d	0	0	0	1	19	d	0	0	0	1	29	d	0	0	0	1
10	a	1	0	0	0	20	a	1	0	0	0	30	a	1	0	0	0
10	b	0	1	0	0	20	b	0	1	0	0	30	b	0	1	0	0
10	c	0	0	1	0	20	c	0	0	1	0	30	c	0	0	1	0
10	d	0	0	0	1	20	d	0	0	0	1	30	d	0	0	0	1

Table 5

Q-matrix for 6 Attribute, 30 Item Test

Item/Alternative		α_1	α_2	α_3	α_4	α_5	α_6	Item/Alternative		α_1	α_2	α_3	α_4	α_5	α_6
1	a	1	0	0	0	0	0	16	a	0	0	0	1	0	0
1	b	1	0	0	0	0	0	16	b	0	0	0	1	0	0
1	c	1	0	0	0	0	0	16	c	0	0	0	0	1	1
2	a	0	1	0	0	0	0	17	a	0	0	0	0	1	0
2	b	0	1	0	0	0	0	17	b	0	0	0	0	1	0
2	c	0	1	0	0	0	0	17	c	0	0	0	0	0	1
3	a	0	0	1	0	0	0	18	a	0	0	0	0	0	1
3	b	0	0	1	0	0	0	18	b	0	0	0	0	0	1
3	c	0	0	1	0	0	0	18	c	1	0	0	0	0	0
4	a	0	0	0	1	0	0	19	a	1	0	0	0	0	0
4	b	0	0	0	1	0	0	19	b	0	0	1	0	0	0
4	c	0	0	0	1	0	0	19	c	0	0	0	0	1	0
5	a	0	0	0	0	1	0	20	a	0	1	0	0	0	0
5	b	0	0	0	0	1	0	20	b	0	0	0	1	0	0
5	c	0	0	0	0	1	0	20	c	0	0	0	0	0	1
6	a	0	0	0	0	0	1	21	a	1	0	0	0	0	0
6	b	0	0	0	0	0	1	21	b	0	1	0	0	0	0
6	c	0	0	0	0	0	1	21	c	1	1	0	0	0	0
7	a	1	0	0	0	0	0	22	a	0	0	1	0	0	0
7	b	0	1	0	0	0	0	22	b	0	0	0	1	0	0
7	c	0	0	1	0	0	0	22	c	0	0	1	1	0	0
8	a	0	1	0	0	0	0	23	a	0	0	0	0	1	0
8	b	0	0	1	0	0	0	23	b	0	0	0	0	0	1
8	c	0	0	0	1	0	0	23	c	0	0	0	0	1	1
9	a	0	0	1	0	0	0	23	d	0	0	0	0	0	0
9	b	0	0	0	1	0	0	24	a	1	0	0	0	0	0
9	c	0	0	0	0	1	0	24	b	0	0	1	0	0	0
9	d	0	0	0	0	0	0	24	c	1	0	1	0	0	0
10	a	0	0	0	1	0	0	25	a	0	1	0	0	0	0
10	b	0	0	0	0	1	0	25	b	0	0	0	1	0	0
10	c	0	0	0	0	0	1	25	c	0	1	0	1	0	0
11	a	0	0	0	0	1	0	26	a	0	0	1	0	0	0
11	b	0	0	0	0	0	1	26	b	0	0	0	0	1	0
11	c	1	0	0	0	0	0	26	c	0	0	1	0	1	0
12	a	0	0	0	0	0	1	27	a	0	0	0	1	0	0
12	b	1	0	0	0	0	0	27	b	0	0	0	0	0	1
12	c	0	1	0	0	0	0	27	c	0	0	0	1	0	1
13	a	1	0	0	0	0	0	28	a	0	0	0	0	1	0
13	b	1	0	0	0	0	0	28	b	1	0	0	0	0	0
13	c	0	1	0	0	0	0	28	c	1	0	0	0	1	0
14	a	0	1	0	0	0	0	29	a	0	0	0	0	0	1
14	b	0	1	0	0	0	0	29	b	0	1	0	0	0	0
14	c	0	0	1	1	0	0	29	c	0	1	0	0	0	1
15	a	0	0	1	0	0	0	30	a	1	0	0	0	0	0
15	b	0	0	1	0	0	0	30	b	0	1	0	0	0	0
15	c	0	0	0	1	1	0	30	c	0	0	1	0	0	0

Table 6a

Convergence Rates for Simulation Conditions with Misconception Correlation of .50

Parameter Type	*E*	*I*	*A*	Mean $\hat{R}$	Median $\hat{R}$	% $\hat{R}$ < 1.1	% $\hat{R}$ <1.5
Item (Intercepts, Misconception Main Effects, Misconception Interactions)	3000	30	3	1.250	1.098	52.00	86.25
			6	1.368	1.143	43.25	78.75
		60	3	1.170	1.068	61.00	91.25
			6	1.238	1.085	55.00	87.50
	10000	30	3	1.155	1.065	62.75	92.75
			6	1.260	1.100	51.75	85.75
		60	3	1.100	1.045	72.25	96.75
			6	1.145	1.060	65.00	93.50
Item (Ability Main Effects)	3000	30	3	1.155	1.085	61.50	91.75
			6	1.190	1.108	58.50	88.50
		60	3	1.083	1.053	74.00	98.50
			6	1.088	1.053	72.50	98.25
	10000	30	3	1.153	1.080	60.75	92.50
			6	1.183	1.095	59.25	89.25
		60	3	1.075	1.048	76.75	98.75
			6	1.080	1.048	75.00	98.50
Structural (All)	3000	30	3	1.183	1.118	62.50	89.50
			6	1.725	1.448	29.75	63.25
		60	3	1.093	1.055	75.25	96.25
			6	1.338	1.198	44.25	79.00
	10000	30	3	1.240	1.160	53.00	85.00
			6	1.853	1.538	23.25	55.00
		60	3	1.155	1.105	60.75	91.75
			6	1.440	1.250	36.75	72.50
Examinee (Ability)	3000	30	3	1.220	1.125	48.25	88.00
			6	1.233	1.138	47.00	85.75
		60	3	1.163	1.090	57.00	92.25
			6	1.170	1.095	55.75	91.50
	10000	30	3	1.223	1.125	49.50	86.50
			6	1.230	1.133	45.50	86.50
		60	3	1.150	1.083	59.75	93.25
			6	1.170	1.090	57.75	92.00

Note. A variation of Gelman and Rubin's (1992) $\hat{R}$ statistic is used to assess convergence.

Table 6b

Convergence Rates for Simulation Conditions with Misconception Correlation of .25

Parameter Type	E	I	A	Mean $\hat{R}$	Median $\hat{R}$	% $\hat{R}$ < 1.1	% $\hat{R}$ < 1.5
Item (Intercepts, Misconception Main Effects, Misconception Interactions)	3000	30	3	1.253	1.098	52.25	86.25
			6	1.393	1.158	41.50	77.00
		60	3	1.170	1.068	61.00	91.50
			6	1.243	1.090	53.50	87.00
	10000	30	3	1.155	1.065	62.75	92.75
			6	1.263	1.105	50.75	85.50
		60	3	1.103	1.045	71.50	96.75
			6	1.148	1.060	64.75	93.5
Item (Ability Main Effects)	3000	30	3	1.148	1.085	62.00	92.50
			6	1.180	1.098	58.00	89.75
		60	3	1.070	1.040	77.50	99.50
			6	1.078	1.043	76.25	98.75
	10000	30	3	1.145	1.075	62.25	93.25
			6	1.180	1.095	59.75	89.25
		60	3	1.070	1.045	77.75	99.5
			6	1.073	1.045	76.00	99.00
Structural (All)	3000	30	3	1.180	1.113	62.50	89.75
			6	1.510	1.300	37.25	70.75
		60	3	1.098	1.060	74.25	95.75
			6	1.215	1.115	54.75	87.50
	10000	30	3	1.243	1.165	55.00	84.75
			6	1.643	1.400	29.50	64.50
		60	3	1.128	1.080	67.00	94.25
			6	1.305	1.170	44.25	80.25
Examinee (Ability)	3000	30	3	1.223	1.115	48.00	87.25
			6	1.228	1.133	46.75	86.00
		60	3	1.163	1.093	58.25	92.00
			6	1.170	1.098	55.75	91.25
	10000	30	3	1.215	1.125	47.75	87.75
			6	1.228	1.138	47.00	86.00
		60	3	1.168	1.093	56.25	91.25
			6	1.185	1.100	54.75	90.75

Note. A variation of Gelman and Rubin’s (1992) $\hat{R}$ statistic is used to assess convergence.

Table 7

Number of Parameters in Simulation Conditions by Number of Items (I) and Attributes (A)

I	A	Item Intercepts (λ_0)	Misconception Item Parameters ($\lambda_{1(a);}\lambda_{2(a,b)}$)	Ability Item Parameters ($\lambda_{1(\theta)}$)	Structural Parameters (γ)	Total
30	3	90	114	30	6	240
	6	90	114	30	22	256
60	3	180	228	60	6	474
	6	180	228	60	22	490

Table 8a

Estimation Accuracy for Item Parameters across Magnitude of Main Effects Factor: Misconception Correlation of .50

τ	E	I	A	$Bias\ (\hat{\tau})$	$RMSE\ (\hat{\tau})$	$r(\hat{\tau},\tau)$	$SE(\hat{\tau})$
λ_0	3000	30	3	-0.009 (0.023)	0.013 (0.003)	.983 (.004)	0.103 (0.002)
			6	-0.012 (0.024)	0.015 (0.004)	.981 (.004)	0.107 (0.003)
		60	3	-0.010 (0.014)	0.010 (0.001)	.860 (.002)	0.094 (0.001)
			6	-0.010 (0.013)	0.011 (0.001)	.986 (.002)	0.095 (0.001)
	10000	30	3	-0.002 (0.011)	0.004 (0.001)	.995 (.001)	0.057 (0.001)
			6	-0.003 (0.012)	0.004 (0.001)	.994 (.001)	0.058 (0.001)
		60	3	-0.004 (0.009)	0.003 (0.000)	.996 (.001)	0.052 (0.000)
			6	-0.004 (0.008)	0.003 (0.000)	.996 (.000)	0.052 (0.000)
$\lambda_{1(\alpha)}$	3000	30	3	0.002 (0.026)	0.030 (0.006)	.657 (.056)	0.149 (0.003)
			6	0.003 (0.030)	0.046 (0.010)	.561 (.073)	0.173 (0.004)
		60	3	0.005 (0.014)	0.021 (0.003)	.707 (.034)	0.133 (0.001)
			6	0.000 (0.017)	0.028 (0.004)	.658 (.039)	0.147 (0.002)
	10000	30	3	0.000 (0.012)	0.008 (0.002)	.853 (.024)	0.083 (0.001)
			6	-0.001 (0.016)	0.015 (0.003)	.767 (.046)	0.099 (0.002)
		60	3	0.002 (0.008)	0.006 (0.001)	.879 (.015)	0.073 (0.001)
			6	0.000 (0.010)	0.008 (0.001)	.850 (.021)	0.082 (0.001)
$\lambda_{2(\alpha,b)}$	3000	30	3	0.014 (0.088)	0.098 (0.038)	.423 (.250)	0.245 (0.013)
			6	0.018 (0.134)	0.228 (0.098)	.313 (.249)	0.324 (0.024)
		60	3	0.007 (0.047)	0.055 (0.017)	.541 (.133)	0.209 (0.007)
			6	0.020 (0.067)	0.110 (0.034)	.440 (.164)	0.258 (0.010)
	10000	30	3	0.005 (0.042)	0.027 (0.012)	.658 (.162)	0.143 (0.005)
			6	0.018 (0.070)	0.078 (0.035)	.479 (.211)	0.209 (0.013)
		60	3	0.000 (0.027)	0.015 (0.005)	.750 (.078)	0.117 (0.002)
			6	0.009 (0.033)	0.031 (0.010)	.652 (.105)	0.152 (0.004)
λ_θ	3000	30	3	0.005 (0.014)	0.042 (0.006)	.810 (.060)	0.040 (0.040)
			6	0.006 (0.015)	0.044 (0.008)	.811 (.054)	0.040 (0.042)
		60	3	0.011 (0.011)	0.038 (0.005)	.867 (.027)	0.035 (0.036)
			6	0.011 (0.013)	0.039 (0.006)	.864 (.027)	0.036 (0.035)
	10000	30	3	0.001 (0.007)	0.022 (0.003)	.935 (.019)	0.021 (0.021)
			6	0.002 (0.008)	0.023 (0.003)	.926 (.021)	0.022 (0.022)
		60	3	0.004 (0.006)	0.019 (0.002)	.954 (.009)	0.019 (0.019)
			6	0.004 (0.006)	0.020 (0.003)	.953 (.009)	0.019 (0.019)

Table 8b

Estimation Accuracy for Item Parameters across Magnitude of Main Effects Factor: Misconception Correlation of .50

