PSYCHOMETRIKA
2013
DOI: 10.1007/s11336-013-9350-4

COMBINING ITEM RESPONSE THEORY AND DIAGNOSTIC CLASSIFICATION MODELS: A PSYCHOMETRIC MODEL FOR SCALING ABILITY AND DIAGNOSING MISCONCEPTIONS

LAINE BRADSHAW AND JONATHAN TEMPLIN

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY, THE UNIVERSITY OF GEORGIA

Traditional testing procedures typically utilize unidimensional item response theory (IRT) models to provide a single, continuous estimate of a student's overall ability. Advances in psychometrics have focused on measuring multiple dimensions of ability to provide more detailed feedback for students, teachers, and other stakeholders. Diagnostic classification models (DCMs) provide multidimensional feedback by using categorical latent variables that represent distinct skills underlying a test that students may or may not have mastered. The Scaling Individuals and Classifying Misconceptions (SICM) model is presented as a combination of a unidimensional IRT model and a DCM where the categorical latent variables represent misconceptions instead of skills. In addition to an estimate of ability along a latent continuum, the SICM model provides multidimensional, diagnostic feedback in the form of statistical estimates of probabilities that students have certain misconceptions. Through an empirical data analysis, we show how this additional feedback can be used by stakeholders to tailor instruction for students' needs. We also provide results from a simulation study that demonstrate that the SICM MCMC estimation algorithm yields reasonably accurate estimates under large-scale testing conditions.

Key words: diagnostic classification models, item response theory, diagnosing student misconceptions, multidimensional measurement model, nominal response.

The Scaling Individuals and Classifying Misconceptions (SICM) model is a nominal response psychometric model designed by synthesizing the frameworks of item response theory (IRT) and diagnostic classification models (DCMs; Rupp, Templin, & Henson, 2010). This synthesis requires key changes to each modeling framework. Unlike existing applications of DCMs (e.g., de la Torre & Douglas, 2008; Lee, Park, & Taylan, 2011; Templin & Bradshaw, under review; Templin & Henson, 2006), the categorical latent variables, or attributes, in the SICM model are defined as *misconceptions* instead of skills. Misconceptions, or conceptions incongruous with expert or scientific understandings, are latent traits that systematically yield incorrect answers to test items. Unlike skill-based attributes in typical DCMs, possession of a misconception results in a *decreased* chance of a correct response. The SICM model alters the nominal response (NR) IRT model (Bock, 1972) by specifying categorical misconceptions, along with a continuous ability, as predictors of responses to multiple-choice items. By combining IRT and DCM features in this way, the SICM model more aptly characterizes existing cognitive theories about students' traits that influence item responses than either individual modeling framework: The model measures not only a composite ability indicated by correct answers, but also identifies distinct misconceptions manifested through incorrect answers.

Before presenting the SICM model, we first discuss the need for psychometric methodology for modeling misconceptions as latent variables. Next, we test the SICM Markov chain Monte Carlo (MCMC) estimation algorithm and evaluate properties of the new model through a simulation study. We then demonstrate the SICM model in practice through an empirical data analysis before concluding with a discussion of our findings.

Requests for reprints should be sent to Laine Bradshaw, Department of Educational Psychology, The University of Georgia, 323 Aderhold Hall, Athens, GA 30602, USA. E-mail: laineb@uga.edu

Electronic Supplementary Material The online version of this article (doi:10.1007/s11336-013-9350-4) contains supplementary material, which is available to authorized users.

Published online: 02 August 2013

1. Assessing Misconceptions

Learning theorists and educational researchers have developed strong research bases regarding misconceptions learners develop as they make sense of novel concepts (e.g., Confrey, 1990; Jendraszek, 2008; Smith, diSessa, & Roschelle, 1993). For example, consider the following statements: (a) Two multiplied by ¼ is greater than two divided by ¼; (b) If three random and independent draws are made from letters of the alphabet, LPB is a more likely result than ZZZ; (c) If a brick and an apple are dropped from the same height, the brick will reach the ground first. Each statement is a manifestation of a common student misconception: (a) Multiplication always yields a larger number, and division always yields a smaller number (Bell, Swan, & Taylor, 1981); (b) The result of one independent draw impacts the probability of subsequent draws (e.g., Borovcnik, Bentz, & Kapadia, 1991); (c) Heavier objects fall faster (e.g., Hestenes, Wells, & Swackhamer, 1992).

The presence of a misconception impedes learning; thus, the awareness and identification of misconceptions are essential for teaching (e.g., National Council of Teachers of Mathematics, 2001). Seeking to identify students' misconceptions, researchers have designed assessments known as *concept inventories* (Hestenes et al., 1992). Concept inventories, or *distractor-driven assessments* (Sadler, 1998), utilize multiple-choice items with incorrect options that are common erroneous answers reflective of known student misconceptions. The item writers' choices for incorrect options are often informed by extensive qualitative research. Repeated interactions with students reveal not only the common misconceptions students develop, but also common incorrect answers students give to open-ended questions when they have certain misconceptions. Student responses, observed through classroom settings or formal interviews, provide bases for incorrect options on concept inventories.

Hestenes et al. (1992) refined this test development practice through their innovation of the Force Concepts Inventory (FCI). The FCI has been credited with reforming physics education by connecting teaching practice to student understandings, and it is one of the most widely administered assessments in science education (Evans et al., 2003). Since the development of the FCI, many concept inventories have been developed in science, technology, engineering, and mathematics (STEM) related domains, including astronomy (e.g., Astronomy and Space Science Concepts Inventory; Sadler et al., 2010), engineering (e.g., Electromagnetic Concepts Inventory; Evans et al., 2003), chemistry, (e.g., Chemical Concepts Inventory; Mulford & Robinson, 2002), and statistics (e.g., Statistical Reasoning Assessment; Garfield & Chance, 2000; Probability Reasoning Questionnaire; Khazanov, 2009).

Concept inventories have been administered in classrooms to conduct research using pre-post comparisons to evaluate the effectiveness of teaching practices that help students overcome commonsense conceptions that are at odds with science (e.g., Hake, 1998). However, present psychometric methodologies do not incorporate the effects of misconceptions in the parameterization of the item response probability, nor do they provide statistical estimates of misconceptions as latent variables to measure and study these student traits. Traditional psychometric methods (i.e., Classical Test Theory and IRT) used to analyze concept inventory data focus on measuring a single continuous ability. To provide information about misconceptions, one of two strategies has been used: (1) The misconceptions are assessed or "diagnosed" by a subscore for each misconception that was calculated by tallying the number of times an option that measures a given misconception was selected by a student (e.g., Khazanov, 2009), or (2) IRT option characteristic curves (OCCs) corresponding to misconceptions are examined on an individual item and student basis (e.g., Sadler, 1998).

Limitations exist for each of these strategies. When a misconception is measured by a small number of items, tallies of misconceptions may be unreliable, as subscores have been shown to be (Haberman, Sinharay, & Puhan, 2009; Sinharay, Haberman, & Punhan, 2007). Researchers caution using subscores to make instructional decisions, as decisions based on unreliable subscores

counterproductively may misguide instructional strategies and resources (Tate, 2004). Further, the use of misconception subscores does not directly answer the question of whether or not a student possesses the misconception. A subsequent decision must be made to assign a cut-off score—analogous to setting a standard (e.g., Cizek, Bunch, & Koons, 2004)—to classify a student as having the misconception or not.

Although IRT has advantages over CTT in terms of quantifying variable error in overall ability estimates and providing information about how well the items function to measure the overall ability, IRT similarly lacks usefulness with regards to measuring misconceptions or evaluating the test's or an item's effectiveness for measuring misconceptions. Post-hoc analysis of OCCs for each student and item is tedious, and similar cautions are warranted for making instructional decisions based on individual item responses. Most importantly, examining OCCs does not use information aggregated across items to make a decision about, or to identify, which misconceptions students have. Thus, both CTT and IRT methods can provide descriptive information regarding misconceptions; however, they cannot directly diagnose whether a student has each misconception nor provide information about how well an item measures a misconception. Diagnosing misconceptions is critical for informing teaching practices, and item information is critical for constructing tests that yield valid diagnoses of misconceptions.

Another strategy that has been used to identify "bugs" in students' reasoning uses a statistical pattern classification approach. For a test of fraction subtraction skills, Tatsuoka (1985) used Rule Space methodology to determine if student component score response pattern locations in a Cartesian coordinate system coincide with locations of response patterns expected if students had certain misconceptions. Like subscores and IRT ability estimates, this methodology does not utilize misconceptions to predict the item response, provide estimates of misconceptions as latent student traits, nor yield item information regarding misconceptions.

Although concept inventory design follows current recommendations for test construction or *assessment engineering* (Luecht, 2013) where the construct delineation, task development, and psychometric model selection occurs sequentially to avoid construct morphing due to model-fit prioritizing, the final step of appropriate psychometric model selection is not possible when such methodology does not yet exist. In lieu of this gap in psychometric theory, we developed the Scaling Individuals and Classifying Misconceptions or SICM model. Our goal was to create a psychometric framework to quantitatively study misconceptions, specifically designing the SICM model to meet the practical need of diagnosing student misconceptions and the theoretical need to statistically falsify cognitive theories regarding misconceptions. The empirical data analysis presented later in this paper utilizes data from the FCI to demonstrate how the SICM model can capitalize on the careful design of this test to produce reliable estimates of misconceptions and also yield item parameters that explain how the misconception presence systematically influences OCCs.

