PSYCHOMETRIKA
2012
DOI: 10.1007/S11336-012-9296-Y

A MULTICOMPONENT LATENT TRAIT MODEL FOR DIAGNOSIS

SUSAN E. EMBRETSON

GEORGIA INSTITUTE OF TECHNOLOGY

XIANGDONG YANG

EAST CHINA NORMAL UNIVERSITY

This paper presents a noncompensatory latent trait model, the multicomponent latent trait model for diagnosis (MLTM-D), for cognitive diagnosis. In MLTM-D, a hierarchical relationship between components and attributes is specified to be applicable to permit diagnosis at two levels. MLTM-D is a generalization of the multicomponent latent trait model (MLTM; Whitely in Psychometrika, 45:479–494, 1980; Embretson in Psychometrika, 49:175–186, 1984) to be applicable to measures of broad traits, such as achievement tests, in which component structure varies *between* items. Conditions for model identification are described and marginal maximum likelihood estimators are presented, along with simulation data to demonstrate parameter recovery. To illustrate how MLTM-D can be used for diagnosis, an application to a large-scale test of mathematics achievement is presented. An advantage of MLTM-D for diagnosis is that it may be more applicable to large-scale assessments with more heterogeneous items than are latent class models.

Key words: item response theory, cognitive diagnosis, multidimensional measurement models.

1. Introduction

The development and implementation of models for cognitive diagnosis has been an active area of psychometric research. Although these models can be applied to the same item response data as traditional item response theory (IRT) models, the goal is different. That is, while traditional IRT models are applied to estimate one or more global indices of examinees' trait or competency levels, the goal in applying cognitive diagnostic models is to estimate examinees' possession of the attributes and skills that are represented in the items. When applied to achievement tests, for example, the traditional IRT models provide estimates of broad competency levels in a subject area while the diagnostic models can provide estimates of specific skill mastery by the examinees. Thus, the diagnostic models provide information that is relevant to remedial instruction on the skills represented in the test. When applied to psychological tests, given a plausible theory about the sources of item complexity, the diagnostic models can assess relative possession of attributes that are involved in responding to items.

All diagnostic models require substantive information about the skills and attributes involved in specific items. Thus, attribute structure matrices, often referred to as Q matrices, link the postulated attributes or skills to each item. The diagnostic models contain item parameters and classes or traits that relate item response probabilities to the attributes. Diagnostic models are most useful when items vary in which attributes are involved in item responses. Items on some existing tests, particularly tests that are designed to measure a narrow trait or competency level, may vary only in a few precisely defined attributes that impact item responses. However,

*Preparation of this paper was partially supported by *Institute of Educational Sciences* Grant R305A100234, Susan Embretson, PI.

Requests for reprints should be sent to Xiangdong Yang, Department of Educational Psychology, East China Normal University, Shanghai, China. E-mail: xdyang50@hotmail.com

Published online: 06 December 2012

other tests may be explicitly designed to represent a broad number of attributes. For example, standards-based achievement tests, such as the *National Assessment of Educational Progress* (NAEP) and most state accountability tests, contain diverse items that represent different aspects of competency as specified in a blueprint. Test blueprints for educational achievement tests, as well as for many licensure and certification tests, are typically hierarchically structured, with narrow attributes nested within broad categories. Some blueprints contain three or more levels in the hierarchy.

Several models for cognitive diagnosis have been developed (e.g., DiBello, Stout, & Roussos, 1995; Hartz, 2002; Haertel, 1989; Junker & Sijtsma, 2001; Templin & Henson, 2006; Henson, Templin, & Willse, 2009; von Davier, 2008) in which attribute mastery patterns are represented as latent classes. Some diagnostic classification models contain item parameters and latent variables, and the number of these parameters and variables increase directly with the number of attributes, K. Although these models have been applied successfully (e.g., Roussos, DiBello, Henson, Jang, & Templin, 2010) to tests that are characterized by relatively few attributes, estimation can be infeasible or impractical if the number of attributes is large. Further, differences between examinees may not be reliable when many latent variables are needed to characterize performance. Due to the wide differences between tests in item attribute structures, research effort has been devoted to developing criteria for the optimal design of attribute structures for diagnosis (DiBello et al., 1995; Tatsuoka, 1985).

The current paper presents a model for cognitive diagnosis that has a relatively small number of item parameters and latent variables. The latent variables are dimensions that represent underlying components involved in task solution, each of which can be related to attributes that impact difficulty. This model, the multicomponent latent trait model for diagnosis (MLTM-D), is a generalization of two earlier models, the multicomponent latent trait model (MLTM: Whitely, 1980) and the general component latent trait model (GLTM; Embretson, 1984). Unlike the earlier models, MLTM-D is formulated with a hierarchical structure to represent item differences in the number of underlying components. A brief overview of some currently available IRT models for diagnosis will precede the presentation of MLTM-D. Parameter recovery will be illustrated with a simulation example, and then the interpretability of the parameters for diagnosis will be illustrated with an application to a large-scale test of mathematical achievement.

2. Background

Psychometric models for diagnosis incorporate information for item i about the attributes k involved in a matrix, Q. For some diagnostic models, the entries in Q can be either binary or continuous, while for other models Q is restricted to binary variables. The meaning of the model parameters depends on the validity of Q for measuring substantively important attributes in items. The fit of a diagnostic model depends not only on the appropriateness of the model, but also on the reliability and validity of Q. Recent research (Rupp & Templin, 2008; DeCarlo, 2011) on diagnostic latent class models has shown that a misspecified Q can impact model parameter estimates, class sizes and misclassifications, as well as model fit. Empirical methods to detect misspecified vectors in Q are being developed for specific models (e.g., de la Torre & Chiu, 2009). However, the appropriateness of changing Q based on empirical findings depends on the conceptual plausibility underlying the attributes and skills, as well as on the rater reliability. For Q based on minimal prior empirical or theoretical foundations, changes to improve model fit may be more justifiable than for Q based on well established and empirically supported sets of attributes.

In the following sections, an overview of current diagnostic latent class models and latent trait models is presented. Then diagnosis is compared between models.

2.1. *Diagnostic Classification Models*

Diagnostic classification models (DCM) are latent class models with noncompensatory or compensatory effects of attributes. The original unified model (DiBello et al., 1995), the reparameterized unified model (RUM; Hartz, 2002) and DINA (Haertel, 1989; Junker & Sijtsma, 2001) are noncompensatory models because the probability of item solution is modeled as the product of parameters that represent attributes. However, some DCMs (e.g., Templin & Henson, 2006; Maris, 1999) specify compensatory effects for the attributes, where the parameters combine additively to increase the probability of item solution.