τ	E	I	A	$Bias\ (\hat{\tau})$	$RMSE\ (\hat{\tau})$	$r(\hat{\tau},\tau)$	$SE(\hat{\tau})$
λ_0	3000	30	3	-0.011 (0.021)	0.014 (0.004)	.982 (.004)	0.105 (0.002)
			6	-0.018 (0.025)	0.020 (0.006)	.974 (.006)	0.117 (0.004)
		60	3	-0.012 (0.015)	0.011 (0.001)	.986 (.002)	0.096 (0.001)
			6	-0.011 (0.016)	0.012 (0.002)	.984 (.002)	0.100 (0.001)
	10000	30	3	-0.003 (0.011)	0.004 (0.001)	.995 (.001)	0.058 (0.001)
			6	-0.005 (0.012)	0.005 (0.001)	.993 (.002)	0.063 (0.001)
		60	3	-0.003 (0.008)	0.003 (0.000)	.996 (.001)	0.052 (0.000)
			6	-0.005 (0.008)	0.003 (0.000)	.995 (.001)	0.055 (0.000)
$\lambda_{1(a)}$	3000	30	3	0.008 (0.023)	0.030 (0.007)	.657 (.066)	0.148 (0.003)
			6	0.006 (0.026)	0.049 (0.012)	.547 (.074)	0.175 (0.004)
		60	3	0.006 (0.015)	0.021 (0.003)	.705 (.033)	0.132 (0.001)
			6	0.004 (0.015)	0.028 (0.004)	.663 (.036)	0.145 (0.002)
	10000	30	3	0.002 (0.013)	0.008 (0.002)	.852 (.029)	0.083 (0.001)
			6	-0.001 (0.015)	0.014 (0.003)	.778 (.041)	0.099 (0.002)
		60	3	0.001 (0.008)	0.006 (0.001)	.882 (.015)	0.073 (0.000)
			6	0.002 (0.008)	0.008 (0.001)	.856 (.016)	0.081 (0.001)
$\lambda_{2(a,b)}$	3000	30	3	0.001 (0.089)	0.098 (0.042)	.438 (.230)	0.245 (0.012)
			6	0.006 (0.122)	0.214 (0.092)	.337 (.270)	0.312 (0.022)
		60	3	0.007 (0.049)	0.058 (0.019)	.514 (.148)	0.206 (0.007)
			6	0.005 (0.062)	0.100 (0.027)	.426 (.165)	0.247 (0.010)
	10000	30	3	0.000 (0.047)	0.025 (0.011)	.658 (.169)	0.142 (0.005)
			6	0.015 (0.069)	0.068 (0.028)	.493 (.238)	0.196 (0.011)
		60	3	0.003 (0.027)	0.015 (0.005)	.770 (.077)	0.117 (0.002)
			6	0.002 (0.031)	0.026 (0.008)	.688 (.102)	0.143 (0.004)
λ_θ	3000	30	3	0.006 (0.013)	0.042 (0.007)	.817 (.056)	0.040 (0.039)
			6	0.005 (0.015)	0.043 (0.007)	.811 (.053)	0.042 (0.042)
		60	3	0.010 (0.011)	0.037 (0.006)	.868 (.025)	0.035 (0.035)
			6	0.011 (0.012)	0.038 (0.005)	.862 (.028)	0.035 (0.036)
	10000	30	3	0.002 (0.007)	0.022 (0.003)	.931 (.021)	0.021 (0.021)
			6	0.001 (0.007)	0.023 (0.004)	.926 (.024)	0.022 (0.021)
		60	3	0.003 (0.006)	0.019 (0.002)	.954 (.008)	0.019 (0.019)
			6	0.004 (0.006)	0.020 (0.002)	.953 (.008)	0.019 (0.019)

Table 9

Estimation Accuracy for Item Parameters by Magnitude of Main Effects Factor

τ	Misconception Main Effects ($\lambda_{1(\alpha)}$)	Ability Main Effects (λ_{θ})	$Bias\ (\hat{\tau})$	$RMSE\ (\hat{\tau})$	$r(\hat{\tau}, \tau)$	$SE(\hat{\tau})$
λ_0	Low	Low	-0.008 (0.014)	0.009 (0.002)	.988 (.002)	0.081 (0.002)
		High	-0.010 (0.017)	0.010 (0.002)	.986 (.003)	0.084 (0.002)
	High	Low	-0.005 (0.012)	0.007 (0.001)	.991 (.001)	0.074 (0.001)
		High	-0.008 (0.014)	0.008 (0.001)	.991 (.002)	0.077 (0.001)
$\lambda_{1(a)}$	Low	Low	0.000 (0.016)	0.025 (0.005)	.709 (.043)	0.125 (0.002)
		High	-0.001 (0.018)	0.026 (0.005)	.703 (.043)	0.128 (0.002)
	High	Low	0.006 (0.014)	0.015 (0.002)	.781 (.033)	0.107 (0.001)
		High	0.005 (0.015)	0.016 (0.003)	.776 (.036)	0.109 (0.001)
$\lambda_{2(a,b)}$	Low	Low	0.015 (0.076)	0.109 (0.040)	.437 (.202)	0.234 (0.013)
		High	0.011 (0.078)	0.117 (0.047)	.506 (.183)	0.238 (0.013)
	High	Low	0.003 (0.047)	0.040 (0.015)	.599 (.158)	0.169 (0.006)
		High	0.003 (0.050)	0.045 (0.018)	.603 (.146)	0.176 (0.007)
λ_{θ}	Low	Low	0.003 (0.008)	0.033 (0.004)	.859 (.037)	0.031 (0.032)
		High	0.009 (0.011)	0.027 (0.005)	.915 (.021)	0.026 (0.026)
	High	Low	0.001 (0.009)	0.036 (0.004)	.853 (.039)	0.035 (0.034)
		High	0.008 (0.011)	0.026 (0.005)	.933 (.016)	0.025 (0.025)

Table 10a

Estimation Accuracy for Structural Parameters across Magnitude of Main Effects Factor: Misconception Correlation of .50

τ	E	I	A	$Bias\ (\hat{\tau})$	$RMSE\ (\hat{\tau})$	$SE(\hat{\tau})$	$r(\hat{\tau},\tau)$
γ_0	3000	30	3	-0.014 (0.085)	0.069 (0.053)	0.326 (0.008)	
			6	-0.032 (0.094)	0.081 (0.058)	0.327 (0.008)	
		60	3	-0.004 (0.051)	0.040 (0.032)	0.321 (0.007)	
			6	-0.003 (0.058)	0.045 (0.036)	0.321 (0.007)	
	10000	30	3	-0.008 (0.043)	0.035 (0.027)	0.319 (0.007)	
			6	-0.010 (0.050)	0.039 (0.031)	0.319 (0.007)	
		60	3	-0.003 (0.034)	0.028 (0.020)	0.318 (0.007)	
			6	-0.003 (0.036)	0.029 (0.021)	0.318 (0.007)	
γ_1	3000	30	3	-0.003 (0.072)	0.102 (0.045)	0.101 (0.006)	
			6	-0.062 (0.064)	0.271 (0.113)	0.195 (0.021)	
		60	3	0.003 (0.048)	0.072 (0.031)	0.076 (0.004)	
			6	-0.013 (0.037)	0.134 (0.044)	0.128 (0.008)	
	10000	30	3	0.005 (0.037)	0.053 (0.023)	0.054 (0.004)	
			6	-0.016 (0.031)	0.125 (0.043)	0.097 (0.009)	
		60	3	0.000 (0.027)	0.039 (0.015)	0.041 (0.002)	
			6	-0.008 (0.019)	0.072 (0.026)	0.067 (0.004)	
γ_2	3000	30	3	0.008 (0.059)	0.115 (0.047)	0.115 (0.006)	
			6	0.027 (0.025)	0.302 (0.089)	0.225 (0.015)	
		60	3	-0.002 (0.042)	0.083 (0.037)	0.091 (0.004)	
			6	0.006 (0.014)	0.162 (0.042)	0.152 (0.007)	
	10000	30	3	-0.002 (0.030)	0.060 (0.025)	0.061 (0.004)	
			6	0.007 (0.012)	0.151 (0.041)	0.110 (0.008)	
		60	3	0.001 (0.025)	0.046 (0.019)	0.049 (0.002)	
			6	0.003 (0.007)	0.088 (0.025)	0.080 (0.004)	
$\gamma_{.}$	3000	30	3				.990 (.008)
			6				.974 (.011)
		60	3				.995 (.004)
			6				.993 (.003)
	10000	30	3				.997 (.002)
			6				.993 (.003)
		60	3				.999 (.001)
			6				.998 (.001)

Table 10b

Estimation Accuracy for Structural Parameters across Magnitude of Main Effects Factor: Misconception Correlation of .25

τ	E	I	A	$Bias\ (\hat{\tau})$	$RMSE\ (\hat{\tau})$	$SE(\hat{\tau})$	$r(\hat{\tau},\tau)$
γ_0	3000	30	3	-0.019 (0.085)	0.071 (0.053)	0.328 (0.008)	
			6	-0.073 (0.135)	0.128 (0.094)	0.341 (0.011)	
		60	3	-0.006 (0.065)	0.052 (0.040)	0.322 (0.007)	
			6	-0.015 (0.079)	0.060 (0.054)	0.323 (0.006)	
	10000	30	3	-0.003 (0.044)	0.037 (0.027)	0.319 (0.008)	
			6	-0.020 (0.062)	0.056 (0.038)	0.321 (0.008)	
		60	3	-0.008 (0.033)	0.027 (0.020)	0.317 (0.007)	
			6	-0.004 (0.037)	0.029 (0.024)	0.318 (0.007)	
γ_1	3000	30	3	0.005 (0.071)	0.101 (0.040)	0.099 (0.007)	
			6	-0.025 (0.075)	0.236 (0.079)	0.184 (0.018)	
		60	3	0.007 (0.052)	0.072 (0.030)	0.075 (0.003)	
			6	-0.004 (0.042)	0.119 (0.038)	0.117 (0.006)	
	10000	30	3	0.001 (0.038)	0.052 (0.021)	0.053 (0.004)	
			6	-0.005 (0.033)	0.109 (0.035)	0.090 (0.009)	
		60	3	0.004 (0.028)	0.038 (0.016)	0.040 (0.002)	
			6	-0.002 (0.024)	0.063 (0.021)	0.062 (0.004)	
γ_2	3000	30	3	0.002 (0.063)	0.110 (0.046)	0.111 (0.006)	
			6	0.015 (0.026)	0.209 (0.056)	0.179 (0.010)	
		60	3	-0.004 (0.046)	0.083 (0.034)	0.088 (0.003)	
			6	0.002 (0.016)	0.124 (0.032)	0.122 (0.004)	
	10000	30	3	0.000 (0.034)	0.059 (0.025)	0.060 (0.004)	
			6	0.003 (0.012)	0.103 (0.027)	0.091 (0.004)	
		60	3	-0.001 (0.025)	0.045 (0.018)	0.048 (0.002)	
			6	0.001 (0.009)	0.068 (0.018)	0.065 (0.002)	
$\gamma_{.}$	3000	30	3				.983 (.014)
			6				.972 (.010)
		60	3				.991 (.007)
			6				.992 (.003)
	10000	30	3				.995 (.004)
			6				.993 (.003)
		60	3				.998 (.002)
			6				.997 (.001)

Table 11

Estimation Accuracy for Structural Parameters by Magnitude of Main Effects Factor

τ	Misconception Main Effects ($\lambda_{1(\alpha)}$)	Ability Main Effects (λ_{θ})	$Bias\ (\hat{\tau})$	$RMSE\ (\hat{\tau})$	$SE(\hat{\tau})$	$r(\hat{\tau}, \tau)$
γ_0	Low	Low	-0.027 (0.077)	0.067 (0.049)	0.325 (0.008)	
		High	-0.031 (0.084)	0.073 (0.054)	0.326 (0.008)	
	High	Low	0.000 (0.043)	0.033 (0.027)	0.319 (0.007)	
		High	0.002 (0.043)	0.034 (0.027)	0.319 (0.007)	
$\gamma_{1(a)}$	Low	Low	-0.013 (0.058)	0.122 (0.045)	0.118 (0.011)	
		High	-0.007 (0.059)	0.124 (0.043)	0.124 (0.012)	
	High	Low	-0.005 (0.028)	0.051 (0.019)	0.063 (0.002)	
		High	-0.004 (0.029)	0.054 (0.020)	0.065 (0.003)	
$\gamma_{2(a,b)}$	Low	Low	0.007 (0.035)	0.127 (0.039)	0.128 (0.008)	
		High	0.005 (0.036)	0.126 (0.039)	0.131 (0.008)	
	High	Low	0.002 (0.020)	0.061 (0.018)	0.075 (0.003)	
		High	0.002 (0.020)	0.065 (0.020)	0.077 (0.003)	
$\gamma_{.}$	Low	Low				.986 (.008)
		High				.986 (.007)
	High	Low				.997 (.002)
		High				.996 (.002)

Table 12

Estimation Accuracy for Examinee Ability Parameters across Magnitude of Main Effects Factor

ρ_{tet}	E	I	A	$Bias\ (\hat{\theta})$	$RMSE\ (\hat{\theta})$	$r(\hat{\theta}, \theta)$
.25	3,000	30	3	-0.006 (0.019)	0.708 (0.010)	.697 (.010)
			6	-0.007 (0.019)	0.724 (0.010)	.678 (.011)
		60	3	-0.012 (0.019)	0.590 (0.010)	.802 (.007)
			6	-0.013 (0.021)	0.597 (0.010)	.796 (.008)
	10,000	30	3	-0.003 (0.010)	0.705 (0.006)	.699 (.006)
			6	-0.003 (0.011)	0.721 (0.006)	.680 (.006)
		60	3	-0.004 (0.010)	0.588 (0.005)	.803 (.004)
			6	-0.006 (0.011)	0.594 (0.005)	.799 (.004)
.50	3,000	30	3	-0.006 (0.020)	0.710 (0.011)	.694 (.010)
			6	-0.008 (0.017)	0.728 (0.010)	.675 (.011)
		60	3	-0.011 (0.019)	0.592 (0.010)	.801 (.008)
			6	-0.015 (0.018)	0.599 (0.009)	.795 (.007)
	10,000	30	3	-0.003 (0.010)	0.707 (0.006)	.697 (.006)
			6	-0.003 (0.010)	0.725 (0.005)	.679 (.006)
		60	3	-0.005 (0.010)	0.589 (0.005)	.803 (.004)
			6	-0.006 (0.010)	0.596 (0.005)	.798 (.003)