2. Psychometric Foundations for SICM Model

The SICM model is a generalized linear latent and mixed model (Skrondal & Rabe-Hesketh, 2004) with both continuously- and categorically-distributed random effects. The statistical foundations of the model lie in diagnostic classification models (also known as cognitive diagnosis models; e.g., Leighton & Gierl, 2007), IRT, restricted latent class analysis, and multinomial logistic regression. Models foundational to the SICM model will be overviewed before specifying the SICM model, beginning with a newer class of multidimensional measurement models referred to as diagnostic classification models (DCMs).

2.1. Diagnostic Classification Models

DCMs have been the focus of much recent psychometric research. One recent advance includes unifying the parameterization of various DCMs using a single general log-linear framework. Henson, Templin, and Willse (2009) extended von Davier's (2005) general diagnostic model to include latent variable interaction effects to create the Log-linear Cognitive Diagnosis Model (LCDM). With interaction terms in the LCDM, many commonly used—and seemingly different—DCMs can be understood as constrained versions of the LCDM. Another key advance was being able to estimate complex DCMs with a commercially available software package (i.e., Mplus; Muthén & Muthén, 1998–2012; see Templin & Hoffman, 2013, and Rupp et al., 2010). Both of these recent advances greatly improved the accessibility of DCMs to the psychological and educational measurement and research communities.

DCMs seem well-suited for addressing the growing demands for gaining diagnostic feedback from educational assessments (Huff & Goodman, 2007; National Research Council, 2010; No Child Left Behind Act, 2001). These models use categorical latent attributes to represent multiple skills or abilities measured by a test. Attributes, denoted by α_a where $a = 1, 2, \ldots, A$, are assumed to be dichotomous latent variables: attribute a is either mastered/present ($\alpha_a = 1$) or not mastered/absent ($\alpha_a = 0$). A student's attribute pattern is an A-length vector of binary indicators of mastery. Attributes a student has mastered can be interpreted as abilities the student does not need to improve upon. Conversely, future instruction may be needed with respect to attributes that a student has not mastered. In comparison to an overall ability estimate, the attribute mastery pattern can provide feedback with respect to more fine-grain abilities, which is helpful for tailoring instruction to students' specific needs.

Multidimensional feedback from tests has been sought after for some time, but measuring multiple continuous traits remains difficult to do under practical testing conditions because a sizable number of items are needed for measuring each dimension. Instead of finely locating each examinee in a continuous multidimensional space as IRT models do, DCMs coarsely classify examinees with respect to each trait (i.e., as masters or nonmasters of the trait). Data demands (i.e., test length and sample size needed for parameter calibration) are substantially relaxed for classification according to categorical attributes in comparison to scaling continuous traits (see Bolt & Lall, 2003; Bradshaw & Cohen, 2010; Kunina-Habenicht, Rupp, & Wilhelm, 2012). Thus, trading scaling for classification enables DCMs to provide reliable multidimensional feedback in practical testing settings (Templin & Bradshaw, 2013).

These features make DCMs viable for reporting multidimensional diagnostic information in educational settings where time for testing is limited but multidimensional feedback is needed. However, given the current reliance of testing on measuring an overall ability, DCMs may not fulfill all needs of policy-driven assessment systems centered on scaling examinee ability. The SICM model seeks to bridge the IRT and DCM paradigms by using DCMs to provide diagnostic feedback while still providing an overall measure of ability. In contrast to the DCMs described above, the SICM model's diagnostic feedback is given with respect to attributes defined as misconceptions instead of skills or abilities.

2.2. Multinomial Models

The SICM model is able to measure both a continuous latent ability θ_e and a set of binary misconceptions $\boldsymbol{\alpha}_e$ by utilizing nominal response data from multiple-choice tests. Often in psychometric modeling of educational tests, responses are dichotomized into two categories—correct or incorrect, collapsing all of the incorrect options into one category (a wrong answer) and failing to preserve the uniqueness of each option. Such dichotomization can be viewed as an incomplete theory for modeling the item response if characteristics of the incorrect options

present variations in the item responses (Thissen & Steinberg, 1984; van der Linden & Hambleton, 1997). The item response function of the SICM model incorporates misconceptions as student characteristics to provide a more complete model of the item response.

Various types of models have been developed to predict nominal item responses. In IRT, these models include the nominal response IRT model (NR IRT model; Bock, 1972) and the multiple-choice model (MC model; Thissen & Steinberg, 1984). The MC model extends the NR IRT model to account for guessing by predicting a "do not know" category in addition to the possible categories of response provided by item options. In diagnostic modeling, the nominal response DCM (Templin & Bradshaw, under review) extends the LCDM to a nominal response model, maintaining the general log-linear parameterization. When sample sizes are large, the NR DCM capitalizes on information in the incorrect options, demonstrated by greater classification accuracy when compared to the LCDM for dichotomous responses (Templin & Bradshaw, under review). The multiple-choice deterministic-inputs, noisy and gate model (MC DINA; de la Torre, 2009) is a more constrained nominal response DCM that requires strict assumptions about options and attribute interactions. To date, no empirical data analysis has been reported with the MC DINA. The SICM model can be viewed as a blend of the NR DCM and the NR IRT models: categorical attributes and a continuous ability predict the nominal response. It accounts for guessing with a different parameterization (discussed later) than the MC model does, and it does not employ the constraints of the MC DINA model.

3. The Scaling Individuals and Classifying Misconceptions Model

As a blend of the NR DCM and NR IRT models, the SICM model classifies examinees according to attributes (defined as misconceptions) and measures examinee ability. The model posits there is a continuous trait, denoted θ_e, that largely explains the covariance among item responses. However, unlike in unidimensional IRT, responses to items measuring the same misconception(s) are not independent conditional on θ_e alone. Thus, the model additionally assumes that there exists a set of categorical misconceptions, each of which an examinee does or does not possess, that systematically account for the variations in the selections amongst the incorrect options. Individually denoted by α_{ea}, misconceptions are assumed to be dichotomous latent variables where, for examinee e, misconception a is either present/possessed ($\alpha_{ea} = 1$) or absent/not possessed ($\alpha_{ea} = 0$). Marginally, each misconception has a Bernoulli distribution with the probability an examinee possesses the misconception, p_a (i.e., $\alpha_{ea} \sim B(p_a)$). The misconception pattern $\boldsymbol{\alpha}_e$ is a vector of A binary indicators representing the presence or absence of each misconception (i.e., $\boldsymbol{\alpha}_e = [\alpha_{e1}\alpha_{e2}\ldots\alpha_{eA}]$). As such, $\boldsymbol{\alpha}_e$ has a multivariate Bernoulli distribution (Maydeu-Olivares & Joe, 2005) with a $2^A \times 1$ vector $\boldsymbol{p}$, representing the 2^A pattern probabilities (i.e., $\boldsymbol{\alpha}_e \sim MB(\boldsymbol{p})$).

Given a set of $j = 1, \ldots, J_i$ response categories or possible options for item i, the SICM model is a confirmatory mixture nominal item response model that defines the probability of observing an examinee's nominal response pattern to I items ($\boldsymbol{x_e}$) as

$$P(\boldsymbol{X_e} = \boldsymbol{x_e}) = \int_{-\infty}^{\infty} \sum_{c=1}^{2^A} \upsilon_c \prod_{i=1}^{I} \prod_{j=1}^{J_i} \pi_{n_{ij}|\boldsymbol{\alpha}_c,\theta}^{[x_{ei}=n_{ij}]} P(\theta)\, d\theta. \tag{1}$$

The terms υ_c and $P(\theta)$ are the so-called structural components of the model, describing the distributions of and relationships among the latent variables in the model, with θ and $\boldsymbol{\alpha}$ assumed to be orthogonal as in the original bifactor model (Gibbons & Hedeker, 1992). Theoretically, we expected that students vary in ability even when they possess the same misconception pattern, meaning a significant correlation between ability and misconception pattern was not expected or

modeled. Unlike subfactors in a typical bifactor model, individual misconceptions are not held independent in the SICM model, rather they are correlated as attributes in DCMs are. Each latent class represents a unique misconception pattern such that given A misconceptions, there are 2^A unique patterns. The set of υ_c parameters represent the proportion of examinees within a given misconception pattern (i.e., latent class) c. These terms contain the structural information for all misconceptions including the marginal proportion of examinees for a given misconception $a(p_a)$ and the bivariate association between any pairs of misconceptions (often summarized by a tetrachoric correlation coefficient). The term υ_c is parameterized as a function of the individual misconceptions by a log-linear model (Henson & Templin, 2005; see also Rupp et al., 2010 and Xu & von Davier, 2008). The term $P(\theta)$ is the density function of ability, with $\theta \sim N(0,1)$ for identifiability.

The parameter $\pi_{n_{ij}|\boldsymbol{\alpha}_c,\theta}$ is the conditional probability that an examinee with misconception profile (or in class) c with ability θ selects option j from the set of J_i options for item i (i.e., $x_{ei} = n_{ij}$). The Iverson brackets $[\cdot]$ indicate if $x_{ei} = n_{ij}$, then $[x_{ei} = n_{ij}] = 1$; otherwise, $[x_{ei} = n_{ij}] = 0$. The parameter $\pi_{n_{ij}|\boldsymbol{\alpha}_c,\theta}$ represents the measurement component of the SICM model: It quantifies how the latent variables are related to the observed item responses. In the model, ability is measured by the correct option and misconceptions by incorrect options. Not every incorrect option measures each misconception, so an indicator variable is used to specify when a misconception is measured by an option. Mimicking DCM practices, specifications are set a priori and are described in an item-option by misconception Q-matrix (Tatsuoka, 1990). Cells of the Q-matrix are indicators, where $q_{n_{ij}a} = 1$ if option j on item i measures misconception a, and $q_{n_{ij}a} = 0$ otherwise. An additional indicator $c_{n_{ij}}$ specifies which option measures θ; $c_{n_{ij}} = 1$ if, and only if, option j is the correct answer to item i.