Two generalizations of diagnostic models have been developed that include most DCM because attributes may be specified to have either noncompensatory or compensatory effects, or both. Henson et al. (2009) developed the Log-Linear Cognitive Diagnosis Model (LCDM) for binary item response data. With LCDM, the probability of a correct response is given as follows:

$$P(X_{ij} = 1|\underline{\alpha}_j, \underline{q}_i) = 1/\big(1 + \exp\big(-1\big(\underline{\lambda}_i^T h(\underline{\alpha}_j, \underline{q}_i) - \pi_i\big)\big)\big), \tag{1}$$

where $\underline{\lambda}_i^T$ is a vector of parameters (weights) for item i and $h(\underline{\alpha}_j, \underline{q}_i)$ are linear combinations of attributes in the item and attribute possession and π_i defines the probability of a correct response for examinees who have not mastered any attributes. LCDM obtains generality from $h(\underline{\alpha}_j, \underline{q}_i)$, which can result in single terms, α_{jk}, and interactions, $\alpha_{jk}\alpha_{jk'}$, depending on $\underline{q}_i$. Weighted combinations of single attributes, $\lambda_1\alpha_{j1} + \lambda_2\alpha_{j2} + \cdots + \lambda_K\alpha_{jK}$, represent compensatory effects of attributes while product terms, such as $\lambda_1\alpha_{j1}\alpha_{j2}\alpha_{jK}$, represent noncompensatory effects of attributes. Estimation of LCDM parameters has been linked to algorithms in statistical software (Rupp, Templin, & Henson, 2010).

The General Diagnostic Model (GDM; von Davier & Yamamoto, 2004, 2007; von Davier 2005, 2008), in addition to including both compensatory and noncompensatory DCMs, is formulated to permit ordered polytomous item responses and a general function, $h(\alpha_{jk}, q_{ik})$, to link skills with attribute possession. That is, the probability of response x to item i is given as follows:

$$P(X_{ij} = x|\underline{\alpha}_j, \underline{q}_i) = 1/\big(1 + \exp\big(-1\big(\pi_{ix} + \underline{\lambda}_{ix}^T h(\underline{\alpha}_j, \underline{q}_i)\big)\big)\big), \tag{2}$$

where π_{ix} is an intercept and $\underline{\lambda}_{ix}^T$ is a vector of parameters (weights) for item i, in which the element λ_{ixk} specifies impact of attribute k for category x in item i. For example, if function h denotes product, and the item format is binary, then Equation (2) specifies the noncompensatory RUM. However, the GDM need not be constrained in this way; GDM can be specified to include a variety of diagnostic latent class models as well as models that include latent dimensions (von Davier, 2008) depending on $h(\bullet)$. Further, GDM also includes higher level interactions; thus a great variety of models can be specified. Since GDM is formulated to include both polytomous and binary item responses and more linking functions, GDM is a more general model than LCDM.

The number of item parameters varies between the latent classification models for diagnosis. DINA is a relatively simple model, for example, as the model contains two parameters per item. However, for other models, such as the noncompensatory RUM, the number of item parameters increases with the number of attributes required for each item. For other diagnostic models that can be specified with LCDM or GDM, the number of item parameters is potentially, but not necessarily, much larger if both non-compensatory and compensatory effects are included in the model. Conversely, both GDM (von Davier, 2008) and LCDM (Henson et al., 2009) also permit constraints such that more parsimonious versions of particular models can be specified.

2.2. *Diagnostic Latent Trait Models*

IRT models with latent traits, rather than latent classes, have potential for diagnosis under certain conditions. Stout (2007) noted that multidimensional item response theory (MIRT) models could be interpreted relative to diagnostic entities depending on how the dimensions were

combined. In compensatory MIRT models, the dimensions could represent alternative strategies for item solving because of their additive structure. In noncompensatory MIRT models, the dimensions could represent broad skills, all of which are required in item solution. The multidimensional IRT models are usually more parsimonious than DCMs and they are easier to be estimated.

Under certain conditions, unidimensional IRT models may have implications for diagnosis. The linear logistic test model (LLTM; Fischer, 1973) can be used to link examinee trait levels to the probability of solving items with specified attributes and skills if organized on a single underlying dimension. In LLTM, item difficulty is replaced with a model of item difficulty. The probability that person j passes item i, $P(X_{ij}=1)$ depends on q_{ik}, the score of item i on stimulus feature k in the cognitive complexity of items and η_k, the weight of stimulus feature k in item difficulty in LLTM, as follows:

$$P(X_{ij}=1|\theta_j,\underline{q}_i,\underline{\eta})=\frac{\exp(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k+\eta_0)}{1+\exp(\theta_j-\sum_{k=1}^{K}q_{ik}\eta_k+\eta_0)}. \tag{3}$$

In most applications, K is usually substantially smaller than I, the number of items. Models to relate item stimulus features to item discrimination and item difficulty also have been proposed, such as the 2PL-Constrained model (Embretson, 1999).

Compensatory MIRT models, as introduced by Bock and Aitkin (1981), were originally formulated as exploratory models, where the number and the nature of the latent dimensions were determined by fit to the data. A logistic form of a compensatory MIRT model is given as follows:

$$P(X_{ij}=1|\underline{\theta}_j,\underline{\lambda}_i,\pi_i)=1\Big/\left(1+\exp\left(-1.7\left(\pi_i+\sum_{m=1}^{M}\lambda_{im}\theta_{jm}\right)\right)\right), \tag{4}$$

where π_i is an intercept that represents item easiness, θ_{jm} is the level of attainment of examinee j on latent dimension m and λ_{im} is the weight of latent dimension m in the response to item i. A confirmatory version of Equation (4), in which weights of the dimensions, λ_{im}, contain both specified fixed values (i.e., to 0) and estimated values, could represent the impact of alternative strategies if the involvement of dimension m in item i is specified when $q_{im}=1$, and $q_{im}=0$, otherwise. Maximum likelihood estimation for confirmatory MIRT models based on the Bock–Aiken EM algorithm is relatively efficient, even for large numbers of items, if the number of dimensions is not too large (Cai, du Toit, & Thissen, 2012).

Noncompensatory MIRT models, such as the multicomponent latent trait model (MLTM; Whitely, 1980), and its generalization, the general component latent trait model (GLTM; Embretson, 1984) represent the probability of a correct item response as the product of the probabilities of correct responses to K component subtasks, as follows:

$$P(X_{ij}=1|\underline{\theta}_j,\underline{\beta}_i)=\prod_{m=1}^{M}\big(1/\big(1+\exp\big(-1.7(\theta_{jm}-\beta_{im})\big)\big)\big), \tag{5}$$

where θ_{jm} is the ability of examinee j on component m and β_{jm} is the difficulty of item i on component m. In MLTM, all components must be passed to solve an item. GLTM added a mapping of attributes to β_{im} by replacing it with $\sum_k \eta_{mk}q_{ikm}+\eta_{m0}$ to represent scored attributes. MLTM is a special case of GLTM when q_{ikm} consists of I dummy variables (and $\eta_{m0}=0$) that specify item difficulty, β_{im}, within each component m.