Table 13

Estimation Accuracy for Examinee Ability Parameters by Magnitude of Main Effects Factor

Misconception Main Effects ($\lambda_{1(a)}$)	Ability Main Effects (λ_{θ})	$Bias\ (\hat{\theta})$	$RMSE\ (\hat{\theta})$	$r(\hat{\theta}, \theta)$
Low	Low	-0.006 (0.015)	0.727 (0.008)	.682 (.008)
	High	-0.007 (0.015)	0.560 (0.007)	.827 (.005)
High	Low	-0.006 (0.014)	0.751 (0.008)	.653 (.009)
	High	-0.008 (0.015)	0.580 (0.008)	.811 (.005)

Table 14

Reliability for Examinee Ability (θ) and Classification of Individual Misconceptions (α_a) across Magnitude of Main Effects Factor

ρ_{tet}	E	I	A	θ	$\alpha_{.}$	α_1	α_2	α_3	α_4	α_5	α_6
.25	3,000	30	3	.545 (.015)	.907 (.007)	.894 (.008)	.907 (.008)	.919 (.006)			
			6	.525 (.016)	.806 (.016)	.793 (.019)	.795 (.016)	.812 (.015)	.799 (.016)	.829 (.013)	.809 (.014)
		60	3	.674 (.011)	.987 (.003)	.986 (.002)	.987 (.003)	.987 (.003)			
			6	.668 (.012)	.923 (.006)	.911 (.007)	.907 (.006)	.912 (.007)	.938 (.006)	.937 (.005)	.930 (.005)
	10,000	30	3	.544 (.008)	.906 (.004)	.895 (.004)	.906 (.004)	.918 (.004)			
			6	.525 (.008)	.807 (.009)	.793 (.009)	.793 (.009)	.810 (0.008)	.800 (.009)	.829 (.007)	.812 (.009)
		60	3	.676 (.006)	.987 (.002)	.986 (.001)	.987 (.002)	.987 (0.002)			
			6	.669 (.006)	.924 (.003)	.912 (.003)	.909 (.004)	.914 (0.003)	.939 (.003)	.938 (.003)	.931 (.003)
.50	3,000	30	3	.541 (.015)	.909 (.007)	.897 (.007)	.908 (.008)	.920 (0.007)			
			6	.523 (.016)	.853 (.015)	.843 (.015)	.839 (.015)	.857 (0.015)	.847 (.017)	.875 (.013)	.860 (.015)
		60	3	.674 (.012)	.988 (.003)	.986 (.002)	.989 (.003)	.988 (0.003)			
			6	.667 (.012)	.945 (.006)	.935 (.006)	.931 (.006)	.936 (0.006)	.956 (.004)	.957 (.005)	.951 (.005)
	10,000	30	3	.542 (.008)	.908 (.004)	.897 (.004)	.908 (.004)	.920 (0.004)			
			6	.523 (.009)	.853 (.007)	.842 (.008)	.841 (.008)	.857 (0.007)	.846 (.007)	.872 (.007)	.860 (.007)
		60	3	.675 (.006)	.988 (.002)	.987 (.001)	.988 (.002)	.989 (0.002)			
			6	.669 (.006)	.946 (.003)	.937 (.003)	.933 (.003)	.937 (0.004)	.957 (.003)	.957 (.003)	.952 (.003)

Note. The average of the reliability of the individual misconceptions for a row is denoted by $\alpha_{.}$.

Table 15

Average Reliability for Examinee Ability (θ) and Classification of Individual Misconceptions ($\alpha_.$) by Magnitude of Main Effects Factor

Misconception Main Effects ($\lambda_{1(\alpha)}$)	Ability Main Effects (λ_θ)	θ	$\alpha_.$
Low	Low	.522 (.013)	.855 (.010)
	High	.709 (.007)	.843 (.011)
High	Low	.491 (.013)	.965 (.001)
	High	.687 (.008)	.954 (.002)

Table 16

Correct Classification Rates for Marginal (α_a) and Whole Pattern ($\boldsymbol{\alpha}$) Classification across Magnitude of Main Effects Factor

ρ_{tet}	E	I	A	α_1	α_2	α_3	α_4	α_5	α_6	$\boldsymbol{\alpha}$
.25	3000	30	3	.903	.910	.916				.778
				(.006)	(.005)	(.005)				(.008)
			6	.843	.842	.851	.847	.858	.851	.481
				(.007)	(.008)	(.007)	(.008)	(.007)	(.007)	(.011)
		60	3	.958	.957	.958				.891
				(.003)	(.003)	(.003)				(.005)
			6	.912	.908	.913	.925	.927	.921	.677
				(.004)	(.005)	(.005)	(.004)	(.004)	(.004)	(.008)
	10000	30	3	.905	.911	.918				.781
				(.003)	(.003)	(.003)				(.004)
			6	.847	.847	.854	.850	.862	.855	.492
				(.004)	(.004)	(.004)	(.004)	(.004)	(.004)	(.005)
		60	3	.959	.959	.959				.894
				(.002)	(.002)	(.002)				(.003)
			6	.914	.910	.915	.927	.929	.923	.682
				(.003)	(.003)	(.002)	(.003)	(.002)	(.003)	(.004)
.50	3000	30	3	.904	.910	.917				.782
				(.005)	(.006)	(.005)				(.008)
			6	.864	.863	.868	.867	.877	.871	.573
				(.007)	(.006)	(.007)	(.006)	(.006)	(.006)	(.009)
		60	3	.958	.958	.958				.894
				(.003)	(.003)	(.003)				(.005)
			6	.925	.921	.925	.934	.937	.932	.731
				(.005)	(.005)	(.005)	(.004)	(.004)	(.005)	(.008)
	10000	30	3	.906	.912	.918				.786
				(.003)	(.003)	(.003)				(.004)
			6	.867	.866	.871	.870	.879	.874	.580
				(.003)	(.003)	(.004)	(.003)	(.003)	(.004)	(.005)
		60	3	.959	.959	.960				.896
				(.002)	(.002)	(.002)				(.003)
			6	.926	.922	.926	.936	.938	.934	.735
				(.002)	(.003)	(.003)	(.002)	(.002)	(.002)	(.004)

Table 17

Cohen's Kappa for Marginal (α_a) and Whole Pattern ($\boldsymbol{\alpha}$) Classification across Magnitude of Main Effects Factor

ρ_{tet}	E	I	A	α_1	α_2	α_3	α_4	α_5	α_6	$\boldsymbol{\alpha}$
.25	3,000	30	3	.808	.821	.833				.742
				(.013)	(.012)	(.010)				(.009)
			6	.686	.681	.700	.693	.718	.704	.449
				(.021)	(.027)	(.020)	(.027)	(.018)	(.020)	(.010)
		60	3	.917	.915	.916				.874
				(.006)	(.006)	(.006)				(.005)
			6	.827	.817	.827	.851	.856	.844	.660
				(.010)	(.011)	(.010)	(.008)	(.009)	(.022)	(.008)
	10,000	30	3	.812	.823	.837				.746
				(.006)	(.005)	(.005)				(.005)
			6	.698	.698	.711	.704	.729	.714	.460
				(.011)	(.009)	(.009)	(.010)	(.008)	(.008)	(.005)
		60	3	.919	.917	.918				.877
				(.003)	(.003)	(.004)				(.003)
			6	.831	.823	.832	.855	.859	.848	.665
				(.005)	(.006)	(.005)	(.005)	(.004)	(.005)	(.004)
.50	3,000	30	3	.809	.821	.835				.743
				(.011)	(.012)	(.010)				(.009)
			6	.730	.728	.740	.736	.759	.746	.505
				(.017)	(.014)	(.015)	(.014)	(.013)	(.014)	(.010)
		60	3	.917	.917	.917				.875
				(.007)	(.007)	(.006)				(.005)
			6	.851	.843	.852	.871	.876	.866	.695
				(.009)	(.010)	(.009)	(.009)	(.008)	(.009)	(.008)
	10,000	30	3	.814	.825	.837				.748
				(.006)	(.006)	(.005)				(.005)
			6	.738	.736	.748	.744	.763	.753	.513
				(.007)	(.007)	(.008)	(.007)	(.007)	(.007)	(.005)
		60	3	.919	.918	.920				.878
				(.003)	(.003)	(.003)				(.003)
			6	.854	.846	.854	.874	.878	.869	.700
				(.004)	(.005)	(.005)	(.005)	(.004)	(.004)	(.005)

Table 18

Correct Classification Rates and Cohen's Kappa by Magnitude of Main Effects

Misconception Main Effects ($\lambda_{1(a)}$)	Ability Main Effects (λ_{θ})	α_1	α_2	α_3	α_4	α_5	α_6	$\boldsymbol{\alpha}$
Correct Classification Rate								
Low	Low	.861	.864	.868	.847	.854	.838	.847
		(.005)	(.005)	(.005)	(.005)	(.005)	(.006)	(.005)
	High	.848	.848	.855	.826	.841	.838	.826
		(.005)	(.006)	(.005)	(.006)	(.005)	(.006)	(.006)
High	Low	.970	.970	.970	.959	.960	.958	.959
		(.002)	(.002)	(.002)	(.003)	(.003)	(.003)	(.003)
	High	.959	.957	.963	.946	.950	.947	.946
		(.003)	(.003)	(.003)	(.003)	(.003)	(.003)	(.003)
Cohen's Kappa								
Low	Low	.724	.729	.737	.696	.711	.678	.547
		(.013)	(.012)	(.012)	(.012)	(.012)	(.015)	(.008)
	High	.700	.699	.715	.656	.688	.682	.525
		(.012)	(.014)	(.012)	(.019)	(.012)	(.012)	(.008)
High	Low	.940	.940	.940	.919	.919	.916	.867
		(.004)	(.005)	(.004)	(.005)	(.005)	(.012)	(.004)
	High	.919	.914	.927	.892	.901	.896	.843
		(.005)	(.005)	(.005)	(.006)	(.006)	(.006)	(.005)

Table 19

Reliability for Classification of Individual Misconceptions by Magnitude of Main Effects Factor

Misconception Main Effects ($\lambda_{1(a)}$)	Ability Main Effects (λ_{θ})	α_1	α_2	α_3	α_4	α_5	α_6
Low	Low	.831 (.011)	.843 (.011)	.852 (.011)	.805 (.013)	.822 (.012)	.783 (.014)
	High	.806 (.012)	.804 (.012)	.824 (.011)	.755 (.015)	.792 (.012)	.786 (.013)
High	Low	.997 (.001)	.998 (.001)	.996 (.001)	.998 (.001)	.995 (.001)	.996 (.001)
	High	.990 (.002)	.987 (.002)	.993 (.001)	.983 (.003)	.988 (.003)	.987 (.003)

Table 20

Q-matrix for Literacy Assessment

Item/						Item/						Item/					
Alternative		α_1	α_2	α_3	θ	Alternative		α_1	α_2	α_3	θ	Alternative		α_1	α_2	α_3	θ
1	a	0	1	0	0	11	a	1	0	0	0	21	a	1	0	0	0
1	b	0	0	0	1	11	b	1	0	0	0	21	b	0	0	0	1
1	c	0	1	0	0	11	c	0	0	0	1	21	c	1	0	0	0
1	d	1	0	0	0	11	d	1	0	0	0	21	d	0	1	0	0
2	a	0	1	0	0	12	a	0	0	0	1	22	a	0	0	0	1
2	b	1	0	0	0	12	b	0	1	0	0	22	b	1	0	0	0
2	c	0	0	0	1	12	c	0	0	1	0	22	c	1	0	0	0
2	d	1	0	0	0	12	d	0	0	1	0	22	d	1	0	0	0
3	a	0	0	0	1	13	a	1	0	0	0	23	a	1	0	0	0
3	b	0	0	1	0	13	b	0	1	0	0	23	b	0	0	0	1
3	c	0	0	1	0	13	c	0	0	0	1	23	c	0	1	0	0
3	d	1	0	0	0	13	d	1	0	0	0	23	d	0	1	0	0
4	a	0	0	1	0	14	a	0	0	0	1	24	a	0	1	0	0
4	b	1	0	0	0	14	b	1	0	0	0	24	b	0	1	0	0
4	c	1	0	0	0	14	c	0	0	1	0	24	c	0	1	0	0
4	d	0	0	0	1	14	d	0	0	1	0	24	d	0	0	0	1
5	a	0	1	0	0	15	a	0	1	0	0	25	a	1	0	0	0
5	b	0	1	0	0	15	b	0	0	0	1	25	b	0	0	0	1
5	c	0	1	0	0	15	c	0	1	0	0	25	c	1	0	0	0
5	d	0	0	0	1	15	d	0	1	0	0	25	d	0	1	0	0
6	a	0	0	1	0	16	a	0	0	1	0	26	a	0	0	0	1
6	b	0	0	0	1	16	b	0	0	0	1	26	b	0	0	1	0
6	c	0	0	1	0	16	c	0	0	1	0	26	c	1	0	0	0
6	d	0	0	1	0	16	d	1	0	0	0	26	d	0	1	0	0
7	a	0	0	1	0	17	a	0	1	0	0	27	a	1	0	0	0
7	b	0	0	1	0	17	b	0	1	0	0	27	b	0	1	0	0
7	c	0	0	0	1	17	c	0	0	0	1	27	c	0	0	0	1
7	d	1	0	0	0	17	d	0	1	0	0	27	d	0	1	0	0
8	a	0	0	0	1	18	a	0	0	1	0	28	a	0	0	1	0
8	b	1	0	0	0	18	b	0	1	0	0	28	b	0	0	0	1
8	c	0	0	1	0	18	c	1	0	0	0	28	c	1	0	0	0
8	d	0	0	1	0	18	d	0	0	0	1	28	d	0	1	0	0
9	a	0	1	0	0	19	a	1	0	0	0						
9	b	1	0	0	0	19	b	0	1	0	0						
9	c	0	0	1	0	19	c	0	0	1	0						
9	d	0	0	0	1	19	d	0	0	0	1						
10	a	0	0	0	1	20	a	1	0	0	0						
10	b	0	1	0	0	20	b	0	1	0	0						
10	c	0	1	0	0	20	c	0	0	0	1						
10	d	0	1	0	0	20	d	0	0	1	0						