The SICM model parameterizes $\pi_{n_{ij}|\boldsymbol{\alpha}_c,\theta}$ in (1) for examinee e in class c where $\pi_{n_{ij}|\boldsymbol{\alpha}_c,\theta} = \pi_{n_{ij}|\boldsymbol{\alpha}_e,\theta_e}$ by utilizing a multicategory logistic regression framework (e.g., Agresti, 2002) that models the $J_i - 1$ non-redundant logits with the Jth option as the baseline category as

$$\log\left(\frac{P(X_{ei} = n_{ij}|\boldsymbol{\alpha}_e,\theta_e)}{P(X_{ei} = n_{iJ}|\boldsymbol{\alpha}_e,\theta_e)}\right) = \lambda_{n_{ij},0} + \lambda^*_{n_{ij},\theta}(\theta_e)(c_{n_{ij}}) + \boldsymbol{\lambda}^{T*}_{n_{ij}}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}) \tag{2}$$

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

$$\lambda^*_{n_{ij},\theta}(\theta_e)(c_{n_{ij}}) = \lambda_{n_{ij},\theta}(\theta_e)(c_{n_{ij}}) - \lambda_{n_{iJ},\theta}(\theta_e)(c_{n_{iJ}}); \tag{3}$$

$$\boldsymbol{\lambda}^{T*}_{n_{ij}}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}) = \boldsymbol{\lambda}^{T}_{n_{ij}}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}) - \boldsymbol{\lambda}^{T}_{n_{iJ}}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{iJ}}). \tag{4}$$

Specifying the correct option as the baseline category, denoted n_{iJ}, simplifies (2): In (3), $c_{n_{ij}}$ equals zero for every option n_{ij} where $j \neq J$ because the incorrect options do not measure θ, and in (4), $q_{n_{iJ}}$ always equals zero because the correct option does not measure any misconceptions. Therefore, the $J_i - 1$ equations specifying the log-odds of selecting an incorrect option over the correct option in the SICM model can be equivalently formulated as

$$\log\left(\frac{P(X_{ei} = n_{ij}|\boldsymbol{\alpha}_e,\theta_e)}{P(X_{ei} = n_{iJ}|\boldsymbol{\alpha}_e,\theta_e)}\right) = \lambda_{n_{ij},0} - \lambda_{n_{iJ},\theta}(\theta_e) + \boldsymbol{\lambda}^{T}_{n_{ij}}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}) \tag{5}$$

for every n_{ij} such that $j \neq J$. The conditional probability option n_{ij} is selected is expressed as

$$P(X_{ei} = n_{ij}|\boldsymbol{\alpha}_e,\theta_e) = \frac{\exp(\lambda_{n_{ij},0} - \lambda_{n_{iJ},\theta}(\theta_e) + \boldsymbol{\lambda}^{T}_{n_{ij}}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}))}{\sum_{j=1}^{J}\exp(\lambda_{n_{ij},0} - \lambda_{n_{iJ},\theta}(\theta_e) + \boldsymbol{\lambda}^{T}_{n_{ij}}\boldsymbol{h}(\boldsymbol{\alpha}_e,\boldsymbol{q}_{n_{ij}}))}. \tag{6}$$

The intercept $\lambda_{n_{ij},0}$ is the logit of selecting an incorrect option n_{ij} over the correct option n_{iJ} for an average ability examinee who possesses *none* of the misconceptions measured by option n_{ij}. The more appealing the option is, the larger the intercept will be. The term $\lambda_{n_{iJ},\theta}$ is the loading for ability and is the discrimination parameter for ability as in IRT.

Using notation consistent with the LCDM, the term $\boldsymbol{\lambda}_{n_{ij}}^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_{n_{ij}})$ is a linear combination of misconception main effects and interactions. The term $\boldsymbol{\lambda}_{n_{ij}}$ is a $2^A \times 1$ vector of possible main effects and interactions for a given option n_{ij}. The term $\boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_{n_{ij}})$ is a $2^A \times 1$ vector of indicators with elements that equal one if and only if (a) the option measures each misconception corresponding to the parameter ($q_{n_{ij}a} = 1$ for every relevant a), and (b) the examinee possesses each misconception corresponding to the parameter ($\alpha_{ea} = 1$ for every relevant a). More generally, $\boldsymbol{\lambda}_{n_{ij}}^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_{n_{ij}})$ is

$$\sum_{a=1}^{A} \lambda_{n_{ij},1(a)}(\alpha_{ea} q_{n_{ij}a}) + \sum_{a=1}^{A-1} \sum_{b=a+1}^{A} \lambda_{n_{ij},2(ab)}(\alpha_{ea}\alpha_{eb} q_{n_{ij}a} q_{n_{ij}b}) + \cdots. \tag{7}$$

The first subscript for the effects (λs) specifies the item and option to which the effect refers, the second subscript denotes the level of the effect (intercepts = 0, main effects = 1, two-way interactions = 2, all the way to A-way interactions = A), and the remaining parenthetical subscripts list the misconceptions(s) to which the effect refers. The first summation includes main effects where $\lambda_{n_{ij},1(a)}$ is the main effect for misconception a for the jth option of item i. The double summation specifies two-way interaction effects for options measuring two or more misconceptions, where $\lambda_{n_{ij},2(ab)}$ is the two-way interaction effect between misconceptions a and b for the jth option of item i. The model also incorporates higher-order misconception interactions for options measuring more than two misconceptions. The ellipses denote the third through Ath higher-order interactions, where $\lambda_{n_{ij},A(1,2,\ldots,A)}$ is the A-way interaction effect between all A attributes. Main effects and interactions are discrimination parameters with respect to misconception patterns; as $\boldsymbol{\lambda}_{n_{ij}}^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_{n_{ij}})$ increases, the difference in the probability of answering the item correctly for examinees with and without the misconceptions increases.

To identify the SICM model, for convenience we treated the correct option as the baseline category and set all parameters for this option equal to zero. We also constrained the main effect and interaction parameters to ensure monotonicity for misconceptions and for ability, meaning (a) the possession of a misconception never led to a decrease in the probability of selecting an option measuring that misconception, and (b) an increase in ability never resulted in a decrease in the probability of answering the item correctly.

3.1. Lower Asymptote for SICM Model

The specification of the SICM model in (5) does not provide a lower asymptote for the probability of a correct response to account for guessing on a multiple-choice test, yielding one incorrect option that will be chosen with probability one as ability decreases. We developed an alternative formulation for the SICM model that provides a lower asymptote *without* adding an additional parameter to the model:

$$\log\left(\frac{P(X_{ei} = n_{ij}|\boldsymbol{\alpha}_e, \theta_e)}{P(X_{ei} = n_{iJ}|\boldsymbol{\alpha}_e, \theta_e)}\right) = \lambda_{n_{ij},0} - \exp\left(\lambda_{n_{iJ},\theta}(\theta_e)\right) + \boldsymbol{\lambda}_{n_{ij}}^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_{n_{ij}}). \tag{8}$$

The difference in (5) and (8) is the ability portion of the model is now exponentiated. The intercept of the model in (8) is interpreted as the logit that an examinee with an extremely low ability who possesses no misconceptions will select option j over the correct answer. Holding other parameters constant, as ability decreases, the value of $\exp(\lambda_{n_{iJ},\theta}(\theta_e))$ decreases, meaning the logit of selecting the correct answer also decreases which satisfies the monotonicity assumption for the model. As ability approaches negative infinity, $\exp(\lambda_{n_{iJ},\theta}(\theta_e))$ approaches 0, meaning the logit in (8) approaches $\lambda_{n_{ij},0} + \boldsymbol{\lambda}_{n_{ij}}^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_{n_{ij}})$. Because the DCM item parameters for the base-

line category J are equal to zero (and $-\exp(0) = -1$), this formulation yields a lower asymptote for the correct response where

$$\lim_{\theta \to -\infty} (\pi_{n_{iJ}|\theta_e, \boldsymbol{\alpha}_e}) = \frac{\exp(-1)}{\exp(-1) + \sum_{j=1}^{J-1} \exp(\lambda_{n_{ij},0} + \boldsymbol{\lambda}_{n_{ij}}^T \boldsymbol{h}(\boldsymbol{\alpha}_e, \boldsymbol{q}_{n_{ij}})}. \tag{9}$$

This formulation 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 three-parameter logistic (3PL) IRT model (Baker & Kim, 2004).

3.2. Estimation of the SICM Model

We estimated the SICM model with the lower asymptote (see (8)) with a Fortran-based MCMC estimation algorithm that uses Metropolis–Hastings sampling. Supplemental materials available from the first author detail the specifics of the algorithm. To evaluate the performance of the SICM model and algorithm, a simulation study was conducted and is discussed next. An empirical analysis of data from the Force Concepts Inventory follows.

4. Simulation Study

The SICM model is complex due to the large number of parameters estimated and the different types (i.e., continuous and categorical) of estimated examinee parameters. This simulation study provided information about (a) the performance of the model under realistic testing situations, and (b) the interplay of a continuous ability and a set of categorical misconceptions within a single model. To our knowledge, estimation of continuous and categorical latent variables in the measurement portion of a psychometric model has never been tried at the item option level before, so it was of interest to determine whether the random effect of one type of variable would mask the other's random effect.

4.1. Simulation Study Design

The conditions for the simulation study were informed by existing concept inventories, preliminary data analyses, current large-scale testing practice, and previous simulation research on DCMs. Combining these resources with our own experience estimating DCMs, we chose conditions to reflect our projections for a typical range of characteristics that tests developed for the SICM model would have in practice. The study had four manipulated factors that were fully crossed: number of misconceptions (3 and 6), sample size (3,000 and 10,000), test length (30 and 60 items), and the magnitude of the main effects for ability and misconceptions (labeled as "low" and "high").