Both MLTM and GLTM require subtask responses to an item to identify the component model parameters in a joint modeling of item response patterns. Some attempts to estimate MLTM and GLTM without subtasks have been made, using data augmentation methods (Yang

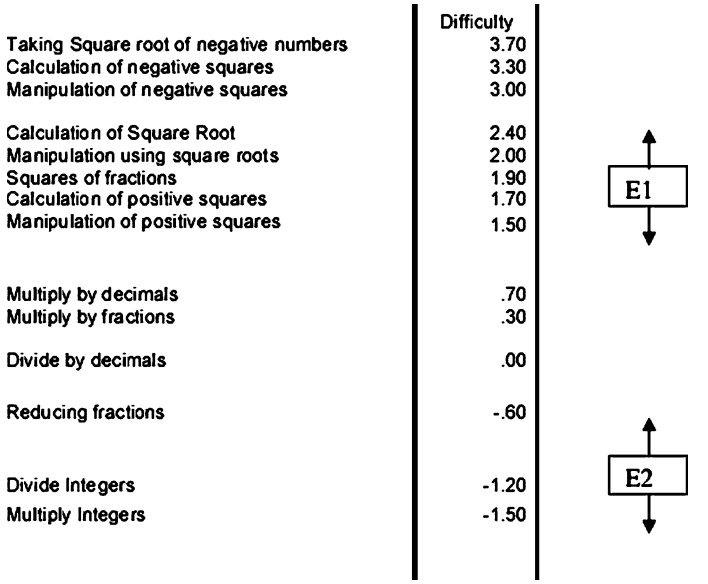


FIGURE 1.
Example of the relationship of attribute difficulty levels to item difficulty and trait level.

& Embretson, 2007), prediction models to identify one or more components (e.g., Embretson, 1995; Embretson & McCollam, 2000) and Bayesian estimates (Bolt & Lall, 2003). These methods have had limited success in recovering parameters; for example, Bolt and Lall (2003) found that MLTM yielded adequate recovery with two moderately correlated components ($\rho \leq 0.30$) when the number of items $I \geq 50$ and the sample size $N \geq 3{,}000$. However, parameter recovery was not adequate in the other conditions in their simulation study.

Diagnosis with Latent Classification and Latent Trait Models For all the latent class models, diagnosis for an examinee with response pattern $\mathbf{X}_j$ can be based on the joint distribution of the latent classes that represent unique patterns of attribute possession. Determining attribute or skill possession by finding the most likely response pattern, however, can be complicated for several reasons. First, the expected probability of any specific class for an examinee is often low due to the large number of latent classes. Even for a moderately complex test, with 10 categories in the blueprint, 1,024 latent classes (2^{10}) are needed to represent fully attribute possession patterns. Second, identifying a single latent class for an examinee is likely to be unreliable. Many latent classes, with somewhat different patterns of attribute possession, will have very similar likelihoods, especially as the number of skills or attributes increases. Third, interventions focused on increasing attribute possession or skill mastery are probably differentiated only at the attribute level. For example, it is easy to envision differentiated intervention for each of the 10 attributes, but not for 1,024 classes with differing patterns of attributes mastery. Thus, applications of diagnostic latent class models are usually focused on the marginal probability of attribute mastery across mastery patterns for each examinee j.

In the IRT latent trait models, θ_j obtains diagnostic meaning by its mapping to specified dimensions or attributes. For the unidimensional IRT models such as LLTM, attributes are mapped to the predicted difficulty levels of items. For example, Figure 1 illustrates how diagnosis may be obtained from a test of arithmetic skills, which are ordered by complexity, as supported by many theoretical perspectives (e.g., National Mathematics Advisory Panel, 2008). Item difficulty predicted by LLTM can locate the arithmetic skills on a latent continuum that is mapped to examinees by θ_j. Given cutline probability γ, mastery is diagnosed for θ_j by the skills represented

in items for which $P_{ij} > \gamma$. However, LLTM may have limited usefulness for diagnosis (see Stout, 2007; von Davier, 2009), especially if the attributes must be linear attributes, ordered on a dimension to permit diagnostic inferences.

Stochastically ordered skills on a latent dimension also may have diagnostic potential. For example, de la Torre and Douglas' (2004) higher order latent trait model that includes direct modeling of attribute possession from latent trait levels. That is, if the model involves a single latent trait, as most fully explicated by de la Torre and Douglas (2004), then attributes are ordered stochastically on a single continuum. The usefulness for diagnosis depends on reliable separation of the attributes on the continuum.

Compensatory MIRT models, when formulated in a confirmatory model defined by Q, also have diagnostic potential. For example, such specifications would be most successful for strategies, which involve combinations of attributes or skills, rather than for representing subgoals, each of which is required in problem solution. Thus, individual attributes would probably not define latent dimensions in most applications, nor is any θ_j mapped to more specific attributes in the MIRT model. Thus, the diagnostic information may be expected only at a higher-order level in a compensatory MIRT model.

In noncompensatory MIRT, such as MLTM, the components also are more likely to represent higher-order competencies than attributes. However, in MLTM, the noncompensatory relationship can represent subgoals, each of which is required in problem solution. For GLTM, mapping the component competencies to attributes is also possible, as in LLTM. Although MLTM and GLTM have some potential advantages as compared to MIRT models, as noted above, adequate estimation for both MLTM and GLTM usually requires subtask data, which are not typical in most testing situations. Thus, a generalization of MLTM-D that would not require subtask data is needed to be applicable to practical testing situations.

3. Specification of the Multicomponent Latent Trait Model for Diagnosis

In this section, MLTM-D will be presented and compared to earlier models. Similar to the general diagnostic models described above, LCDM and GDM, MLTM-D allows for both additive and multiplicative effects of attributes. However, unlike these models, MLTM-D has a hierarchical structure.

Formulation of the Model MLTM-D is a noncompensatory IRT model in which responses to latent components underlie item solving. It is assumed that the test is complex, with items varying in which components underlie the probability of a correct response. The probability that examinee j solves item i, $P(X_{ij} = 1)$, is given by MLTM-D as follows:

$$P_{ij} = P(X_{ij} = 1) = \prod_{m=1}^{M} P_{ijm}^{c_{im}} \tag{6}$$

and

$$P_{ijm} = P(X_{ijm} = 1|\theta_{jm}, \underline{q}_{im}, \underline{\eta}_m) = \frac{\exp(1.7(\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} + \eta_{m0}))}{1 + \exp(1.7(\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} + \eta_{m0}))}, \tag{7}$$

where θ_{jm} is the ability for subject j on component m, q_{imk} is the score for stimulus feature k in component m for item i, η_{mk} is the weight of feature k on component m, η_{m0} is the intercept for component m, and c_{im} is a binary variable for the involvement of component m in item i. If component m is not involved in item i ($c_{im} = 0$), then $P_{ijm} = 1$. GLTM and MLTM (Embretson, 1984) are special cases of Equation (6) when $c_{im} = 1$ for all items.

TABLE 1.
Three examples of component design structure in a C-matrix.