Table 21

Evaluation of Convergence Rates for Selecting a Prior Distribution for Main Effect for Ability

Model	Prior	Item Parameters		Main Effect for Ability		Structural Parameters		Examinee Ability	
$\hat{R}\ Criterion$		1.1	1.5	1.1	1.5	1.1	1.5	1.1	1.5
SICM*	$logN(0{,}0.5)$	.687	.952	.821	1.000	.286	.429	.924	.999
	$logN(0{,}1.5)$	.663	.916	.857	1.000	.143	.429	.913	.998
	$N(.6,.15)$	.675	.934	.679	.964	.143	.571	.922	.999
	$N(.6,.25)$	.753	.964	.929	1.000	.143	.143	.933	.998
NR IRT*	$logN(0{,}0.5)$	1.00	1.000	1.000	1.000			1.000	1.000
	$N(.6,.15)$	.976	1.000	1.000	1.000			1.000	1.000
2-PL IRT*	$logN(0{,}0.5)$	1.00	1.000	1.000	1.000			1.000	1.000
	$N(.6,.15)$	1.00	1.000	1.000	1.000			1.000	1.000

Table 22

Q-matrix and Estimated Item Parameters for Example Item from Literacy Assessment

Q-matrix					Parameter Estimates (Standard Errors)				
Alternative	α_1	α_2	α_3	θ	λ_0	$\lambda_{1(1)}$	$\lambda_{1(2)}$	$\lambda_{1(3)}$	λ_θ
A	0	0	0	1		0	0	0	0.228
B	0	0	1	0	-1.154	0	0	1.705	0
C	1	0	0	0	-0.092	0.681	0	0	0
D	0	1	0	0	0.666	0	0.049	0	0

Table 23

Misconception Pattern-Specific Log-Odds Equations for Modeling Nominal Response Probabilities with the SICM Model for Example Item*

$\boldsymbol{\alpha}_e$	$\log(\frac{P(X_{ei} = B\|\boldsymbol{\alpha}_e, \theta_e)}{P(X_{ei} = A\|\boldsymbol{\alpha}_e, \theta_e)})$	$\log(\frac{P(X_{ei} = C\|\boldsymbol{\alpha}_e, \theta_e)}{P(X_{ei} = A\|\boldsymbol{\alpha}_e, \theta_e)})$	$\log(\frac{P(X_{ei} = D\|\boldsymbol{\alpha}_e, \theta_e)}{P(X_{ei} = A\|\boldsymbol{\alpha}_e, \theta_e)})$
[000]	$\lambda_{B,0} - \exp(\lambda_{A,\theta}(\theta_e))$	$\lambda_{C,0} - \exp(\lambda_{A,\theta}(\theta_e))$	$\lambda_{D,0} - \exp(\lambda_{A,\theta}(\theta_e))$
[001]	$\lambda_{B,0} - \exp(\lambda_{A,\theta}(\theta_e)) + \lambda_{B,1(3)}$	$\lambda_{C,0} - \exp(\lambda_{A,\theta}(\theta_e))$	$\lambda_{D,0} - \exp(\lambda_{A,\theta}(\theta_e))$
[010]	$\lambda_{B,0} - \exp(\lambda_{A,\theta}(\theta_e))$	$\lambda_{C,0} - \exp(\lambda_{A,\theta}(\theta_e))$	$\lambda_{D,0} - \exp(\lambda_{A,\theta}(\theta_e)) + \lambda_{D,1(2)}$
[011]	$\lambda_{B,0} - \exp(\lambda_{A,\theta}(\theta_e)) + \lambda_{B,1(3)}$	$\lambda_{C,0} - \exp(\lambda_{A,\theta}(\theta_e))$	$\lambda_{D,0} - \exp(\lambda_{A,\theta}(\theta_e)) + \lambda_{D,1(2)}$
[100]	$\lambda_{B,0} - \exp(\lambda_{A,\theta}(\theta_e))$	$\lambda_{C,0} - \exp(\lambda_{A,\theta}(\theta_e)) + \lambda_{C,1(1)}$	$\lambda_{D,0} - \exp(\lambda_{A,\theta}(\theta_e))$
[101]	$\lambda_{B,0} - \exp(\lambda_{A,\theta}(\theta_e)) + \lambda_{B,1(3)}$	$\lambda_{C,0} - \exp(\lambda_{A,\theta}(\theta_e)) + \lambda_{C,1(1)}$	$\lambda_{D,0} - \exp(\lambda_{A,\theta}(\theta_e))$
[110]	$\lambda_{B,0} - \exp(\lambda_{A,\theta}(\theta_e))$	$\lambda_{C,0} - \exp(\lambda_{A,\theta}(\theta_e)) + \lambda_{C,1(1)}$	$\lambda_{D,0} - \exp(\lambda_{A,\theta}(\theta_e)) + \lambda_{D,1(2)}$
[111]	$\lambda_{B,0} - \exp(\lambda_{A,\theta}(\theta_e)) + \lambda_{B,1(3)}$	$\lambda_{C,0} - \exp(\lambda_{A,\theta}(\theta_e)) + \lambda_{C,1(1)}$	$\lambda_{D,0} - \exp(\lambda_{A,\theta}(\theta_e)) + \lambda_{D,1(2)}$

Note. A subscript of 0 indicates an intercept and 1 indicates a main effect for a misconception. Item subscripts have been suppressed as only one item is illustrated.

Table 24

Nominal Response Pattern for Two Students with Similar Estimated Abilities in Different Classes

Item	Key	Examinee 403	Examinee 199
1	2	2	2
2	3	3	3
3	1	1	1
4	4	4	4
5	4	4	4
6	2	2	2
7	3	3	3
8	1	1	1
9	4	4	4
10	1	1	1
11	3	3	3
12	1	1	1
13	3	3	3
14	1	1	1
15	2	2	2
16	2	2	2
17	3	3	3
18	4	4	4
19	4	4	4
20	3	3	3
21	2	2	2
22	1	1	1
23	2	1	2
24	4	4	3
25	2	1	2
26	1	1	4
27	3	2	1
28	2	3	4
	Number Correct	24	24
	$(\hat{\theta}_e)$	2.149	1.729
	$(\hat{\alpha}_e)$	[111]	[000]

Note: Bold face type indicates an item missed by at least one of the two examinees.

Table 25

Convergence Rates for Literacy Assessment

Model	Parameter	n_p	Mean $\hat{R}$	Median $\hat{R}$	$\hat{R}$ < 1.1	$\hat{R}$ < 1.5
SICM*	Item	166	1.202	1.048	.687	.952
	Main Effect Ability	28	1.054	1.023	.821	1.000
	Structural	7	3.735	1.756	.286	.429
	Examinee Ability	1097	1.034	1.017	.924	.999
SICM	Item	166	1.098	1.026	.783	.964
	Main Effect Ability	28	1.025	1.012	.964	1.000
	Structural	7	3.349	3.731	.143	.286
	Examinee Ability	1097	1.025	1.013	.963	1.000
NR LCDM	Item	168	1.100	1.023	.869	.976
	Structural	7	3.917	1.429	.286	.571
LCDM	Item	111	3.120	2.562	.297	.378
	Structural	7	4.487	3.854	.000	.143
NR IRT*	Item (Intercept)	84	1.002	1.001	1.000	1.000
	Main Effect Ability	84	1.001	1.001	1.000	1.000
	Examinee Ability	1097	1.002	1.001	1.000	1.000
NR IRT	Item (Intercept)	84	1.002	1.001	1.000	1.000
	Main Effect Ability	84	1.002	1.001	1.000	1.000
	Examinee Ability	1097	1.001	1.000	1.000	1.000
2-PL IRT*	Item (Intercept)	28	1.001	1.000	1.000	1.000
	Main Effect Ability	28	1.001	1.001	1.000	1.000
	Examinee Ability	1097	1.002	1.001	1.000	1.000
2-PL IRT	Item (Intercept)	28	1.001	1.000	1.000	1.000
	Main Effect Ability	28	1.001	1.001	1.000	1.000
	Examinee Ability	1097	1.001	1.001	1.000	1.000

Note. A variation of Gelman and Rubin's (1992) $\hat{R}$ statistic is used to assess convergence. For each model, the number of parameters estimated for each parameter type is denoted by n_p.

Table 26

Average Parameter Values (Standard Errors) from Literacy Assessment

	Item Parameters				Structural Parameters			Examinee Ability
	λ_0	$\lambda_{1(a)}$	$\lambda_{2(a,b)}$	λ_θ	γ_0	$\gamma_{1(a)}$	$\gamma_{2(a,b)}$	θ_e
SICM*	-0.935	1.470		0.507	-0.068	-11.308	11.331	-0.029
	(0.572)	(0.393)		(0.077)	(0.351)	(2.752)	(3.919)	(0.672)
SICM	0.308	1.340		0.501	0.253	-12.747	12.662	0.158
	(0.823)	(0.207)		(0.104)	(0.344)	(4.953)	(3.567)	(0.104)
NR LCDM	-2.059	1.577			-0.194	-7.174	-7.240	
	(0.614)	(0.310)			(0.330)	(2.209)	(2.672)	
LCDM	-2.404	3.948	-7.932		84.635	17.746	-45.956	
	(.883)	(1.306)	(2.141)		(29.759)	(3.829)	(12.830)	
NR IRT*	-0.935			0.805				0.030
	(0.107)			(0.137)				(0.484)
NR IRT	-1.256			0.872				0.001
	(0.125)			(0.141)				(0.442)
2-PL IRT*	0.097			0.709				0.001
	(0.075)			(0.090)				(0.486)
2-PL IRT	0.294			0.627				0.023
	(0.076)			(0.069)				(0.525)

Table 27

Pearson Correlations of Ability Estimates from Different Models

$r(\hat{\theta}_r, \hat{\theta}_c)$	SICM*	SICM	NR IRT*	NR IRT	2-PL IRT*	2-PL IRT
SICM*	1					
SICM	.915	1				
NR IRT*	.731	.732	1			
NR IRT	.730	.735	.988	1		
2-PL IRT*	.817	.818	.943	.941	1	
2-PL IRT	.814	.808	.939	.945	.994	1

Note. Respectively, $\hat{\theta}_r$ and $\hat{\theta}_c$ indicate the estimated ability for the model in the row and column.

Table 28

Percent Agreement of Individual (α_a) and Whole Pattern ($\boldsymbol{\alpha}$) Classification for Different Models

		SICM	SICM*	NR LCDM
α_1	SICM*	90.79		
	NR LCDM	84.50	88.06	
	LCDM	78.49	80.95	88.33
α_2	SICM*	90.79		
	NR LCDM	84.50	88.24	
	LCDM	73.29	76.85	85.69
α_3	SICM*	90.79		
	NR LCDM	84.32	88.06	
	LCDM	21.51	19.05	12.03
$\boldsymbol{\alpha}$	SICM*	90.79		
	NR LCDM	84.05	87.79	
	LCDM	0.00	0.00	0.46

Table 29

Relative Model Fit for Literacy Assessment Analysis

Model	N_p	LL	AIC	Relative Fit Rank	BIC	Relative Fit Rank
SICM	201	-33,681.30	67,762.51	6	68,762.58	5
SICM*	201	-33,623.20	67,646.39	4	68,646.45	3
NR LCDM	175	-36,552.00	73,507.91	7	74, 517.98	7
LCDM	118	-200,599.00	401,487.5	8	402,212.60	8
NR IRT	168	-33,539.30	67,526.67	3	68,646.74	4
NR IRT*	168	-33,619.30	67,686.65	5	68,806.73	6
2-PL IRT	56	-18,712.80	37,649.64	1	38,209.68	1
2-PL IRT*	56	-18,761.90	37,747.77	2	38,307.81	2

Note. N_p denotes the number of model parameters estimated.