4.1.1. Number of Misconceptions In the literature, the number of attributes estimated for LCDM-based DCM applications range from one (Templin & Bradshaw, 2013) to 18 (Henson & Templin, 2004). The number of attributes found in applications and simulations based on the unconstrained LCDM parameterization typically ranges from three to five (e.g., Choi, 2009; de la Torre, 2011; Henson et al., 2009; Templin & Hoffman, 2013; Kunina-Habenicht et al., 2012). We estimated the lower end of this range and slightly extended the upper because current concept inventories often target a larger number of misconceptions.

4.1.2. Sample Size Based on previous simulation studies with the NR DCM (Templin & Bradshaw, under review) and the NR IRT model (DeMars, 2003), we anticipated the complexity of the SICM model would necessitate larger samples. The smaller sample of 3,000 would be considered large, yet attainable for a research application. Although our empirical data analyses that follows is a sample near 10,000 collected for research purposes, this large sample is more commonly available for operational large-scale tests, such as mandated state-level tests.

4.1.3. Test Length The test length of 30 items is typical for DCM applications (e.g., Lee, Park, & Taylan, 2011) and in K-12 benchmark testing (e.g., Henson & Templin, 2008). We added the longer 60 item test conditions to test the SICM model under conditions more typical for end-of-course state testing (e.g., Henson, Templin, & Willse, 2013), perhaps the predominant context where diagnostic feedback is mandated (NCLB, 2001) yet found lacking (Huff & Goodman, 2007; Perie, Marion, Gong & Wurtzel, 2007).

4.1.4. Item Parameters Although we did not have previous research to guide how to specify a combination of continuous and categorical predictors of the item response, we chose reasonable parameters based on our experience and preliminary results. The "high" main effects were chosen to reflect stronger test designs where traits were more highly related to item outcomes, and the "low" main effect were chosen to evaluate how the model performs with weaker tests. "Low" main effects were drawn from $U(0.3, 0.6)$ for ability and from $U(0.75, 1.25)$ for misconceptions; "high" main effects were drawn from $U(0.6, 0.8)$ for ability and $U(1.75, 2.25)$ for misconceptions. We manipulated the magnitudes and relative magnitudes of the main effects for these latent variables by crossing the high and low main effect conditions. For example, the low misconception/high ability main effect conditions had high ability main effects in both an absolute and relative sense, whereas the high misconception/high ability main effect conditions had high ability main effects only in an absolute sense. Item intercepts were drawn from $U(-1, 1)$ and two-way interactions were drawn from $U(0.5, 1)$.

4.1.5. Q-Matrix Each simulated item had four options ($J_i = 4$), as is most typical for items on multiple-choice tests. Each incorrect option was specified to measure either one or two misconceptions. The Q-matrix was balanced, with a mean of 2.1 misconceptions measured per item and 1.13 misconceptions measured per option. Concept inventories commonly measure one misconception per incorrect option, making our simulation Q-matrix more complex than current tests that may be aligned with the SICM model. Generally, as the number of "1" entries in a Q-matrix increases, the more complex the Q-matrix is. Increased complexity increases estimation difficulty due to the number of parameters to estimate and also the type of parameters, as interaction terms are the most difficult parameters to estimate (Kunina-Habenicht et al., 2012); however, increased complexity in turn provides more information with which to classify examinees while holding the number of items constant. This Q-matrix was a compromise between reflecting current concept inventory designs and theoretically testing and demonstrating other potentially useful test designs.

4.1.6. Examinee Traits Examinee ability was drawn from a standard normal distribution, as is common practice in IRT. For all conditions, the tetrachoric correlation between misconceptions was set to 0.5 to reflect that misconceptions were expected to be related, yet distinct, traits. Preliminary analyses suggested misconception correlations were lower than correlations of 0.7 more typically found in applications and simulations for attributes in DCMs (e.g., Rupp & Templin, 2008). Originally, conditions of 0.25 and 0.50 were chosen and 64 simulation conditions were estimated, but results for the 0.25 conditions are not reported herein because estimation accuracy was similar under both correlation conditions.

TABLE 1.
Estimation accuracy for item and structural parameters.

τ	E	I	A	*Bias* $(\hat{\tau})$	*RMSE* $(\hat{\tau})$	$r(\hat{\tau}, \tau)$
Item	3,000	30	3	−0.001 (0.014)	0.028 (0.006)	0.976 (0.005)
			6	−0.002 (0.015)	0.043 (0.007)	0.961 (0.009)
		60	3	0.001 (0.011)	0.021 (0.004)	0.984 (0.002)
			6	−0.001 (0.013)	0.027 (0.005)	0.978 (0.003)
	10,000	30	3	0.000 (0.007)	0.009 (0.003)	0.993 (0.002)
			6	−0.001 (0.008)	0.018 (0.003)	0.986 (0.004)
		60	3	0.000 (0.006)	0.007 (0.002)	0.995 (0.001)
			6	−0.001 (0.006)	0.009 (0.002)	0.993 (0.001)
Structural	3,000	30	3	0.000 (0.002)	0.108 (0.038)	0.990 (0.008)
			6	−0.007 (0.016)	0.238 (0.065)	0.974 (0.011)
		60	3	0.000 (0.001)	0.075 (0.029)	0.995 (0.004)
			6	−0.002 (0.009)	0.124 (0.027)	0.993 (0.003)
	10,000	30	3	0.000 (0.001)	0.056 (0.019)	0.997 (0.002)
			6	−0.002 (0.008)	0.116 (0.026)	0.993 (0.003)
		60	3	0.000 (0.001)	0.042 (0.014)	0.999 (0.001)
			6	−0.001 (0.005)	0.067 (0.015)	0.990 (0.008)

Note. Standard deviations are given in parentheses. Simulation conditions are given by E = number of examinees, I = number of items, and A = number of attributes.

4.1.7. Estimation Fifty replications were estimated for each of the 32 conditions. The MCMC estimation algorithm used a thinning interval of five to retain a posterior chain of length 1,000 after a burn-in period of 5,000 steps. These values were a compromise between reasonable estimation time and adequate convergence rates as evaluated by time series plots, density plots, and Gelman and Rubin's (1992) $\hat{R}$ statistic for sample replications.

4.2. Simulation Results

Results are provided in Tables 1–4 where values were (a) averaged *across* the magnitude of main effects factor or (b) averaged across all other factors and given *by* the magnitude of main effects factor. In short, results indicated the item, structural, and examinee parameters were accurately estimated with the MCMC algorithm. Generally, parameter estimates were most accurate in conditions with more examinees, more items, and fewer misconceptions. 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 relevant simulation studies in the DCM literature (e.g., Choi, 2009; Henson et al., 2009; Kunina-Habenicht et al., 2012; Templin & Bradshaw, under review). Results from varying the magnitude of the main effects provided information about which conditions yielded greater estimation accuracy when continuous and categorical traits are estimated together. Results for estimation accuracy, classification accuracy, and reliability are given subsequently.

4.2.1. Model Parameter Estimates Table 1 includes the bias, root mean squared error (RMSE), and the Pearson correlations of the true (τ) and estimated $(\hat{\tau})$ model parameters. The RMSEs for item parameters were less than 0.05 for all conditions. Additional improvement in accuracy indicated by RMSE was negligible as the test length increased, number of misconceptions decreased, and sample size increased. The estimation accuracy of the structural parameters was most affected by the number of misconceptions, which was expected because the complexity of

TABLE 2.
Correct classification rates for individual misconception (α_a) and pattern ($\boldsymbol{\alpha}$) classification.

E	I	A	α_1	α_2	α_3	α_4	α_5	α_6	$\boldsymbol{\alpha}$
3,000	30	3	0.904	0.910	0.917				0.782
			(0.005)	(0.006)	(0.005)				(0.008)
		6	0.864	0.863	0.868	0.867	0.877	0.871	0.573
			(0.007)	(0.006)	(0.007)	(0.006)	(0.006)	(0.006)	(0.009)
	60	3	0.958	0.958	0.958				0.894
			(0.003)	(0.003)	(0.003)				(0.005)
		6	0.925	0.921	0.925	0.934	0.937	0.932	0.731
			(0.005)	(0.005)	(0.005)	(0.004)	(0.004)	(0.005)	(0.008)
10,000	30	3	0.906	0.912	0.918				0.786
			(0.003)	(0.003)	(0.003)				(0.004)
		6	0.867	0.866	0.871	0.870	0.879	0.874	0.580
			(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.004)	(0.005)
	60	3	0.959	0.959	0.960				0.896
			(0.002)	(0.002)	(0.002)				(0.003)
		6	0.926	0.922	0.926	0.936	0.938	0.934	0.735
			(0.002)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.004)

$\lambda_{1(a)}$	λ_θ	α_1	α_2	α_3	α_4	α_5	α_6	$\boldsymbol{\alpha}$
Low	Low	0.861	0.864	0.868	0.854	0.838	0.847	0.599
		(0.005)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)	(0.007)
	High	0.848	0.848	0.855	0.841	0.838	0.826	0.579
		(0.005)	(0.006)	(0.005)	(0.005)	(0.006)	(0.006)	(0.007)
High	Low	0.970	0.970	0.970	0.960	0.958	0.959	0.878
		(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.004)
	High	0.959	0.957	0.963	0.950	0.947	0.946	0.857
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)

Note. Standard deviations are given in parentheses. Simulation conditions are given by E = number of examinees, I = number of items, and A = number of attributes.

the structural model grows quickly as the dimensionality of the test increases. RMSEs for structural parameters were less than 0.11 when three misconceptions were measured. Improvement in structural parameter estimation was seen for the 6-misconception conditions as the number of items or examinees increased, although the correlation among true and estimated parameters was high (greater than 0.974) for all conditions.