Blocks	C1	C2	C3
Set 1			
1	1	0	0
2	0	1	0
3	1	1	1
Set 2			
1	1	0	0
2	1	1	0
3	1	1	1
Set 3			
1	1	1	0
2	0	1	1
3	1	0	1

MLTM-D is applicable to items that may be meaningfully characterized as involving the responses to two or more underlying components, c_{im}, where item difficulty within each component in turn is governed by item attributes, q_{imk}, which are nested within components. Thus, two different types of structures are specified in MLTM-D. Assume that items are placed within blocks b according to their involvement of the same combination of components. The component structure matrix, $C_{b\times m}$, specifies the involvement of component m in items within block b. The number of possible item blocks involving the same combination of components is $2^M - 1$, to exclude the pattern of no components being involved in an item.

The attribute structure matrix, $Q_{i\times K_m}$, contains scores on the K_m attributes that impact item difficulty for the I_m items for which $c_{im} = 1$ on component m. Thus, item difficulty is modeled as a weighted combination of attributes that are relevant for the component. Additive models of the impact of task features on item difficulty have been widely used in studies of the response process aspect of validity (see Gorin, 2007, for examples).

Variables in $Q_{i\times K_m}$ may be binary or continuous. Consequently, covariates can also be included in $Q_{i\times K_m}$, to serve as control variables for other features that impact item difficulty on component m. A special case occurs when $K_m = I_m$ and $Q_{i\times K_m}$ consists of binary variables to represent the items on component m. In this case, the component attribute structure is a Rasch model, with η_{mk} estimating I_m item component difficulties. Other special cases include the possible addition of terms in $Q_{i\times K_m}$ to represent interactions between attributes or between attributes and covariates.

Table 1 presents examples of component structures in which item blocks involve varying combinations of three underlying components. In Set 1, the items in the first block involve only the first component while items in the third block involve all three components. In Set 2, Block 1 is a subset of Block 2 and Block 2 is a subset of Block 3. In Set 3, yet another structure is specified. Attributes, in contrast, vary within components. That is, item difficulty at the component level is modeled by a weighted combination of attributes.

Examinee estimates of mastery, θ_{jm}, are obtained at the component level so that patterns of component competencies can be obtained for each examinee, j. In turn, each component θ_{jm} may also be linked to the more specific skills or attributes that determine component item difficulty level. If attributes are linearly ordered, as in the unidimensional LLTM shown in Figure 1, each examinee may be located on the common measurement scale with attributes by exceeding a specified mastery level γ_m for each component. Or, if combinations of attributes are stochastically ordered, as in de la Torre and Douglas (2004), predictions about mastery patterns are feasible given reliable separation of attributes on the dimension. Thus, linking the component

trait scores to the attributes permits additional interpretations of the specific skills that examinees have mastered.

Distinguishing Between Attributes and Components To apply MLTM-D, a theoretically and empirically plausible theory of the underlying components and the attributes or skills must be available. That is, scores to operationalize the theory must be available for both the involvement of each component m in each item i, c_{im}, and the values of the attributes, q_{imk}, that impact the difficulty of component m in item i. The component design must vary between items to obtain model identification, which is elaborated further below.

Components are distinguished from attributes by the nature of the theory underlying the test. For tests with heterogeneous item content, such as broad tests of achievement or certification tests, theories with hierarchical structures are particularly appropriate. Such tests typically contain large numbers of items (e.g., 60 to 200 items) and repetition of more narrowly defined attributes or skills. For example, the blueprints for most state achievement tests follow Webb's (1999) criterion of a minimum of six items per reportable score unit.

Different types of theories with a hierarchical structure can be specified in MLTM-D. For example, nested categories, such as designated in a blueprint of a standards-based achievement tests (e.g., NAEP and most state educational achievement tests) is one type of theory that can be specified. Consider the three items on Table 2 that are variants of items on a test used in a Midwestern state to assess overall 8th grade mathematics achievement. The test blueprint contains four standards, which define the areas in which competency is tested (Number/Computation, Algebra, Geometry, and Data). Within each standard, more narrow skills are specified as indicators, each containing from 2 to 7 items. For example, an indicator of the Algebra standard in the 8th grade test is "*Translates between the numerical, tabular, graphical and symbolic representation of linear relationships with integer coefficients and constants*."

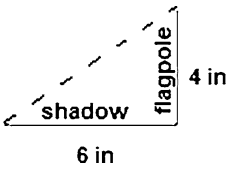
Since complex items may involve more than one standard or indicator, items such as shown on Table 2 were scored by a panel of mathematics and curriculum experts for the involvement of multiple standards and indicators (Poggio, 2011). Up to 56 % of the items on a scored test form involved multiple standards or indicators. Item 1 on Table 2 involves two standards, Algebra and Geometry and one specific indicator within each of the standards. Item 2 and Item 3 involve only one standard, Algebra, but different indicators within the standard. For MLTM-D, c_{im} can represent the standard in item i and q_{imk} can represent the indicators as binary scores within the standard involved in the I_m items. Alternatively, the items can be scored for global variables, such as complexity level, which would replace the indicator scoring of q_{imk}.

Also, it is possible to apply MLTM-D when plausible attributes are not available for one or more components by specifying $Q_{i\times K_m}$ to define a Rasch model within the component, as described above.

For another example, cognitive process analyses can be applied to the same items to estimate the impact of processing stages. In this case, as in cognitive experiments, the difficulty of each processing component is postulated to be impacted by different attributes. The mathematics test described briefly above also was scored for five processing stages (Translation, Integration, Solution Planning, Solution Execution, and Decision Processing), based on an empirically-supported theory of mathematical problem solving (Embretson & Daniel, 2008; Daniel & Embretson, 2010). In this case, $c_{im} = 1$ if processing component m is involved in an item. Item 1 on Table 2 is complex and involves all processing stages except Decision Processing (i.e., the correct answer can be obtained without examining the response alternatives). Item 2, in contrast, involves Encoding, Solution Planning and Solution Execution, but neither Integration (i.e., the equation is given) nor Decision Processing. Item 3 involves Encoding, Integration and Decision Processing, but not Solution Planning or Solution Execution (i.e., the examinee must identify an equation, but need not solve it). Within components, the K_m attributes that are postulated to

TABLE 2.
Three examples of items from a state mathematics achievement test.

Item 1
The scale drawing below shows a flagpole and its shadow.



This drawing shows the height of the flagpole to be 4 inches (in). The actual height of the flagpole is 24 feet (ft). What is the length, in feet, of the actual shadow?
(A) 16 ft
(B) 26 ft
(C) 24 ft
(D) X 36 ft

Item 2
What is the value of x in the equation $3 + x/5 = 8$?
(A) X $x = 25$
(B) $x = 6$
(C) $x = 27$
(D) $x = 10$

Item 3
Joni is going to run a 5,000 meter race that is split into 2 unequal segments. The second segment of the race is 2,000 meters shorter than the first segment of the race. Which equation could be used to find the length of the first segment (m)?
(A) $m + m + 2{,}000 = 5{,}000$
(B) $m + m = 5{,}000$
(C) $m - 2{,}000 = 5{,}000$
(D) X $m + (m - 2{,}000) = 5{,}000$

impact item difficulty are scored as q_{imk} for item i when $c_{im} = 1$. For example, computational burden and procedural complexity are scored only if Solution Execution is involved in an item. The impact of task features on item difficulty is additive, similar to the general linear model approaches used to test hypotheses about task complexity in cognitive psychology. The difficulties of the other components are impacted by different variables.