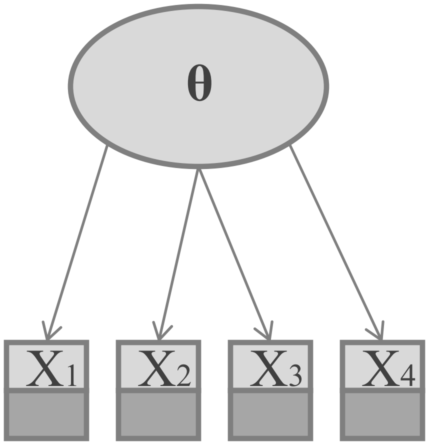


Figure 1. Unidimensional IRT model. The diagram shows a single continuous trait, indicated by the ellipse, being measured by four observed variables (i.e., items), indicated by rectangles. θ is the continuous latent ability and the shading that bisects the observed variables indicates the dichotomous nature of the scored response variable to item i (X_i).

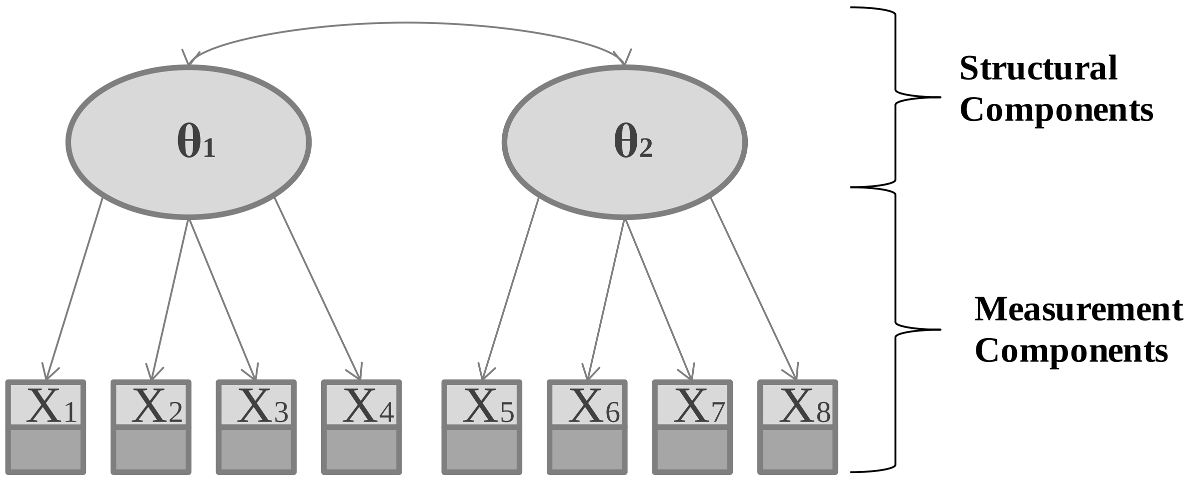


Figure 2. Multidimensional IRT model. This figure shows a MIRT model with two continuous latent variables (also known as traits or dimensions: θ_1, θ_2) being measured by eight dichotomous items, indicated by the bisected rectangles. The undirected path between traits indicates the correlation between the traits (i.e., the structural components of the model). The structural components of the model include trait variances and covariances. The measurement components describe how the observed variables are related to the traits.

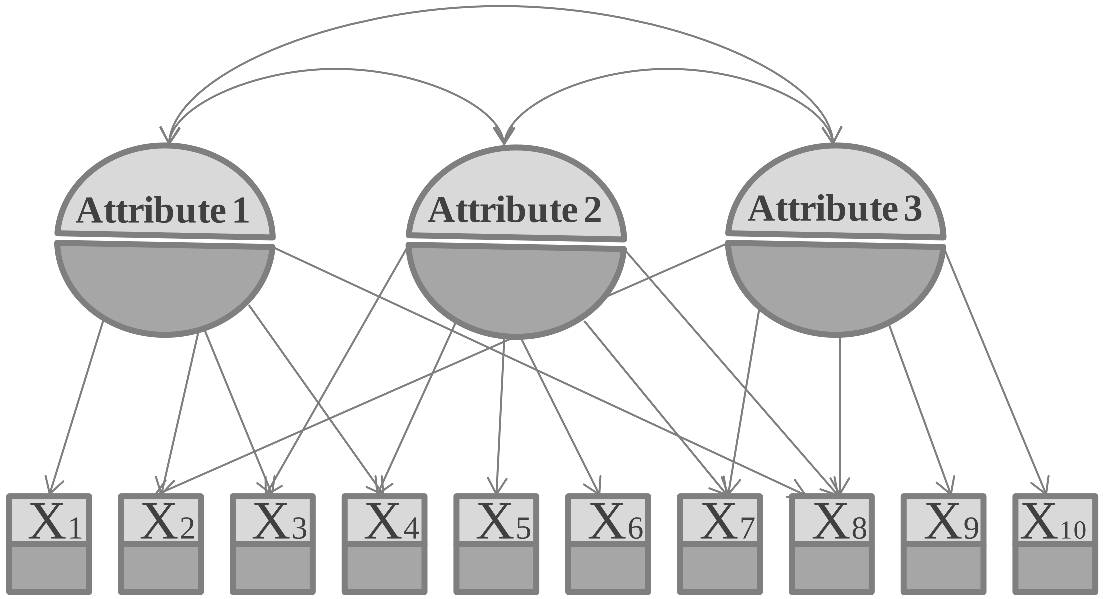
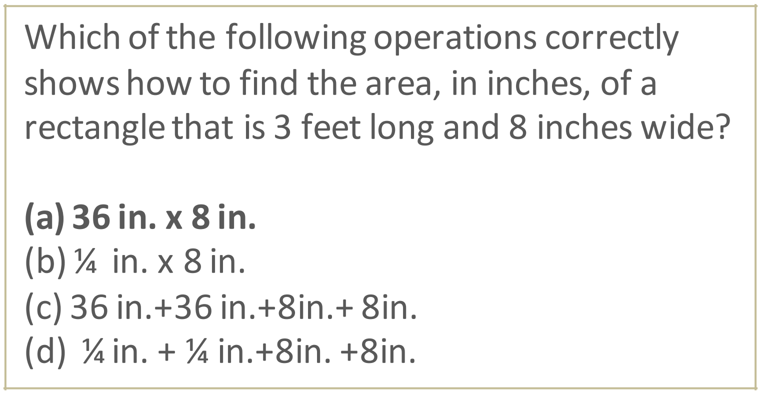


Figure 3. Diagnostic Classification Model. This figure shows a typical path diagram of an example DCM with three dichotomous attributes measured by ten dichotomous items. The diagram shows Attributes 1 and 3 are measured by five items and Attribute 2 by six items.



Which of the following operations correctly shows how to find the area, in inches, of a rectangle that is 3 feet long and 8 inches wide?

(a) 36 in. x 8 in.
(b) ¼ in. x 8 in.
(c) 36 in.+36 in.+8in.+ 8in.
(d) ¼ in. + ¼ in.+8in. +8in.

Figure 4. Example item. This figure contains a sample item about measurement in mathematics. The correct answer is Alternative A, indicated by bold face type.

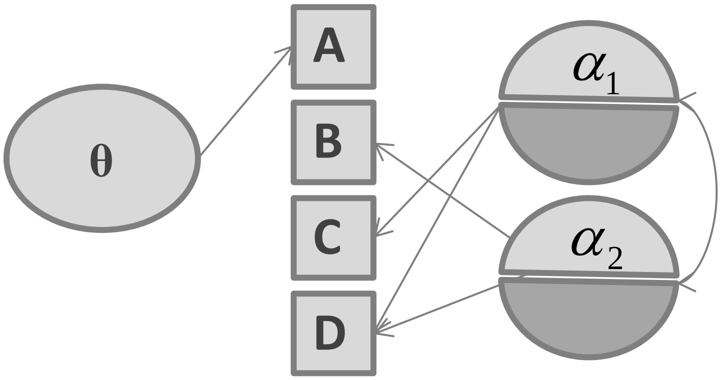


Figure 5. Scaling Individuals and Classifying Misconceptions model. This figure shows a path diagram for the example item in Figure 4 being modeled with the SICM model. The correct answer, Alternative A, is measuring a continuous trait, θ. The incorrect alternatives are measuring specific combinations of two dichotomous attributes. These attributes represent misconceptions or errors. For this item, α_1 was the inability to find the area of a rectangle, and $\alpha_2$was the inability to make conversions among units.

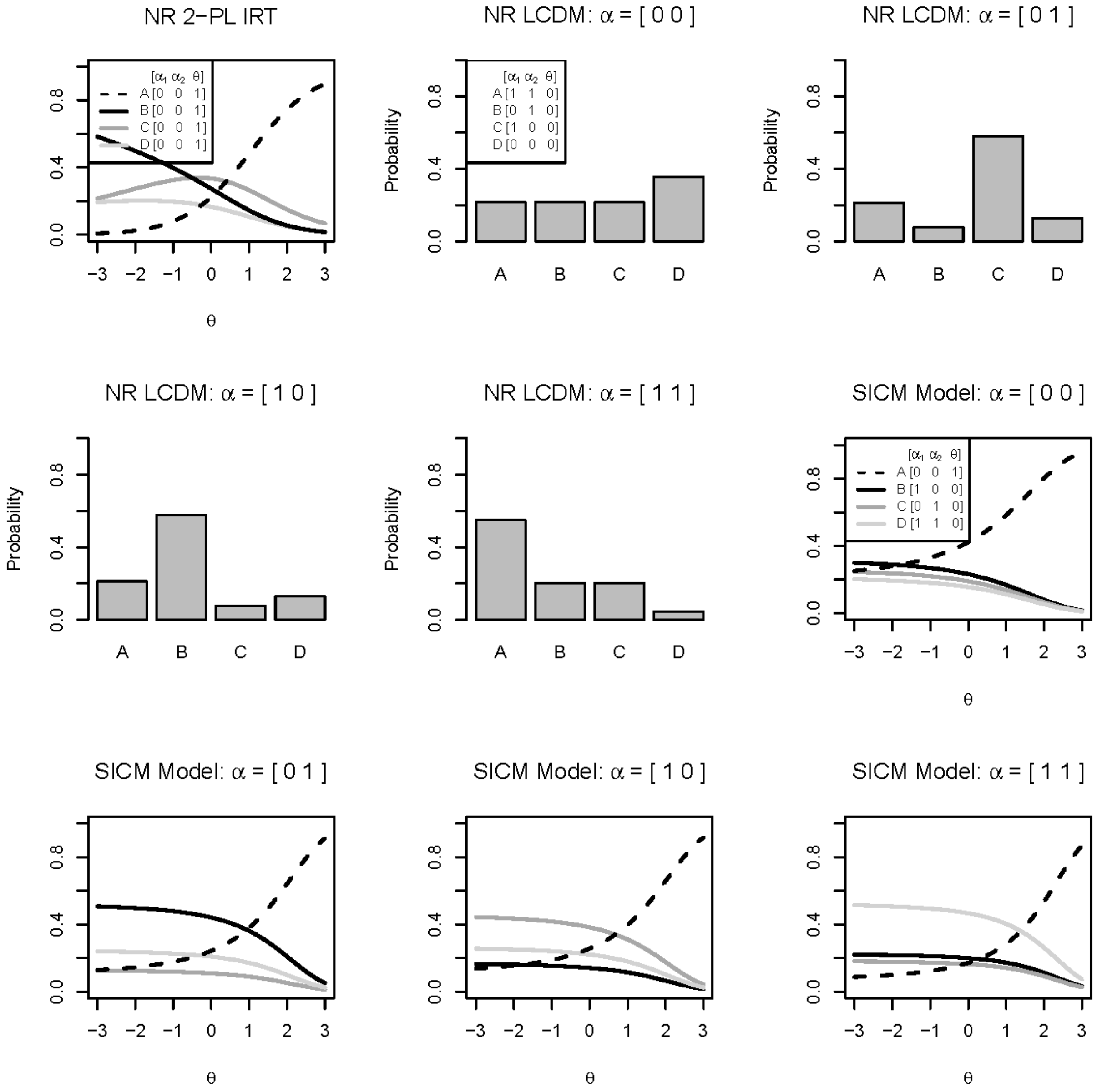


Figure 6. Contrasting models by item response probabilities. Item response probabilities for the the example item in Figure 4 are given to illustrate that the NR 2-PL IRT model provides the nominal response probability as a function of ability (θ), the NR LCDM as a function of attribute pattern (**α**), and the SICM model as a function of ability (θ) and misconception pattern (**α**).