4.2.2. Examinee Parameter Estimates Consistent with psychometric model research, the results for the accuracy of classifications (Table 2) and reliabilities of ability and misconceptions (Table 3) were less affected by the number of examinees responding to the test and more affected by the length of the test, number of misconceptions, and quality of the items. Classifications of misconceptions were generally accurate, as indicated by high correct classification rates (CCR > 0.90) of misconceptions for most conditions. The CCRs dropped below a value of 0.90 for conditions where 6 misconceptions were measured with 30 items. Improvement was seen in CCRs when main effects for misconceptions ($\lambda_{1(\alpha)}$) were high instead of low. Classification accuracy was almost identical for the 3,000 versus 10,000 examinee conditions.

The reliability of examinee ability estimates and classifications was evaluated using a reliability metric that is theoretically analogous for categorical and continuous traits (Templin & Bradshaw, 2013). The top portion of Table 3 gives results averaged across the magnitude of main

TABLE 3.
Reliability for examinee ability (θ) and individual misconceptions (α_a).

E	I	A	θ	α_1	α_2	α_3	α_4	α_5	α_6	$\alpha_.$
3,000	30	3	0.541	0.897	0.908	0.920				0.909
			(0.015)	(0.007)	(0.008)	(0.007)				(0.007)
		6	0.523	0.843	0.839	0.857	0.847	0.875	0.860	0.853
			(0.016)	(0.015)	(0.015)	(0.015)	(0.017)	(0.013)	(0.015)	(0.015)
	60	3	0.674	0.986	0.989	0.988				0.988
			(0.012)	(0.002)	(0.003)	(0.003)				(0.003)
		6	0.667	0.935	0.931	0.936	0.956	0.957	0.951	0.945
			(0.012)	(0.006)	(0.006)	(0.006)	(0.004)	(0.005)	(0.005)	(0.006)
10,000	30	3	0.542	0.897	0.908	0.920				0.908
			(0.008)	(0.004)	(0.004)	(0.004)				(0.004)
		6	0.523	0.842	0.841	0.857	0.846	0.872	0.860	0.853
			(0.009)	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
	60	3	0.675	0.987	0.988	0.989				0.988
			(0.006)	(0.001)	(0.002)	(0.002)				(0.002)
		6	0.669	0.937	0.933	0.937	0.957	0.957	0.952	0.946
			(0.006)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
$\lambda_{1(a)}$	λ_θ									
Low	Low		0.522	0.831	0.843	0.852	0.805	0.822	0.783	0.850
			(0.013)	(0.011)	(0.011)	(0.011)	(0.013)	(0.012)	(0.014)	(0.010)
	High		0.709	0.806	0.804	0.824	0.755	0.792	0.786	0.820
			(0.007)	(0.012)	(0.012)	(0.011)	(0.015)	(0.012)	(0.013)	(0.011)
High	Low		0.491	0.997	0.998	0.996	0.998	0.995	0.996	0.998
			(0.013)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
	High		0.687	0.990	0.987	0.993	0.983	0.988	0.987	0.991
			(0.008)	(0.002)	(0.002)	(0.001)	(0.003)	(0.003)	(0.003)	(0.002)

Note. $\alpha_.$ denotes the average of the reliability of the individual misconceptions. Standard deviations are given in parentheses. Simulation conditions are given by $E =$ number of examinees, $I =$ number of items, and $A =$ number of attributes.

effect conditions where reliability ranged from 0.523 to 0.675 for ability and, on average, from 0.853 to 0.988 for misconceptions ($\alpha_.$). The reliabilities of the misconceptions were high uniformly; however, the reliability of ability estimates reached values near 0.70 only when the main effects for ability were high, as seen in the bottom portion of Table 3. Although further disaggregated results are not presented within due to space, reliabilities for ability were higher than 0.75 when isolating the high main effect ability conditions that also contained 60 items.

4.3. Simulation Study Discussion

The results of the simulation study provide evidence for the efficacy of the estimation algorithm and demonstrate the accuracy and reliability of SICM model estimates under various testing conditions. The only concern raised by the simulation results was the reliability of ability estimates in some conditions. The results showed that higher reliability can be achieved when the test includes items that are adequately related to the traits (i.e., when main effects for misconceptions and ability are reasonably strong) and when more items are available for estimation. To provide a frame of reference for these simulated conditions, the empirical data analysis that follows yielded main effects that would be considered "medium" or somewhere in the middle of our simulated high and low main effects for ability and misconceptions ($\hat{\lambda}_{.\theta} = 0.526$; $\hat{\lambda}_{.\alpha} = 1.348$).

However, main effects in practice may be larger if this test was designed from the SICM framework a priori because test developers would have the opportunity to revise items to improve relationships between options and latent traits. Nonetheless, the SICM model requires more items to estimate stable abilities in comparison to the unidimensional Rasch model where only 35 items were needed to yield a reliability of 0.80 using the same reliability metric under a set of simulated conditions (Templin & Bradshaw, 2013). We emphasize that both the quality and quantity of items is vital given the strong influence on the ability estimate reliability.

Results disaggregated by the strength of the main effects reiterated the reliance of estimation accuracy on quality items. Accuracy and reliability of the estimated latent traits (both ability and misconceptions) were greatest when main effects were high in an absolute sense, and estimation improved only slightly when the main effect was low for the other trait (i.e., high relative magnitude did not substantially improve estimation). Thus, we did not observe the different latent variable types masking one another. When estimating the SICM model in practice, the larger concern for estimation regarding main effects is the strength of the main effects in an absolute sense. The model requires items to have strong relationships with ability and misconceptions to provide enough information to both classify and scale examinees.

These simulation results should be interpreted with the understanding that the estimation model was the correct model for the data and that others factors may impact the accuracy of SICM model estimates in practice. Q-matrices for other tests may be different in terms of complexity and accuracy. Subsequent studies may vary the complexity of the Q-matrix to study both simpler and more complex test designs. Future research may also investigate the model under conditions when perfectly accurate Q-matrices are not used. In practice, a perfectly accurate Q-matrix is one where the misconception-option relationships specified a priori by content experts are all correct. We did not examine the impact of Q-matrix misspecification in this study. We also did not examine scenarios where main effects of misconceptions and ability were mixed within a test. We instead designated absolute and relative magnitudes across the test.

5. The SICM Model Illustrated through an Empirical Data Analysis

We analyzed a Force Concept Inventory (FCI) data set of 10,039 examinees to demonstrate the SICM model in an empirical setting. The examinees were high school students enrolled in an Advanced Placement, honors, or regular education physics course in the United States from 1995–1999. The responses were pretest FCI data, meaning students had not yet had any instruction to correct misconceptions at the time they completed the inventory. The goal of the FCI was to identify misconceptions students have about Newtonian force, as well as to measure their overall facilities in reasoning with Newtonian force concepts. The most common method for scoring the test is by creating total scores for ability and subscores for misconceptions. As an alternative, the SICM model, being well-aligned with the purpose of this assessment, was used to provide categorical attribute feedback about misconceptions and scaled estimates of ability.

Hestenes et al. (1992) list 31 naïve conceptions that are probed by at least one option on the FCI. For this analysis, we selected the first three misconceptions on the list that were measured by at least five items to measure as categorical latent variables. These misconceptions will be referred to as: Misconception 1, *impetus dissipation* (α_1), Misconception 2, *gradual/delayed impetus build-up* (α_2), and Misconception 3, *only active agents exert force* (α_3). The first two misconceptions regard impetus, which can be thought of as a motion-generating property of an object. These two notions of impetus reflect misunderstandings of Newton's Second Law that states that the acceleration and direction of an object in motion changes as a function of net force and mass, not through dissipation (decrease) or build-up (increase). The third misconception reflects the misunderstanding that an object must be in motion in order to exert force upon another

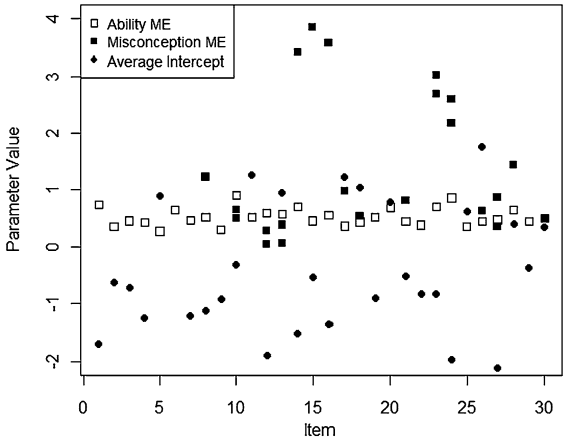

FIGURE 1.
SICM Item Parameters for FCI. The values of three types of item parameters are plotted: option-specific intercepts (averaged within item; Average Intercept), item-specific ability main effects (Ability ME), and option-specific misconception main effects (Misconception ME).

object (e.g., if a moving bicycle runs into a stationary tree, the tree exerts no force on the bicycle). Further explanations of these misconceptions can be found in Hestenes et al. (1992). Sixteen of the 30 items on the FCI measured 1 of the 3 misconceptions used for this analysis. Each of the 3 misconceptions was measured by 6 items and 10, 7, and 6 options, respectively.

To estimate the SICM model using the MCMC algorithm, we used a minimum burn-in of 2,000 and fixed the posterior chain-length to 1,000 (after thinning with an interval of five) and then allowed the algorithm to run until 90 % of the parameters converged. Convergence was determined by an $\hat{R}$ statistic (Gelman & Rubin, 1992) less than 1.5. A lognormal (0, 0.5) prior was placed on the main effect for ability to reflect a range of values within which we expected the parameters to be. After 10,000 steps, 94.4 % of the parameters had converged. In the following sections, we provide the fit of the data with the SICM model in comparison with a logical competing model and then describe and discuss the SICM model estimates.