Number of Parameter Estimates The number of parameters to be estimated in MLTM-D depends on how the model is applied. The greatest number of parameter estimates is obtained when the full component MLTM-D with no attribute structure is applied. In this case, the number of parameters depends on the number of components specified in each item. For the structure shown as Set 1 in Table 1, for example, a full component model with no attributes would estimate three parameters each for items in Block 3 and one parameter each for items in Block 1 or Block 2.

When parameters are estimated for K_m attributes in a component, the number of parameters is reduced if $K_m < I_m$. In the extreme case of one attribute per component, the total number of item parameters is the twice the number of components, consisting of an intercept and a weight for each component. The model can also be applied with greater numbers of attributes within

each component. Thus, the hierarchical structure of MLTM-D can accommodate a large number of attributes with relatively few item parameters. Further, combinations of components with or without attribute structures also may be applied.

4. Model Identification and Estimation

Model Identification Model identification for MLTM-D depends on (1) fixing the measurement scale within each component, (2) the structure of the components between blocks of items $C_{b\times m}$ and (3) the structure of attributes in $Q_{i\times K_m}$ within each component. The measurement scale may be identified using standard IRT procedures, such as fixing either means and variances of the person distributions (e.g., $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \Sigma)$), where the diagonal of Σ equals 1, or by fixing the means of the item difficulties and item discriminations (e.g., $M_b = 0$ and $M_a = 1$).

Model identification at the component level depends on the component block structure where each block b contains items with the same combination of components as shown on Table 1. MLTM-D can be expressed at the component level as modeling items within blocks, where the log probability of item i within block b is the following:

$$\ln P^*_{i(b)j} = \sum_m c_{i(b)m} \ln P_{ijm}, \tag{8}$$

where $c_{i(b)m}$ is a binary variable for the involvement of component m for an item within block b and $\ln P_{ijm}$ is the log probability that component m on item i is responded to correctly by examinee j.

To explicate the relationship to model identification, for convenience assume that items differ in the involvement of components, as given by $c_{i(b)m}$, but not in difficulty within components or blocks. Define $P^B_{b\times j}$ as a matrix with elements $\ln P(X_{jb} = 1)$, the log probabilities that examinee j responds correctly to an item in block b. Let $P^C_{m\times j}$ contain elements $\ln P(X_{jm} = 1)$, the log probability that examinee j responds correctly to component m when it is contained in item i and $C_{b\times m}$ is the component block structure matrix as defined above. Then it can be shown that

$$P^B_{b\times j} = C_{b\times m} P^C_{m\times j}. \tag{9}$$

Suppose that $B = M$, such that the number of blocks equals the number of components. Then, if $C_{b\times m}$ is of full rank,

$$P^C_{m\times j} = C^{-1}_{b\times m} P^B_{b\times j}, \tag{10}$$

where the log probabilities of the components may be obtained as linear combinations of the log probabilities in the blocks, a just determined model. If $B > M$, then there must be at least one submatrix of $C_{b\times m}$, such that $C^*_{b\times m}$ is of full rank and, in this case, the remainder of $C_{b\times m}$ overdetermines components.

Within components, model identification depends on $Q_{i\times K_m}$ consisting of K_m independent vectors. If I_m is the number of items involving component, and if $I_m = K_m$, then $Q_{i\times K_m}$ must be of full rank.

Estimation of Item Parameters Based on Equation (6), the probability of getting a particular response pattern $\mathbf{x}_j$, $\mathbf{x}_j = \{X_{1j}, X_{2j}, \ldots, X_{nj}\}$, for n items, $i = 1, 2, \ldots, n$, is given as

$$P(\mathbf{x}_j) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(\mathbf{x}_j|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} P(\mathbf{x}_j|\boldsymbol{\theta}) g(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{11}$$

where $P(\mathbf{x}_j|\boldsymbol{\theta}) = \prod_{i=1}^{n} P_{ij}^{X_{ij}}(1-P_{ij})^{1-X_{ij}}$ is the probability of getting $\mathbf{x}_j$ given $\boldsymbol{\theta}$, and $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_M\}$ is the vector of ability parameters of a randomly drawn examinee on the M item components involved in the set of items and $g(\boldsymbol{\theta})$ is the probability density function of $\boldsymbol{\theta}$. Normally, it assumes that $\boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix with rank M.

After observing the response matrix $\mathbf{X}$ for a sample of examinees N, $j = 1, 2, \ldots, N$, the likelihood equation parameter η_{mk} to be solved under the MMLE-EM algorithm is given as

$$\frac{\partial l}{\partial \eta_{mk}} = 0 = \int_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \left(\frac{\tilde{r}_i - \tilde{N} P_{ij}}{P_{ij}(1-P_{ij})} \frac{\partial P_{ij}}{\partial \eta_{mk}} \right) \right] g(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

$$= \int_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \left(\frac{\tilde{r}_i - \tilde{N} P_{ij}}{P_{ij}(1-P_{ij})} c_{im} \prod_{h=1, h\neq m}^{M} P_{ijh}^{c_{ih}}(-1.7 q_{imk}) P_{ijm}(1-P_{ijm}) \right) \right] g(\boldsymbol{\theta})\, d\boldsymbol{\theta} \quad (12)$$

for $\tilde{N} = [\frac{\sum_{j=1}^{N} L(\mathbf{x}_j|\boldsymbol{\theta})}{P(\mathbf{x}_j)}]$, $\tilde{r} = \frac{\sum_{j=1}^{N} \mathbf{x}_{ij} L(\mathbf{x}_j|\boldsymbol{\theta})}{P(\mathbf{x}_j)}$.

In implementation, the multiple integrals in Equation (12) can be approximated numerically by applying the M-fold Hermite–Gauss quadrature as follows:

$$\frac{\partial l}{\partial \eta_{mk}} \approx \sum_{z_m=1}^{Z} \cdots \sum_{z_2=1}^{Z} \sum_{z_1=1}^{Z} \times \left[\sum_{i=1}^{n} \left(\frac{\tilde{r}_{i \bullet z_1 z_2 \cdots z_M} - \tilde{N}_{z_1 z_2 \cdots z_M} P_{ij}}{P_{ij}(1-P_{ij})} \frac{\partial P_{ij}}{\partial \eta_{mk}} \right) \right] W(X_{z_1}) W(X_{z_2}) \cdots W(X_{z_M}) \quad (13)$$

where X_{z_m} is a quadrature point on mth dimension in the M-dimensional space, and $W(X_{z_m})$ is the associated weight. The parameter $\tilde{r}_{i \bullet z_1 z_2 \cdots z_M}$ represents the number of examinees with ability vector $\mathbf{X}$, $\mathbf{X} = \{X_{z1}, X_{z2}, \ldots, X_{zM}\}$ who are expected to respond correctly to the item given the sample data. It is an entry in a Q^M dimensional array in which each cell corresponds to an M-tuple of quadrature points for a given item. $\tilde{N}_{z_1 z_2 \cdots z_M}$ is the expected number of examinees with ability vector $\mathbf{X}$ at the same entry and is normalized to the sample size.