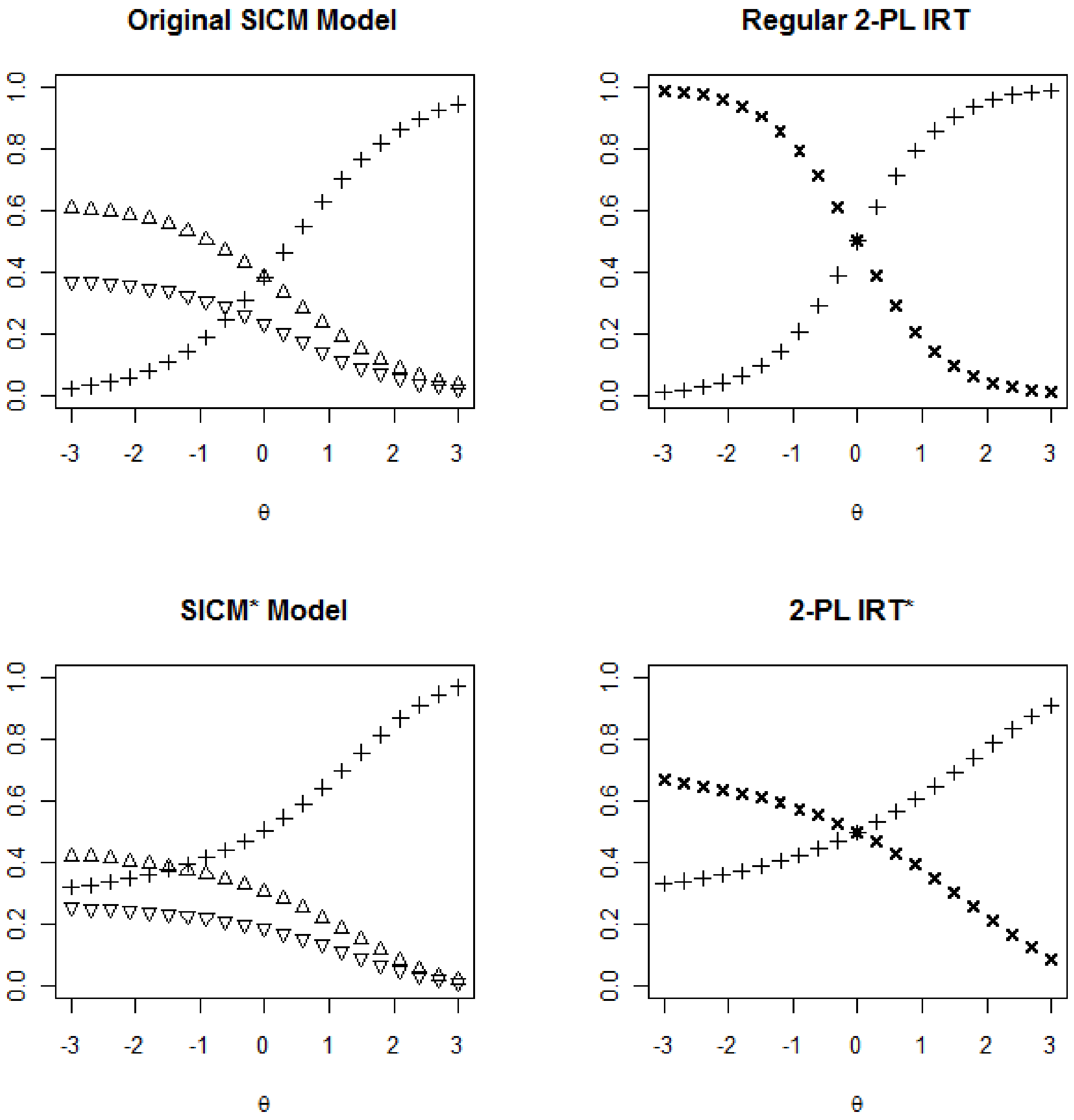


Figure 7. Comparison of the SICM model and the 2-PL IRT model with and without a lower asymptote. This figure shows the trace lines for examinees with no misconceptions for an item with three alternatives measuring the same misconception in each incorrect alternative. The trace line for the correct answer is denoted with + lines. In the SICM models, the trace line for the misconception with the higher and lower intercept are denoted with upward and downward facing triangles, respectively. The trace line for the incorrect answer in the 2-PL IRT models are denoted with x lines.

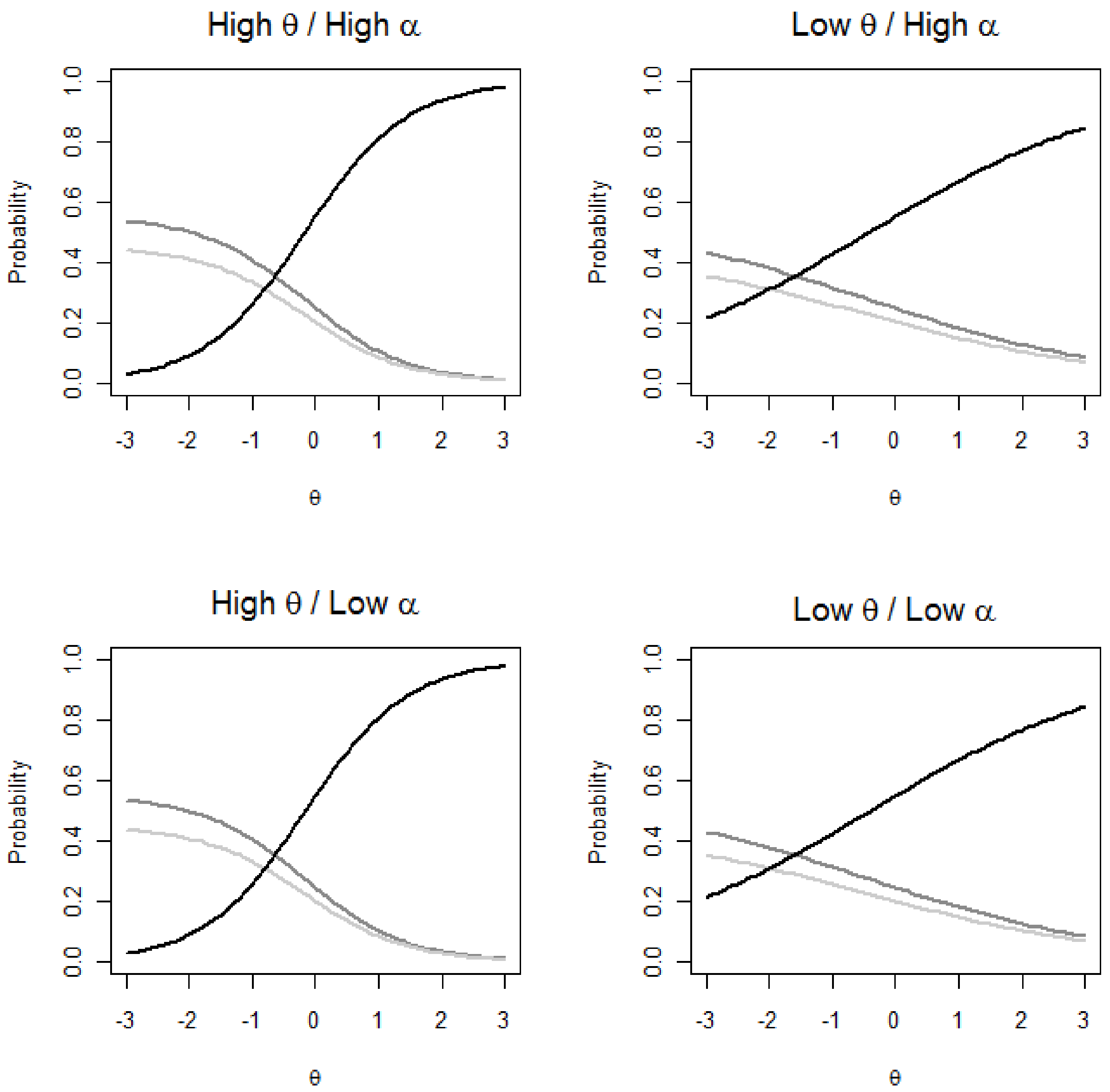


Figure 8. Simulation design: Main effects when misconception is absent. In this figure, response probabilities are given for an example item for each of the four variations of the magnitude of main effects factor. In this item, the correct answer is C (black curve), and the same misconception is measured by incorrect alternatives A (dark gray curve) and B (light gray curve). Response probabilities are for examinees that do not possess the misconception.

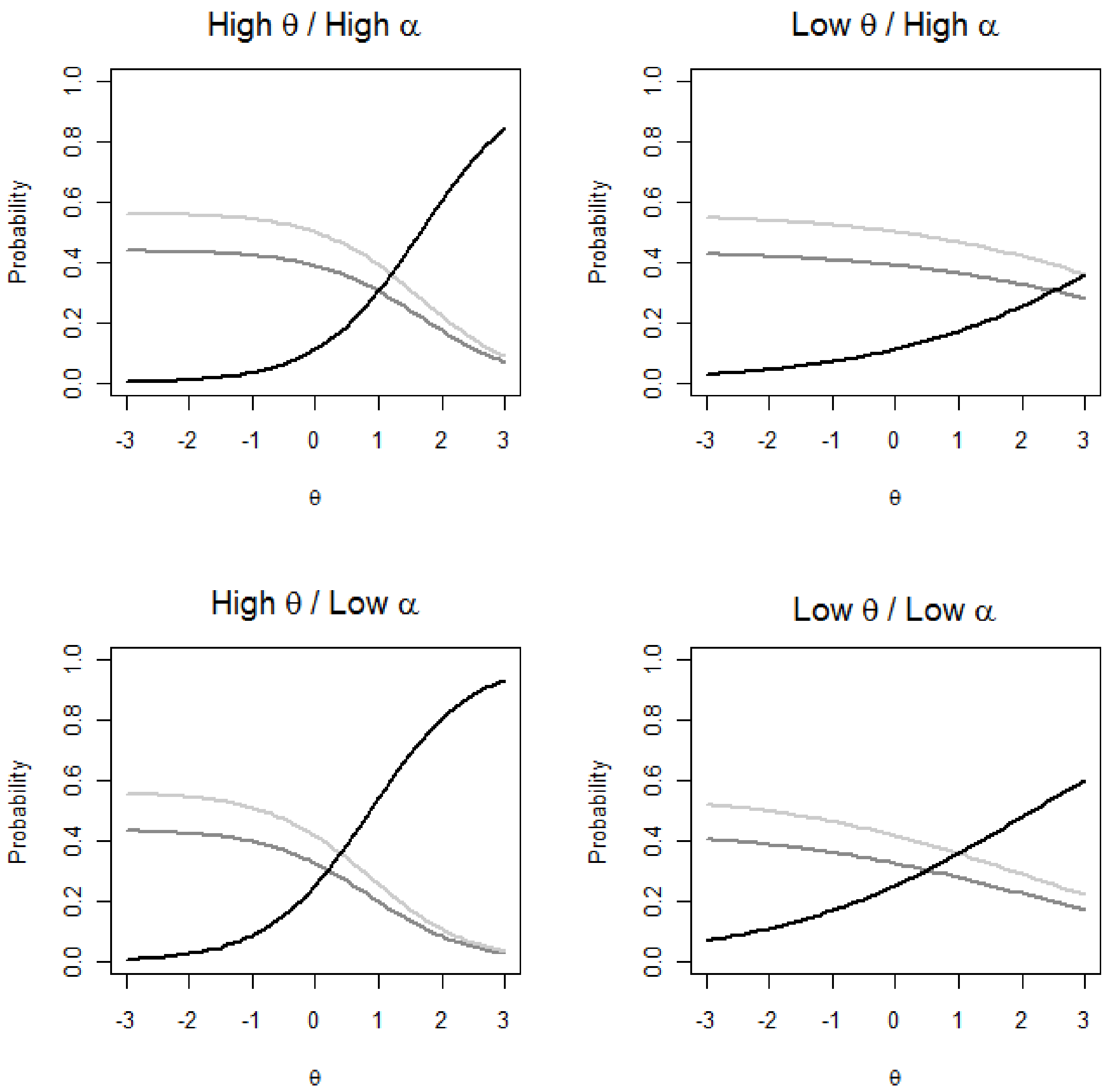


Figure 9. Simulation design: Main effects when misconception is present. This figure shows the trace lines for the same item alternatives under the four conditions as Figure 8, but for examinees who do possess the misconception measured by incorrect alternatives *A* and *B*.

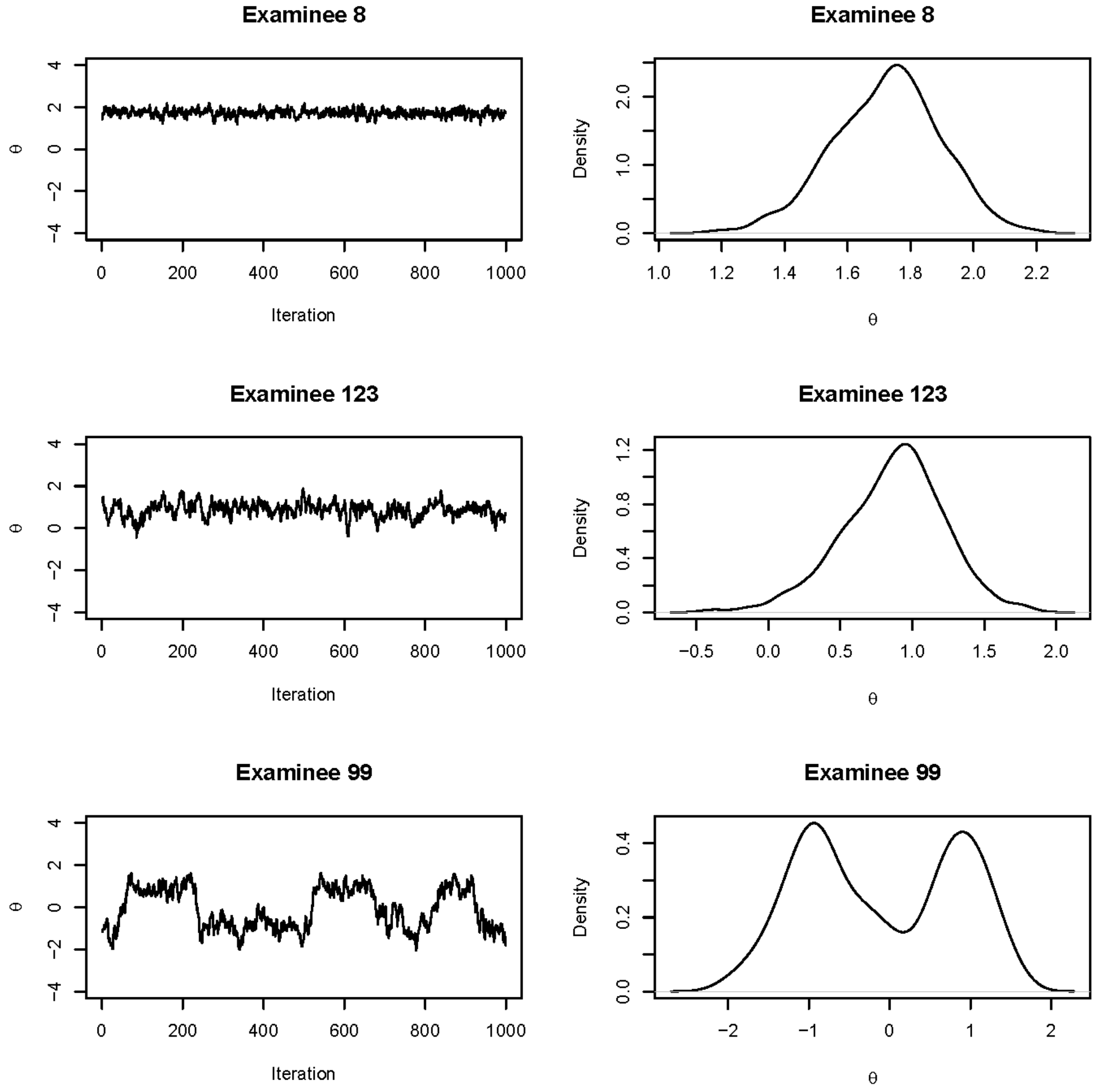


Figure 10. Trace and density plots. This figure shows the trace plots (on left) and the density plots (on right) of three examinees from the same condition.