5.1. Results

5.1.1. Model Data Fit The SICM model with a lower asymptote and the NR IRT model with a lower asymptote (formed by exponentiation of the ability portion of the model, analogous to the parameterization used in the SICM model) were used to analyze the FCI data. According to the deviance information criterion (DIC; Spiegelhalter, Best, Carlin & van der Linde, 2002), the SICM model (DIC = 740, 642.3) better represented the phenomena reflected by the FCI data than the NR IRT model did (DIC = 754, 776.4). The relative model-data fit results suggest the test is measuring more than a single continuous trait because the model including latent misconceptions as predictors yielded more parsimonious model-data fit than the model that only predicted nominal responses with a continuous trait. Notably, ability estimates in the SICM model had a correlation of 0.869 with the NR IRT model estimates. The following sections illustrate the types of information the SICM model provides about the FCI items and examinees.

5.1.2. Item Parameters Figure 1 provides the estimated ability and misconception main effects for each item, as well as the average intercept value at the item level. The mean intercept was -0.464 ($SE = 0.062$), the mean main effect for ability was 0.526 ($SE = 0.020$), and the mean misconception main effect was 1.348 ($SE = 0.178$). For average ability examinees, possessing a misconception increased the probability of selecting the incorrect response associated with that

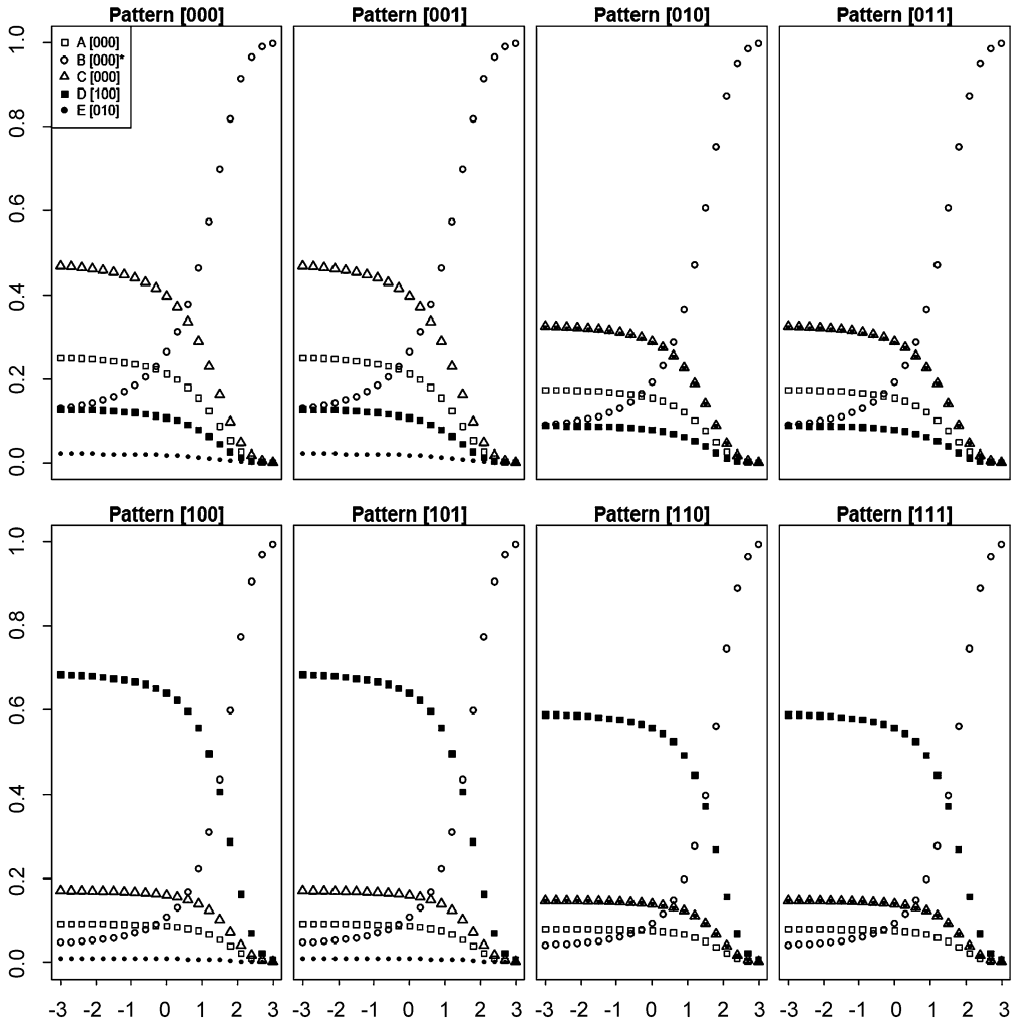

FIGURE 2.
Option characteristic curves for FCI Item 23. The nominal response probabilities differ by the pattern of misconceptions that examinees have (e.g., the *top left* graph is for Pattern [000] meaning these examinees possess no misconceptions). Q-matrix entries, given in the *legend* as $[q_{n_{ij},\alpha_1} q_{n_{ij},\alpha_2} q_{n_{ij},\alpha_3}]$, show D measures Misconception 1 and E measures Misconception 2. Denoted by $*$, B is the correct answer.

misconception on average (accounting for class membership weights) by 10.8 %, 10.5 %, and 29 % for Misconceptions 1–3, respectively.

To illustrate the item response function (IRF) of the SICM model, we present results for Item 23,[1] one of the more complex items that measured two misconceptions. The correct answer to this item is B ($\hat{\lambda}_{B,\theta} = 0.701$ ($SE = 0.019$)). Option D reflects the impetus dissipation misconception (α_1). Option E reflects the gradual/delayed impetus build up misconception (α_2). Each of these misconceptions had large main effects ($\hat{\lambda}_{D,1(1)} = 2.697$, $\hat{\lambda}_{E,1(2)} = 3.027$) providing some empirical evidence that the options measure the misconceptions. Figure 2 provides the IRF for Item 23, illustrating that the SICM model provides each DCM-like class ($[\alpha_1\alpha_2\alpha_3]$) a unique set of NR IRT-like OCCs. Note the features of the SICM model OCCs for incorrect options that are

[1]This item, along with all items on the current version of the FCI, is available by request from Halloun et al. (1995).

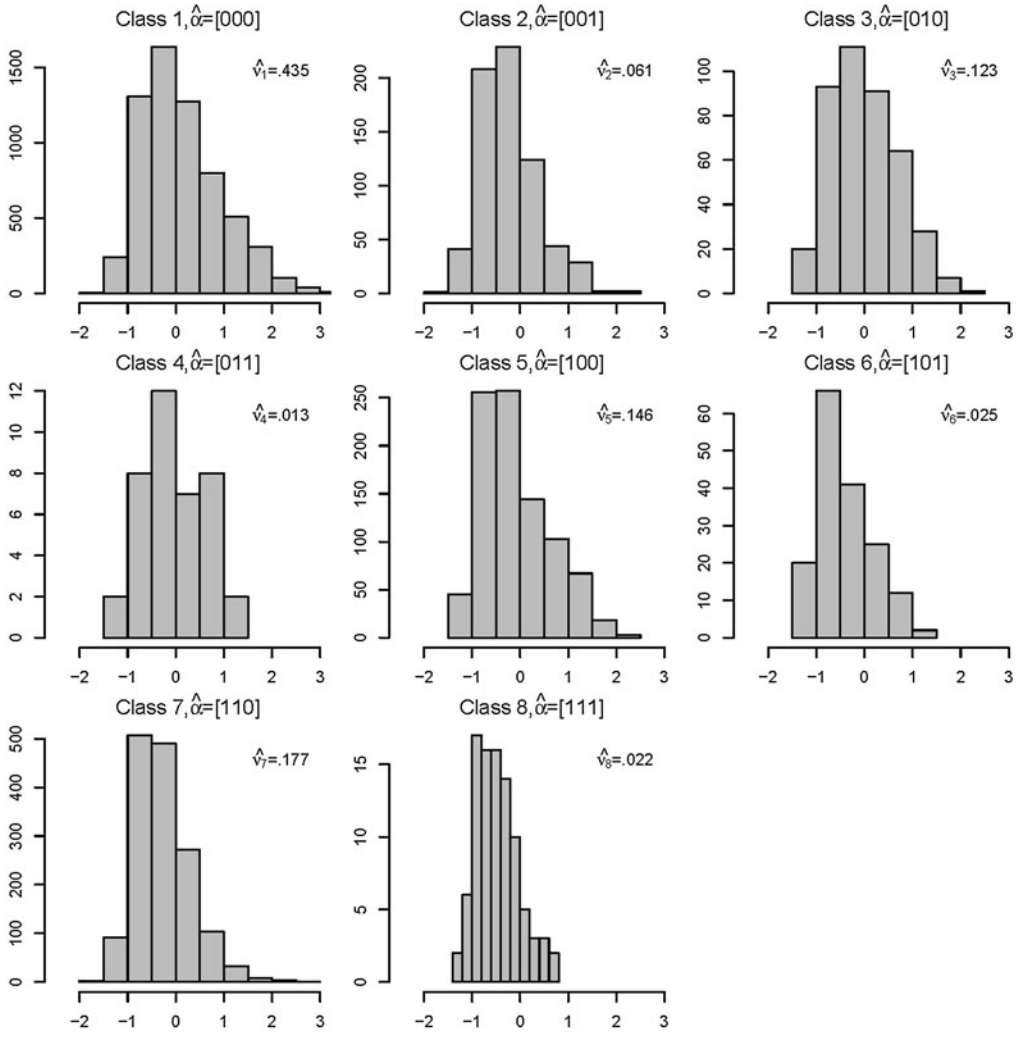

FIGURE 3.
Examinee ability distribution by misconception pattern. Each histogram within this figure provides the estimated ability ($\hat{\theta}$) frequencies for a given class, where classes are defined by unique misconception patterns ($\hat{\alpha}$). The distribution of class membership is also provided, where ν_c is proportion of examinees in class c. For example, 17.7 % of examinees have misconception pattern [110].

distinct from NR IRT model: they have an upper asymptote, decrease monotonically, and do not intersect. The *order* of the probability an examinee selects a given incorrect option is dependent upon his or her misconceptions and is invariant with respect to his or her ability. Put differently, the order varies across classes and is invariant within a class.