5. Demonstration of Parameter Recovery

Parameter recovery can be demonstrated by a simulation study. Data were generated under MLTM-D using a SAS macro written specifically for the current study. The macro is capable of generating data under both a full and restricted MLTM-D based on user-specified $C_{b\times m}$ and $Q_{i\times K_m}$.

In the current simulation, items were assumed to contain up to three latent components. For all replications, component abilities for 1,000 examinees were generated from a multivariate normal distribution ($\boldsymbol{\theta} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$). Two conditions were specified for Σ to represent uncorrelated versus correlated dimensions. That is, for uncorrelated dimensions, $\Sigma = \mathbf{I}$; and for correlated dimensions, diag $\Sigma = \mathbf{1}$ with moderate correlations between dimensions m and m' specified ($\sigma_{\{m,m'\}} = \{0.3, 0.4, 0.5\}$).

The generated test consisted of 60 items with the component structure $C_{b\times m}$ shown in Table 3. Thus, six different blocks of within-item component structures were specified and each type contains 10 items. Table 3 shows that Component 3 is always embedded with another component.

The item generation parameters were set to yield component item difficulties $\sim N(-1, 1)$ as obtained from the attribute design matrix. That is, component item difficulties should be easy to

TABLE 3.
Item-component configuration for simulated data.

Type	C1	C2	C3	Number of items
1	1	0	0	10
2	0	1	0	10
3	1	1	0	10
4	1	0	1	10
5	0	1	1	10
6	1	1	1	10

moderate when multiple components are involved in items so that the probability of a correct item response is sufficiently high for the majority of the examinees in a given sample. For illustration, suppose that an item has three components and each component difficulty is 0, then, for an examinee with component abilities of 0 for each of the three components, the probability of this examinee answering the item correctly, given the MLTM-D, is $0.5^*0.5^*0.5 = 0.125$, which would result in insufficient information in the response data to recover the model parameters.

Data were generated for the full hierarchical structure of MLTM-D with two attribute design factors specified for each item component. Let q_{im1} and q_{im2} be the value of the two design factors on the component m in item i; then the item component difficulty is given as

$$\beta_{im} = \eta_{m0} + \eta_{m1} q_{im1} + \eta_{m2} q_{im2} + \varepsilon_{im}, \tag{14}$$

where the q_{imk} are continuous variables from a normal distribution ($\sim N(0, 1)$ for $k = 1, 2$), η_{m1} and η_{m2} are weights, and η_{m0} is an intercept for component m. For the conditions with prediction error, $\varepsilon_{im} \sim N(0, \sigma_m^2)$. For convenience, assume that $\text{corr}(q_{im1}, q_{im2}) = 0$ for all m, and that for two different components, m and m', $\text{corr}(q_{imk}, q_{im'k}) = 0$. Thus, for each component, $\sigma_m^2 = 1 - \eta_{m1}^2 - \eta_{m2}^2$. To ensure that the predicted item component difficulty β_{im} was easy or moderately difficult, $\eta_{m0} = -1$ for all m components. Two levels of prediction error were specified for the simulations: (1) fixed item effects with no error, where the variance of ε_{im} was specified as 0, and (2) random item effects, with the error variance for ε_{im} specified as 0.10 but the value of ε_{im} for items varied over replications. For the condition without prediction error, the two design factors q_{im1} and q_{im2} perfectly predict item component difficulty.

Data were generated to illustrate parameter recovery for the two conditions of prediction error, using both correlated and uncorrelated dimensions. Item response data were generated for each replication by predicting the item probabilities from MLTM-D using true parameter values with randomly selected q_{imk} and ε_{im} and then comparing the probabilities to a uniform random variable with distribution from 0 to 1. A total of 80 replications were generated, with 20 replications specified for each combination of prediction error and dimension correlation. Each replication contained 1,000 observations.

Estimates for the model parameters were obtained by specifying the MLTM-D as a nonlinear mixed model. SAS macros were written to implement the component design. The measurement scale was fixed by specifying the persons as random effects arising from a standard multivariate normal distribution ($\boldsymbol{\theta} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$), where Σ was specified as diag $\Sigma = 1$ and $\sigma_{m,m'}$ freely estimated for the m components. MML estimates were obtained using adaptive quadrature points.

Overall root mean square error (RMSE) was computed for the set of parameters for each condition. RMSE differed little from the mean estimated parameter errors for data generated without prediction error. Summed across parameters, RMSE was 0.0316 and 0.0332 for data generated, respectively, with uncorrelated and correlated dimensions. As expected, RMSE was higher for data generated with prediction error. RMSE was 0.0793 and 0.0748 for data generated, respectively, with uncorrelated and correlated dimensions.

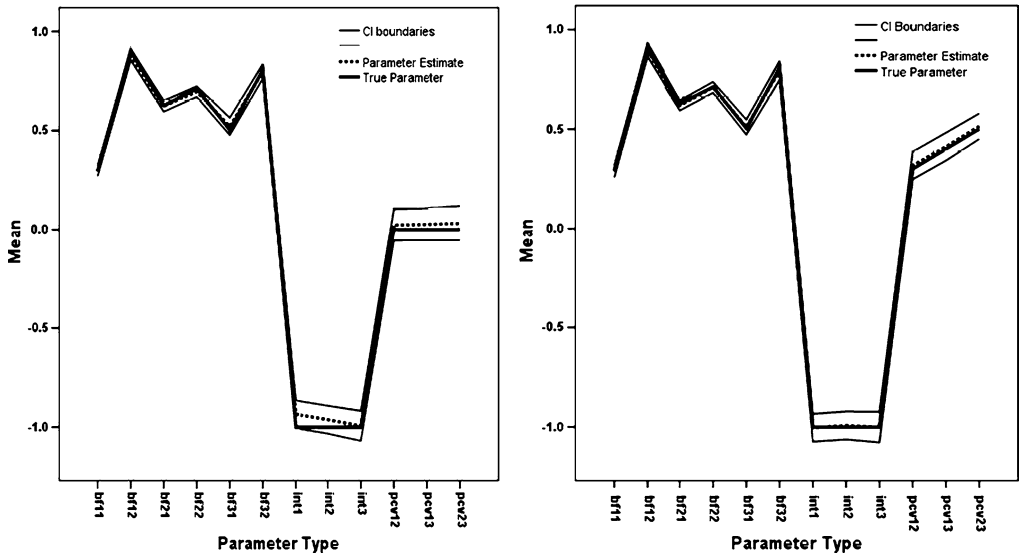


FIGURE 2.
Simulation means and confidence intervals for MLTM-D attribute weights for three components.