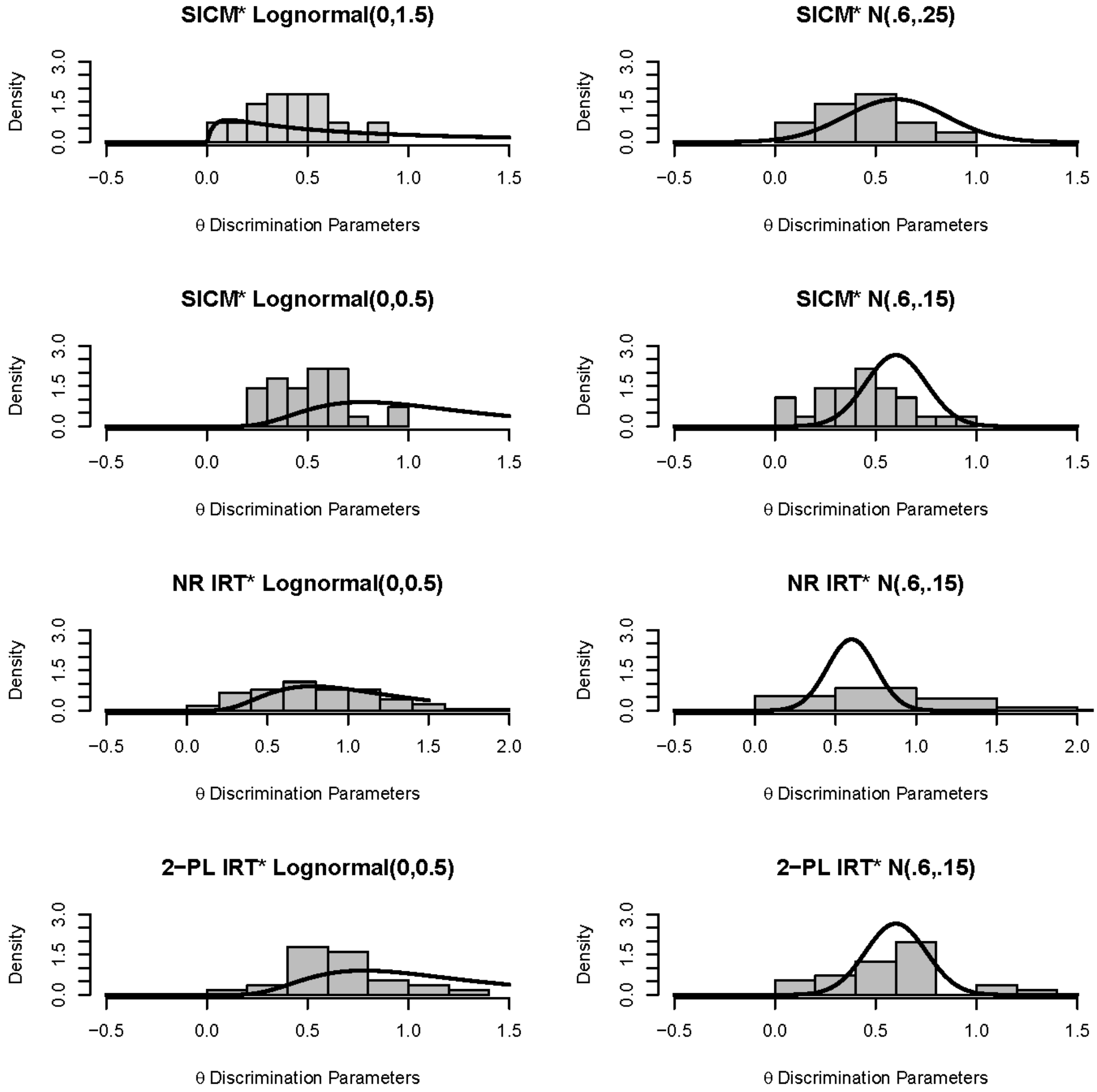


Figure 11. Evaluation of prior distributions for the main effect for ability. This figure shows a histogram of the estimated values of the discrimination parameters for ability with an overlaid density plot of the respective prior distribution used to estimate the parameters displayed in the histogram.

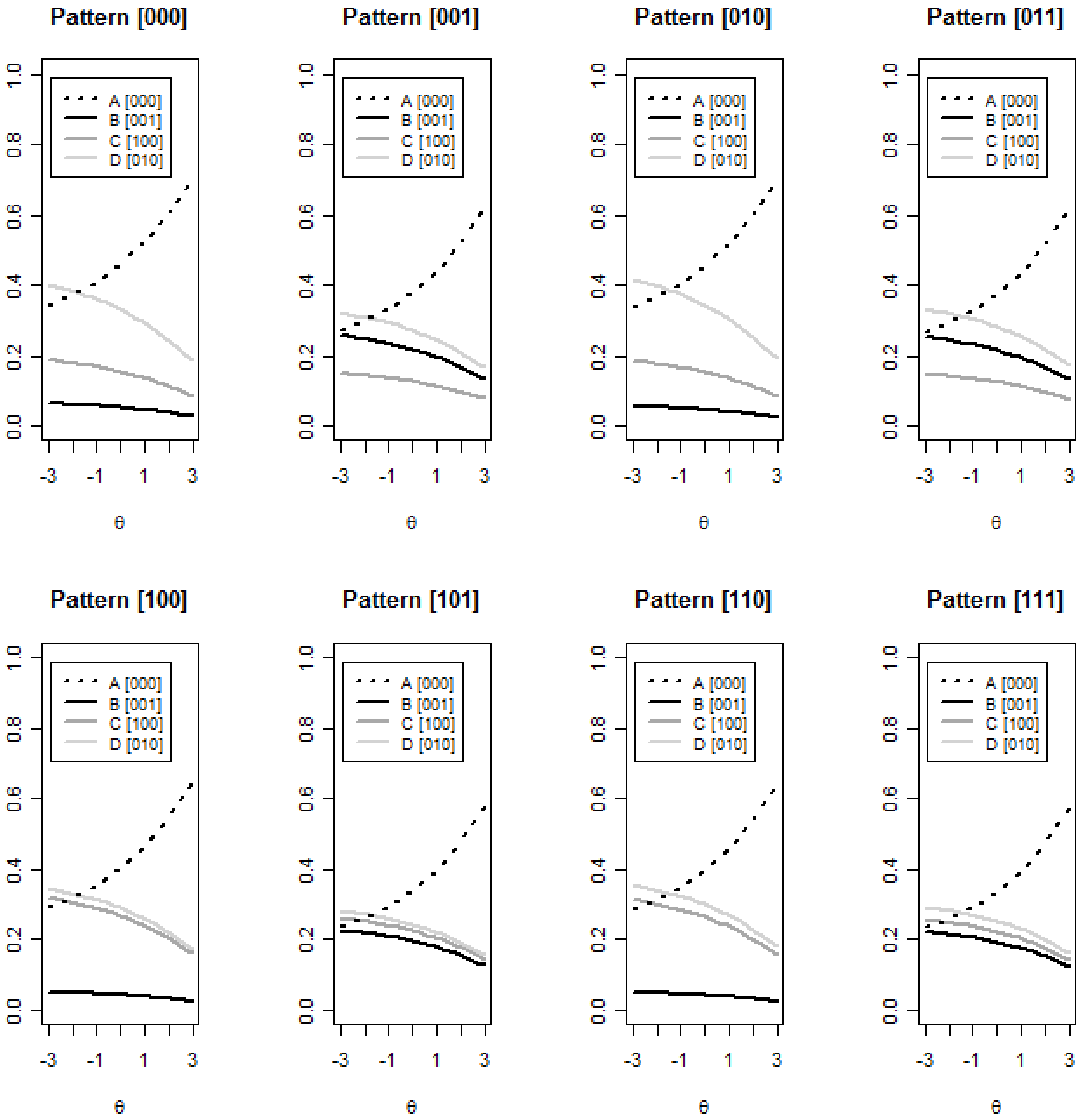


Figure 12. Trace lines for predicted nominal response probabilities for an item on the literacy assessment. This figure illustrates how the nominal response probabilities differ by the pattern of misconceptions that examinees have (e.g., the top left graph is for pattern [000] meaning these examinees possess no misconceptions). The legend shows that *A* is the correct answer because it measures zero misconceptions ([000]) and similarly *B*, *C*, and *D* are incorrect alternatives that each measure a different misconception.

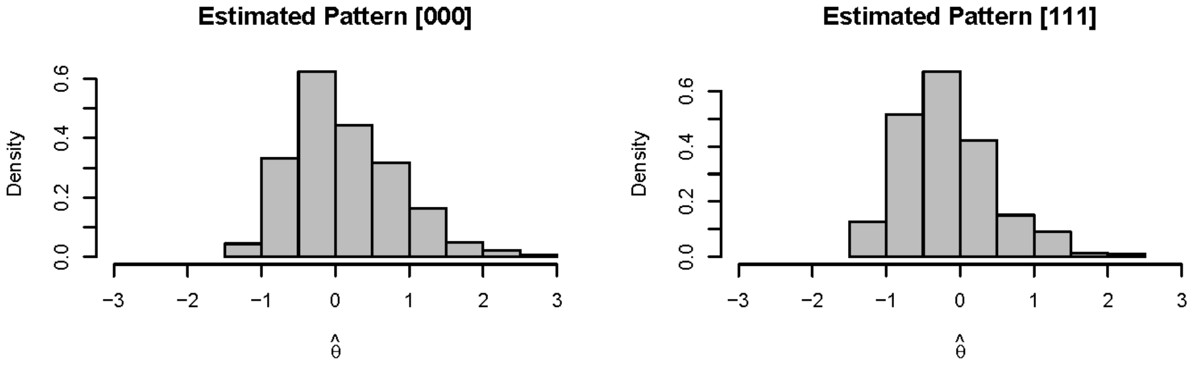


Figure 13. Estimated pattern by ability estimate for literacy assessment. This figure shows a histogram of the estimated values of the ability parameters for each of the two patterns that existed empirically, according to estimation using the SICM* model.