When an examinee possesses a misconception, the examinee should be more likely to select options corresponding to the misconception. Figure 2 shows that Option D is more effective at discriminating among students who do and do not have Misconception 1 than Option E is at discriminating among examinees who do and do not have Misconception 2. However, each option has a large impact on the nominal response probabilities illustrated by the OCCs.

5.1.3. Examinee Ability and Classifications Figure 3 summarizes the estimated examinee parameters in the sample. For each misconception pattern, ability frequencies are plotted, and the

distribution of class membership is noted by ν_c, the proportion of examinees in class c. Misconception pattern [000] was the most common, with 43.5 % of examinees having no misconceptions ($\hat{\nu}_1 = 0.435$). Approximately 12 % of examinees possessed Misconception 3, reflecting the lower class membership in patterns containing this misconception: [001], [011], [101], and [111]. Respectively, 36.84 % and 33.34 % of examinees possessed Misconceptions 1 and 2. The structural ν_c parameters also provide the tetrachoric correlations among the misconceptions: Misconceptions 1 and 2 had a correlation of 0.509, and Misconception 3 had near zero correlations with 1 and 2. The lack of correlation can be understood as follows: about half of the examinees with Misconception 3 did possess Misconception 1 and the other half did not, and an analogous statement can be made for Misconception 2. Figure 3 also shows that examinees' abilities vary within misconception patterns.

To illustrate the practical utility of the estimates provided to examinees from the SICM model, we discuss results for two examinees. In terms of commonly-provided total scores and overall ability estimates, Examinee J and Examinee K's results are similar. They had total scores of 7 and 8, respectively, and ability estimates of $\hat{\theta}_J = -0.083$ and $\hat{\theta}_K = -0.093$. However, because these two examinees selected different types of incorrect answers, they were classified as having different misconception patterns ($\hat{\alpha}_J = [001]$, $\hat{\alpha}_K = [010]$). From the SICM model estimates, we can conclude that both of these examinees have near average ability, but Examinee J needs instruction relevant to Misconception 3, whereas Examinee K needs instruction relevant to Misconception 2. If only a total score or an overall ability estimate was given, conclusions may be drawn that these two examinees need similar instruction to improve their skills, where the SICM model misconception classifications distinguish instructional needs.

For each set of items measuring a misconception, two types of estimates are given in Table 4 for the two examinees: a subscore and the SICM estimated probability that the examinee possesses the misconception. These results demonstrate that subscores are not consistent with SICM estimates of misconceptions. For example, Examinee K selected an option reflective of Misconception 1 for three of six items and has a low probability of (0.110) of possessing Misconception 1. Meanwhile, Examinee J selected an option reflective of Misconception 3 for three of the six items and has a high probability (0.690) of possessing Misconception 3. If observed score methods coupled with standard-setting methods were used where a designated cut-off subscore of three, for example, was utilized to diagnose misconceptions, different conclusions would be reached about which examinees possessed each misconception.

5.2. Data Analysis Discussion

Alignment of psychometric and cognitive theory is important for providing both accurate estimates of examinees' cognitive ability and advancing cognitive theories. Modeling the presence of misconceptions in addition to an overall ability, the SICM model better reflects empirical theories of learning held by physics educators who developed the FCI. Due to characteristics of the test design where some misconceptions were measured by only one to three items, some misconceptions specified by the developers could not be modeled in this analysis. However, the analysis demonstrated how a subset of misconceptions that were measured by a reasonable number of items could be modeled in the SICM framework.

The results illustrated the utility the SICM model estimates have beyond CTT or IRT estimates for examinees. For an IRT model with an estimated discrimination parameter, *which* items an examinee answers correctly matters. The same total scores can yield different ability estimates, or different total scores can yield the same ability estimates. The SICM model uses information not only from *which* items an examinee answers incorrectly, but also *why* an examinee answers items incorrectly. As a result, two examinees could have the exact same scored response

TABLE 4.
Results for two examinees.

		Examinee (e)	
		J	K
Items Measuring Misconception 1	**Option**	**Response**	
	$(q_{n_{ij}1} = 1)$	X_{ei}	X_{ei}
Item 12	C, D	B	C[a]
Item 13	A, B, C	C[a]	C[a]
Item 14	E	D	A
Item 23	D	B	D[a]
Item 24	C, E	C[a]	A
Item 27	B	C	D
CTT estimate	Subscore	2	3
SICM (DCM-like) estimate	$P(\alpha_1 = 1) = \hat{\alpha}_1$	0.093	0.110
Items Measuring Misconception 2	**Option**	**Response**	
	$(q_{n_{ij}2} = 1)$	X_{ei}	X_{ei}
Item 8	D	E	D[a]
Item 10	B, D	A	D[a]
Item 21	D	E	D[a]
Item 23	E	B	D
Item 26	C	D	C[a]
Item 27	D	C	C
CTT estimate	Subscore	0	4
SICM (DCM-like) estimate	$P(\alpha_2 = 1) = \hat{\alpha}_2$	0.189	0.751
Items Measuring Misconception 3	**Option**	**Response**	
	$(q_{n_{ij}3} = 1)$	X_{ei}	X_{ei}
Item 15	D	D[a]	C
Item 16	D	D[a]	A
Item 17	E	D	A
Item 18	A	D	C
Item 28	B	B[a]	B[a]
Item 30	A	D	B
CTT estimate	Subscore	3	1
SICM (DCM-like) estimate	$P(\alpha_3 = 1) = \hat{\alpha}_3$	0.690	0.038
Overall			
CTT estimate	Total Score	8	7
SICM (IRT-like) estimate	$\hat{\theta}_e$	−0.083	−0.093
SICM (DCM-like) estimate	$\hat{\boldsymbol{\alpha}}_e$	[001]	[010]

[a]Indicates a response that is reflective of the corresponding misconception.

pattern and be classified as possessing very different sets of misconceptions. Future research is needed to examine how students, teachers, and other stakeholders are able to interpret and apply these results, as well as which reporting formats best facilitate interpretation and use.

The SICM model results also quantify relationships among misconceptions and base-rates of misconceptions in the examinee population. Content experts can use these empirical estimates to advance their understandings about misconceptions and in turn about teaching strategies. For example, our results showed two of these misconceptions were quite common among beginning physics students. Substantial prevalence of a misconception may support integrating instruction

for overcoming the misconception directly into the curriculum, as opposed to supplying remediation for a selection of students.

The SICM model additionally helps inform cognitive science by providing a mechanism to improve the measurement of misconceptions. Unlike a CTT approach to misconception measurement where each option contributes equally to the subscore, fixed item parameters allow options to contribute differentially to the measure of a misconception. These misconception estimates are random effects where measurement error is quantified, so the practical utility of the scores can be evaluated considering this error. Further, the test itself can be improved because items and options can be empirically evaluated. Item parameters can identify options that are weakly related to the misconception, and these options can be revised or removed in the test development process. For example, Figure 1 shows weak main effects for misconceptions on Items 12 and 13, indicating that these items are not eliciting reasoning expected from students who have the corresponding misconception. Improved measurement of misconceptions can improve facilities we have to study misconceptions further in different populations and settings. Thus, a bidirectional argument can be made for utilizing the SICM model for data from tests similar to the FCI: cognitive theory informs the model to better estimate examinee traits, and the model parameters in turn inform cognitive theory to better understand examinee traits.

6. Conclusion

The SICM model presents a solution to a realistic need in and growing demand for assessment systems: to gain more feedback from tests about what students do not understand (Perie et al., 2007). Ranking individuals and providing diagnostic feedback are two "commonly co-occurring" purposes of a test that may be viewed as "fundamentally antithetical purposes" in commonly used testing paradigms (Wainer, Vevea, Camacho, Reeve, Rosa & Nelson, 2001, p. 342). However, the SICM model enables a test to serve both of these purposes by providing a type of diagnostic feedback that complements a scaled score. The efficacy of the SICM model under various testing conditions was demonstrated through a simulation study. Results suggested that the SICM model, coupled with specific test design characteristics, can enable diagnostic score reports detailing misconceptions in addition to providing a scaled ability estimate that is typically provided by current modeling and testing procedures. In practice, the diagnoses of misconceptions can be used for remediation purposes, and scaled examinee abilities can be used for comparative and accountability purposes.

The SICM model also provides a means to parametrically evaluate and quantitatively study phenomena of interest described in empirical cognitive theories: the interplay of student misconceptions and student ability. The SICM model fulfills a gap in psychometric theory by providing a model for data from concept inventories designed to measure misconceptions through incorrect options. As latent predictors of the item response, misconceptions are integral to the SICM model, and the binary measure of a misconception provided by the SICM model circumvents the need to cut continuous subscores to yield dichotomous classifications of examinees according to whether or not they possess a misconception. The SICM model utilizes the categorical classification methods of DCMs to provide a statistically reliable measure of a misconception as a categorical latent variable.