Figure 2 shows the mean parameter estimates and confidence intervals for each parameter for uncorrelated and correlated dimensions, on the left and right sides respectively, collapsed over replications. It can be seen that the true value for each parameter estimate is contained within the 95 % confidence interval for all parameters. Thus, the simulation results indicated that parameters can be successfully recovered in the conditions that were studied with 1,000 simulated examinees. Since MLTM-D is multidimensional, parameter estimation with smaller sample sizes is not recommended without supporting data.

6. Application to Mathematical Achievement Data

In this section, an application of MLTM-D to a high-stakes test of mathematics achievement is presented. The test had been administered to all 8th grade students for state accountability with respect to four nationally-aligned standards in mathematics. The test consisted of 86 multiple choice items that had been scaled with a three-parameter logistic IRT model. Examinee competency levels are evaluated with respect to proficiency categories which are set by experts based on a substantive analysis of the items. For the items on this particular test form, the proficiency threshold corresponded to a trait level associated with 67.4 % correct on the test as a whole. The overall test results are used to evaluate teachers and schools for instructional quality, as well as to evaluate the preparation of students for the next grade level. Diagnosing the sources of poor performance is especially important for students who are below the proficiency threshold so that appropriate remedial instruction may be given. However, the unidimensional IRT model that is applied to the test assesses only global competency levels and does not assess more specific sources of poor performance. MLTM-D was applied to this achievement test to examine its potential to provide diagnostic information that would be relevant to intervention.

6.1. Method

Testing Materials The test contained mathematics items that were constructed according to a blueprint to represent four standards areas: (1) Number/Computation, (2) Algebra, (3) Geome-

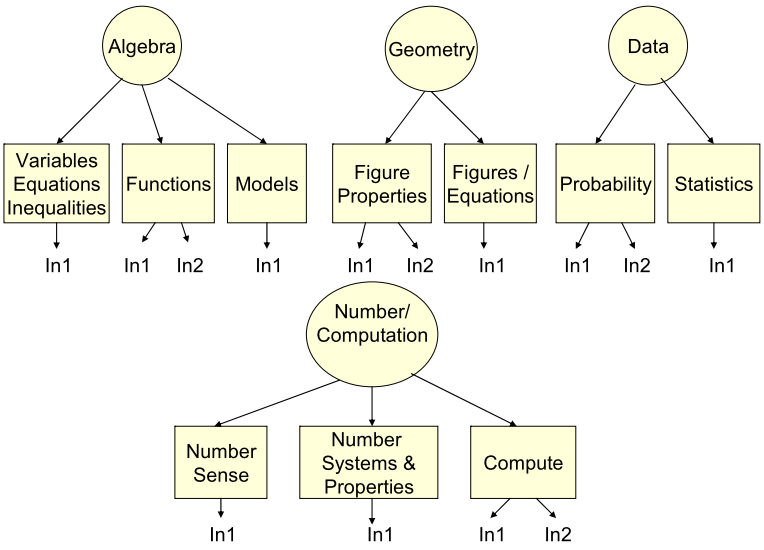


FIGURE 3.
Hierarchical structure of standards and indicators on a high-stakes mathematical achievement test.

try, and (4) Data Analysis, with the hierarchical structure shown on Figure 3. Nested within each standard are benchmarks (e.g., Number Sense with Number/Computation) and nested within benchmarks are indicators (e.g., In1) which define specific skills within the category. The 86 items on the test had been constructed to represent 25 separate indicators; thus each indicator is represented by several items ($M = 3.44$). Item content varied from equations with few words to complex word problems.

Procedure The test was administered online to all students in the state at the end of the school year. A random sample of 3,000 students was obtained for this application of MLTM-D.

Scoring Component and Attribute Structures Although each item had been constructed to fulfill a single indicator in the test blueprint, the complex word problems often involved two or more indicators from different standards areas. Thus, the 86 items were scored by a mathematician for the involvement of each indicator (Poggio, 2011). Items involving multiple indicators predominated, with 56 % of the items involving two or more indicators. The component structure matrix was obtained by scoring standards as components; each item was scored as involving a component, $c_{im} = 1$, if it involved one or more indicators of a standard and $c_{im} = 0$ otherwise. It was found that 36 % of the items involved multiple components and that nine distinct component patterns were identified. Table 4 presents the resulting component structure matrix, $C_{b\times m}$ along with the number of items for each pattern block. Also shown on Table 4 are the number of items per component, $24 \leq c_{\bullet m} \leq 37$. Thus, although the components differ somewhat in the number of items in which they are involved, the differences were not substantial.

Attributes for items were scored within components to obtain $Q_{i\times K_m}$. For the current application, attributes were scored only for the Number/Computation component. The indicators were scored as attributes, along with three variables to reflect the problem context; reading grade level, visualization, and number of words.

6.2. Results

The analyses with MLTM-D and comparison models were preceded by a classical test theory item analysis. The mean p-value was moderately high ($M = 0.70$), as typical on standards-based

TABLE 4.
Component structure matrix and descriptive statistics for MLTM-D component item parameters for the mathematical achievement test.

Block	Pattern				Number of items
	C1	C2	C3	C4	
1	1	0	0	0	11
2	1	0	0	1	1
3	1	1	0	0	17
4	1	1	0	1	5
5	1	1	1	1	7
6	0	0	0	1	11
7	0	0	1	0	17
8	0	1	0	0	16
9	0	1	1	0	1
# of items	37	46	25	24	

achievement tests that are targeted to distinguish proficient from nonproficient students. The biserial correlations were generally high ($M = 0.53$), indicating high item discrimination, although two items had small biserial correlations ($r_{\text{bis}} < 0.20$), indicating poor fit for these items.

Parameter Estimates and Fit The MLTM-D item parameters were estimated with MML. A macro was developed to use with the SAS nonlinear mixed modeling procedure, where items were specified as fixed variables. Examinees were specified as random variables arising from a standardized multivariate normal distribution ($\boldsymbol{\theta} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$). Integration across the random variables was achieved using Gauss–Hermite quadrature. MLTM-D was specified in the normal metric as in Equation (7). For the first set of analyses, Σ was specified as an identity matrix for simplicity in implementing multidimensional quadrature; hence, a constant slope was estimated for each component to reflect differences in variances and discrimination. Component item difficulty parameter estimates were subsequently rescaled to a slope of 1.0.

Three different MLTM-D models were specified and estimated: (1) the full component model with estimated item component difficulties but no attribute structure ($-2\ln L = 250{,}637$; $AIC = 250{,}917$), (2) a restricted hierarchical MLTM-D with components and attributes ($-2\ln L = 251{,}594$; $AIC = 251{,}818$) and (3) a null model, in which all items are equally difficult within the components ($-2\ln L = 263{,}046$; $AIC = 263{,}060$).

For the second set of analyses, the same models were also estimated with Σ specified as free to allow estimation with correlated latent dimensions, except diag $\Sigma = \mathbf{1}$ to set the scale of measurement. The following results were obtained: (1) the full component model ($-2\ln L = 242{,}945$; $AIC = 243{,}237$), (2) the restricted hierarchical MLTM-D ($-2\ln L = 243{,}826$; $AIC = 244{,}062$) and (3) the null model ($-2\ln L = 255{,}439$; $AIC = 255, 641$). These results indicate that MLTM-D estimated with correlated dimensions fits the mathematical achievement substantially better than MLTM-D estimated with uncorrelated dimensions.