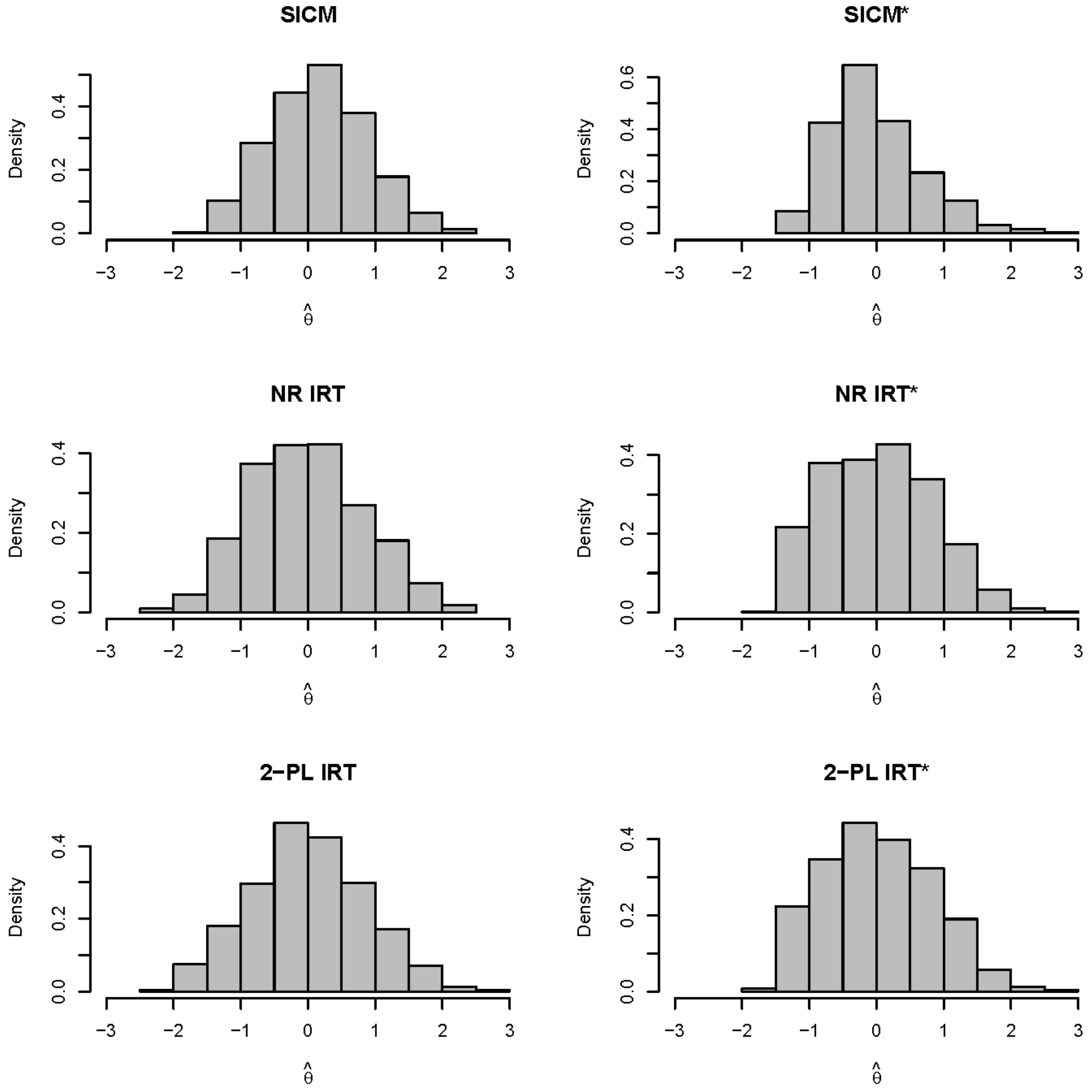


Figure 14. Comparison of ability distributions by estimation model. This figure shows a histogram of the estimated abilities for examinees for the six models that estimated an ability parameter for examinees.

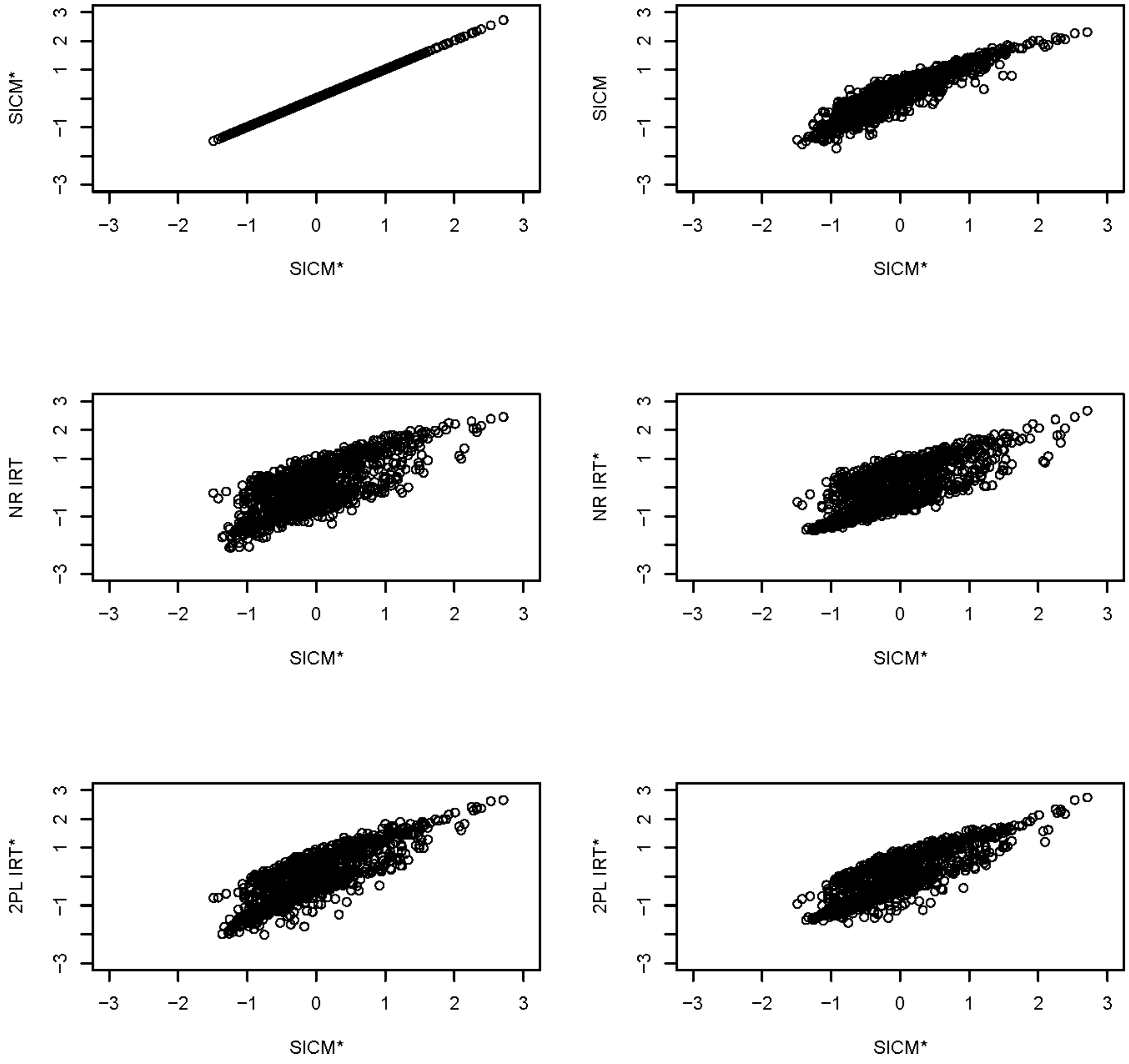


Figure 15. Comparison of ability estimates for SICM* model versus other estimation models. This figure compares the ability estimates of the SICM model with a pseudo-lower asymptote to the other 5 models that estimated an ability parameter for examinees.

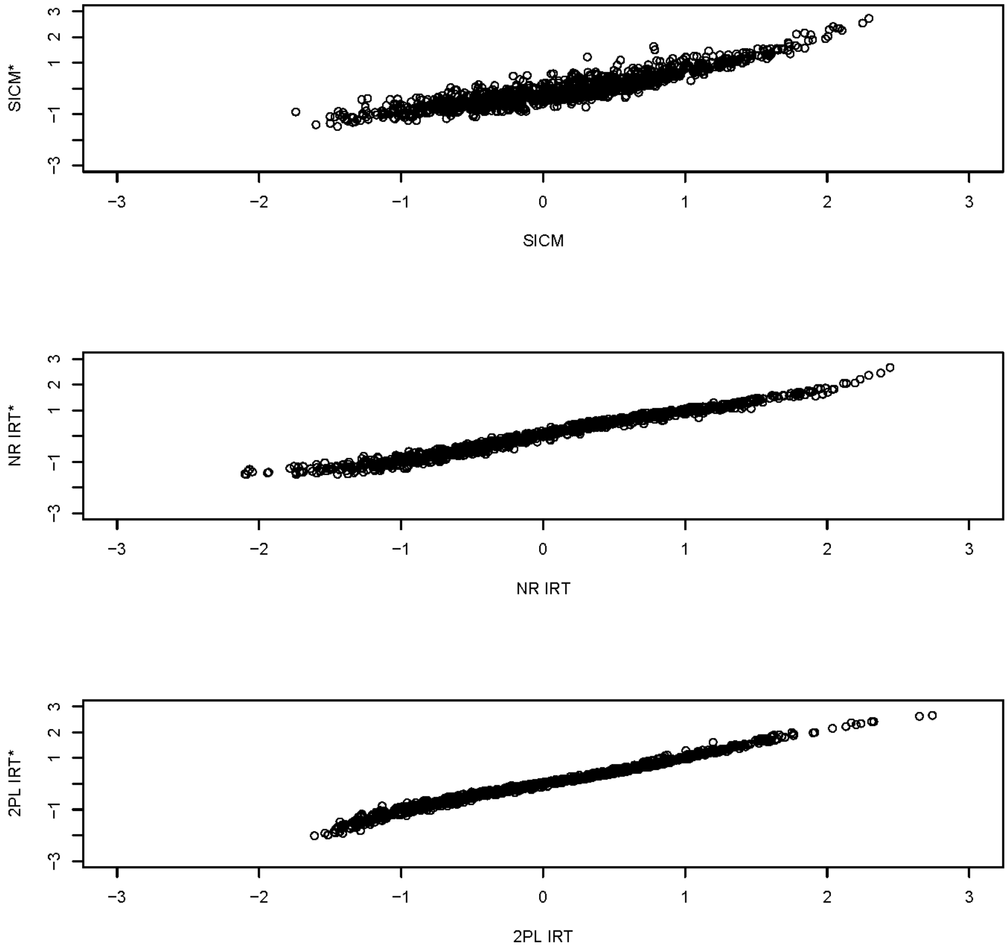


Figure 16. Comparison of ability estimates for models with pseudo-lower asymptotes. This figure compares the ability estimates of the SICM model, the NR IRT model, and the 2-PL IRT model with the versions of these models that have a lower asymptote (as indicated by *).

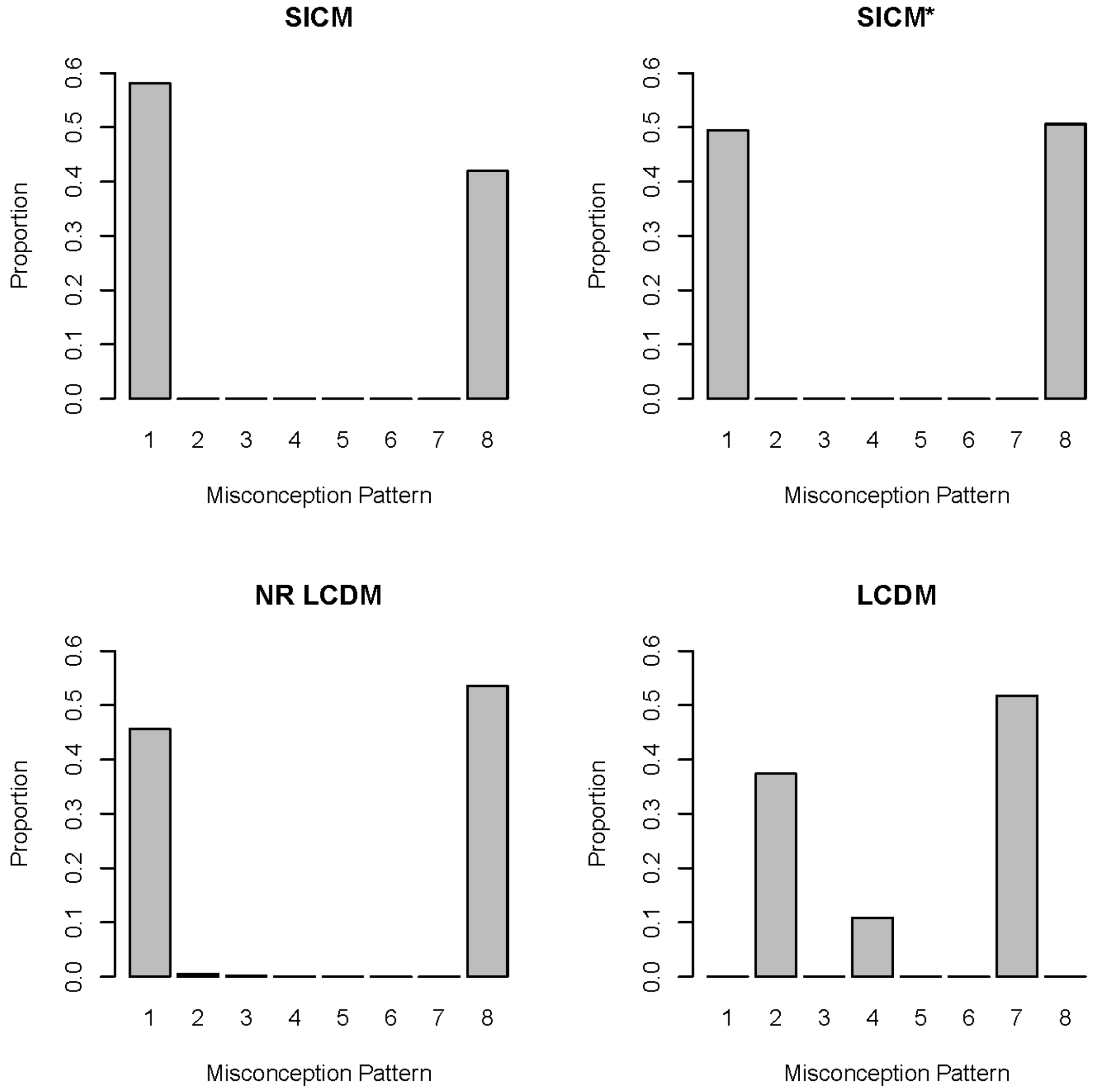


Figure 17. Comparison of examinee classifications. This figure compares the classifications of estimates with respect to the misconception they possess for the four models that measured misconceptions.