Although the SICM model provides a framework for modeling misconceptions as latent variables, future research can expand the SICM model by studying alternate ways to include misconceptions in the parameterization of the item response function. Like a bifactor model, the SICM model assumed traits contributing to variance beyond the primary trait were independent with the primary trait. We did not expect misconceptions to be highly correlated with ability because misconceptions often reflect a degree of sophistication and independence in reasoning

that is not present in lower ability examinees. However, in the future, an estimation algorithm that allows for the correlation of overall ability with misconceptions would allow this assumption to be empirically tested. A second suggestion may be to test whether ability interacts with each misconception at the item level, although this would be difficult to estimate under practical testing conditions. Finally, the misconceptions could be modeled without ability in the model. Although categorical latent variables typically are skills instead of misconceptions for DCMs, another model in the DCM family, the nominal response DCM, could also be applied to estimate misconceptions. The NR DCM may be of interest to researchers who are solely concerned with identifying misconceptions to inform instruction and remediation and are not concerned with scaling examinees' overall ability. These different parameterizations could further expand the psychometric tools available to answer empirical questions about learners' misconceptions.

In addition to psychometric model refinements, future research is needed on test construction methods for writing items and designing tests that diagnose misconceptions. Research in this area will benefit from the collaboration of psychometricians and content experts. The development of tests from the SICM framework a priori is critically important to maximize the model's usefulness. The FCI demonstrated that existing tests created in this fashion, even in the absence of available psychometric theory, may yield reasonable model-data fit when estimated retrospectively with the SICM model. Additionally, we expect if this analysis was treated as a pilot test administration, then test revisions guided by the SICM model results would improve fit and estimation precision. As demonstrated by the FCI data, current designs of concept inventories can serve as initial guides for test construction methodologies for the SICM model, but we anticipate the SICM model can help sharpen this design process by providing empirical feedback about the content validity of the option-attribute alignment.

Through this article, we provided information explaining how the SICM model can be estimated and applied. We hope future test development projects can build upon this information to leverage the model in practical settings to provide actionable information about where students' misunderstandings lie. For researchers seeking to develop a new test a priori using the SICM framework, we end with offering these pieces of advice: (1) Misconceptions must be stable latent traits of the examinee, at least during a testing occasion; the SICM model cannot reliably measure unsystematic errors that students inconsistently make. (2) Misconceptions are not synonymous with a lack of ability; a student who lacks a scientifically-accepted understanding may or may not possess a specific misconception at odds with the desired understanding. Therefore, misconceptions in the SICM model should not be defined as simply the absence of individual subabilities comprising the overall ability (i.e., a large collection of "inabilities"), else the set of misconceptions would be redundant with ability. (3) Retrofitting DCMs often yields classification of examinees into all-or-none attribute patterns. This may indicate that the test is measuring a unidimensional construct (Templin & Henson, 2006), but it is important to distinguish whether this is occurring because of cognitive or psychometric theory. The lack of multidimensionality may be a result of cognitive theory: The misconceptions may not actually exist as stable latent traits. Alternately, the multidimensionality may actually exist, but the assessment may not be capturing it because (a) the test lacks content validity in that options are not measuring what they purport to measure (Borsboom & Mellenbergh, 2007), or (b) the model lacks enough information to estimate the parameters (i.e., needs a larger sample or more items). (4) Simulation results presented provide guidelines for test and sampling conditions, such as sample size and number of items needed. Note large samples will be needed to estimate the parameters, and as the number of options per item increase, the sample size will also need to increase. (5) Existing concept inventories cited can provide insights for writing items to measure misconceptions with wrong answers and for conducting validity studies to verify whether options are eliciting misconceptions. (6) Pilot studies can statistically flag items that exhibit model-data misfit and need to be revised or removed. However, in initial stages of test development, unique statistical considerations are of interest for screening items. Given the purpose of the test, items that provide

information for measuring a single continuous trait and a set of multidimensional categorical traits are needed. No metric that marginalizes across ability and misconceptions was developed herein, but items should be expected to show multidimensionality and will need to be screened differently than they would with a unidimensional IRT model. Hopefully, further guidance on test development will be available in the future as results from collaborative efforts to design tests in this manner are disseminated.

Acknowledgements

This research was supported by the National Science Foundation grants DRL-0822064; SES-0750859; and SES-1030337. The opinions expressed are those of the authors and do not necessarily reflect the views of NSF.

References

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

Baker, F.B., & Kim, S.-H. (2004). *Item response theory: parameter estimation techniques* (2nd ed.). New York: Dekker.

Bell, A., Swan, M., & Taylor, G. (1981). Choice of operation in verbal problems with decimal numbers. *Educational Studies in Mathematics*, *12*, 399–420.

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

Bolt, D., & Lall, V.F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, *27*, 395–414.

Borovcnik, M., Bentz, H.J., & Kapadia, R. (1991). A probabilistic perspective. In R. Kapadia & M. Borovcnik (Eds.), *Chance encounters: probability in education* (pp. 27–33). Dordrecht: Kluwer.

Borsboom, D., & Mellenbergh, G. (2007). Test validity in cognitive diagnostic assessment. In J.P. Leighton & M.J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 85–115). Cambridge: Cambridge University Press.

Bradshaw, L., & Cohen, A. (2010). *Accuracy of multidimensional item response model parameters estimated under small sample sizes*. Paper presented at the annual American Educational Research Association conference in Denver, CO.

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

Cizek, G.J., Bunch, M.B., & Koons, H. (2004). Setting performance standards: contemporary methods. *Educational Measurement, Issues and Practice*, *23*(4), 31–50.

Confrey, J. (1990). A review of the research on student conceptions in mathematics, science, and programming. In C. Cazden (Ed.), *Review of research in education* (Vol. 16, pp. 3–56). Washington: American Educational Research Association.

de la Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*(3), 163–183.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: an analysis of fraction subtraction data. *Psychometrika*, *73*, 595–624.

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

Evans, D.L., Gray, G.L., Krause, S., Martin, J., Midkiff, C., Natoros, B.M., & Wage, K. (2003). Progress of concept inventory assessment tools. In *Proceedings of the 33rd ASEE/IEEE frontiers in education conference*. TT4G1-T4G8.

Garfield, J., & Chance, B. (2000). Assessment in statistics education: issues and challenges. *Mathematical Thinking and Learning*, *2*(1&2), 99–125.

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

Gibbons, R.D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

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

Hake, R. (1998). Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*, 64–74.

Halloun, I., Hake, R. R., Mosca, E. P., & Hestenes, D. (1995). *Force concept inventory (revised)* (unpublished instrument). Retrieved from http://modeling.asu.edu/R&E/Research.html.

Henson, R., & Templin, J. (2004). *Modifications of the Arpeggio algorithm to permit analysis of NAEP* (unpublished manuscript).

Henson, R., & Templin, J. (2005). *Hierarchical log-linear modeling of the joint skill distribution* (unpublished manuscript).

Henson, R., & Templin, J. (2008). *Implementation of standards setting for a geometry end-of-course exam*. Paper presented at the annual meeting of the American Educational Research Association, New York.

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. A., Templin, J., & Willse, J. T. (2013, under review). *Adapting diagnostic classification models to better fit the structure of existing large scale tests* (manuscript under review).

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

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). London: Cambridge University Press.

Jendraszek, P. (2008). *Misconceptions of probability among future mathematics teachers: a study of certain influences and notions that could interfere with understanding the often counterintuitive principles of probability*. Saarbrucken: VDM Verlag Dr. Müller.

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

Kunina-Habenicht, O., Rupp, A.A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59–81.

Lee, Y.-S., Park, Y.S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*(2), 144–177.

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

Luecht, R. (2013). Assessment engineering task model maps, task models, and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology*, *14*, 1–38.

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.

Mulford, D.R., & Robinson, W.R. (2002). An inventory for alternate conceptions among first-semester general chemistry students. *Journal of Chemical Education*, *79*(6), 739–751.

Muthén, L.K., & Muthén, B.O. (1998–2012). *Mplus user's guide* (6th ed.). Los Angeles: Muthén & Muthén.

National Council of Teachers of Mathematics (NCTM) (2001). *Principles and standards for school mathematics*. Reston: National Council of Teachers of Mathematics.

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: The National Academies Press.

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

Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system: a policy brief*. Washington: The Aspen Institute Education and Society Program. Available at www.aspeninstitute.org.

Rupp, A.A., & Templin, J. (2008). Unique characteristics of cognitive diagnosis models: a comprehensive review of the current state-of-the-art. *Measurement*, *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*, 010111.

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

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal and structural equation models*. Boca Raton: Chapman & Hall/CRC.

Smith, J.P., diSessa, A.A., & Roschelle, J. (1993). Misconceptions reconceived: a constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences*, *3*(2), 115–163.

Spiegelhalter, C.P., Best, N.G., Carlin, B.P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *64*(4), 583–640.

Tate, R.L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education*, *17*(2), 89–112.

Tatsuoka, K.K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics*, *10*(1), 55–73.

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: Erlbaum.

Templin, J., & Bradshaw, L. (2013). The comparative reliability of diagnostic model examinee estimates. *Journal of Classification*, *30*(2), 251–275.

Templin, J., & Bradshaw, L. (2013, 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., & Hoffman, L. (2013, in press). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*.

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

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). Princeton, NJ: Educational Testing Service.

Wainer, H., Vevea, J.L., Camacho, F., Reeve, B.B., III, Rosa, K., Nelson, L., et al. (2001). Augmented scores—"borrowing strength" to compute score based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Mahwah: Erlbaum.

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model to NAEP data* (RR-08-27). Princeton, NJ: Educational Testing Service.

Manuscript Received: 2 AUG 2012
Final Version Received: 2 JAN 2013