Table 5 presents descriptive statistics for the item parameters from the full component model estimated with correlated dimensions. As expected for the noncompensatory MLTM-D producing an average item p-value of 0.70, the mean component item difficulty for each component is relatively low. Nonetheless, the standard deviations indicate substantial variability in difficulty among the items within each component. The estimates for the correlations between the dimensions ranged from 0.620 to 0.732, indicating moderate correlations as typically obtained for cognitive or achievement dimensions.

The variant models were compared to examine the relative strength of the more restricted attribute model, as compared to the full component model. Using the null model as a baseline model, an incremental fit index (Embretson, 1999), Δ, was computed as follows:

TABLE 5.
Means and standard deviations of examinee component competency estimates for the mathematical achievement test.

Component	MML parameter estimates					EAP	
	Item parameters		Dimension correlations			Person estimates	
	Mean	SD	C1	C2	C3	Mean	SD
C1	−2.094	1.594	1.000			0.059	0.859
C2	−0.981	0.769	0.620	1.000		0.056	0.720
C3	−1.569	1.178	0.707	0.732	1.000	0.029	1.022
C4	−1.977	1.588	0.642	0.627	0.633	0.038	0.872

$$\Delta = (\ln L_{\text{null}} - \ln L_{\text{restr}})/(\ln L_{\text{null}} - \ln L_{\text{full}}). \tag{15}$$

The resulting values were 0.922 and 0.930, respectively, for MLTM-D estimates with diagonal and full Σ. These results indicated good fit of MLTM-D with a restricted component, since Δ is similar in magnitude to a squared multiple correlation.

The *expected a posteriori* (EAP) method was used to estimate the component competency levels for examinees. Macros for EAP estimates were developed for both SAS and SPSS. Table 5 presents descriptive statistics on the component competency estimates that were obtained using the rescaled item parameter estimates. It can be seen that the means are close to zero for each component, but that the variances differ somewhat between components.

Item fit was assessed by a goodness of fit test in which expected frequencies of correct item responses were compared to observed frequencies. Predicted item probabilities were obtained from the noncompensatory MLTM-D for each examinee on each item using the parameter estimates from the full component model estimated with correlated dimensions. Then, for each item, the range of predicted probabilities was categorized into fixed intervals and examinees were classified into the intervals to represent examinees with similar expectations for item success. Examinees in categories with fewer than 10 observations were collapsed into the next higher category, which resulted in a median of 8 categories per item. The predicted and observed frequency of examinees with correct responses was obtained for each category within each item. Similar to the Mislevy and Bock (1990) procedure, item goodness of fit, G^2, was computed by comparing the observed to expected frequency of correct responses across categories within items. Then G^2 was compared to the χ^2 distribution with *df* equal to the number of categories. Significance was assessed at $p < 0.01$ to minimize error from multiple comparisons.

For 84 items, fit was adequate ($p > 0.01$). For the remaining two items, the probability of G^2 was less than 0.02. These items had poor item biserial correlations ($r < 0.20$) in the CTT analyses. The observed proportions correct in the categories was correlated with the MLTM-D expectation across all categories and items. A high correlation ($r = 0.927$) was obtained. Thus, overall model fit appeared to be adequate.

Comparison with Other Models To provide a comparison with other diagnostic models, two DCMs were estimated on the mathematical achievement data. First, the DINA model was estimated with four attributes, using the same pattern for the standards that defined the four traits in MLTM-D. Bayesian estimates for DINA were obtained as a special case of LCDM using a MCMC procedure (Henson et al., 2009). Estimates of the slip parameter ($M = 0.179$, $SD = 0.049$) and the guessing parameter ($M = 0.535$, $SD = 0.211$) were obtained for each item, for a total of 172 parameters. Overall fit indices for DINA, based on the central tendency of the likelihood distributions, can be computed from MCMC results. These values ($-2 \ln L = 246{,}781$; $AIC = 247{,}125$) indicated that DINA did not fit as well as the full MLTM-D with correlated dimensions. However, this comparison should be interpreted cautiously due to the different estimation methods.

A full LCDM with both compensatory effects and noncompensatory effects was also fit to the data, for a total of 337 parameters. The parameters include 86 intercepts plus 256 weights to represent compensatory effects and noncompensatory effects, as designated by the component structure. LCDM includes DINA, which would be represented as the highest order interaction effect and an intercept for each item. For the full LCDM, 123 of 252 effect parameters were significant, supporting a relatively simple parameter structure for many, but not all, items. Overall indices based on the likelihood ($-2\ln L = 244{,}732$; $AIC = 245{,}407$) indicated that LCDM fit better than DINA. However, a comparison of the AIC between models indicated that neither LCDM nor DINA fit better than the full MLTM-D with correlated person dimensions ($AIC = 243{,}237$). Again, this comparison should be interpreted cautiously due to the different estimation methods.

Finally, the impact of guessing in the mathematical achievement data was examined by estimating the 3PL model for the 86 items ($-2\ln L = 249{,}786$; $AIC = 250{,}302$) with 258 parameters. The mean estimated lower asymptote ($M = 0.234$) was close to the expected value for guessing with four response options. As compared to MLTM-D with correlated dimensions and 146 parameters, the AIC for the 3PL model was substantially larger, indicating worse fit.

Diagnosis MLTM-D permits diagnosis at both the component and the attribute level. For the mathematical achievement test, students at the cutline are particularly important to diagnose for possible interventions. The component competency levels of 64 students at the cutline of 67.4 % correct were examined for diagnostic potential. Histograms of the four component competencies were prepared. The standard deviations for component competency ranged from 0.376 to 0.532 within this group, indicating substantial variability for all four components even though the examinees had the same overall level of performance. Further, the distributions of competency levels on Component 1, Component 2, and Component 3 appeared as bimodal distributions, while the distribution of competency levels on Component 4 was positively skewed. Thus, the distributions are consistent with different patterns of component competency for individuals at the same overall level of performance.

Figure 4 presents the component competency scores for two examinees, along with 67 % confidence intervals. It can be seen that component patterns vary substantially and given a cutline δ_k for component proficiency, differential recommendations for intervention could be made. For consistency, the same δ_1 was set for all four components. For competencies at $\delta_k = -0.80$ for each component, the expected proportion of the proportion of items solved, weighted by block size, equaled the proportion of items solved at the overall competency cutline (i.e., 0.674). Figure 4 shows that the examinees differ in the components for which their scores are not above the proficiency level. Examinee ID2025 falls below the cutline on C2 and C3, while examinee ID1814 falls below the cutline on C4 and marginally on C3. Other examinees within this group (not shown) have different patterns of component competency. Hence, although the examinees have the same overall competency level at just below the full test cutline, differential recommendations for remedial instruction would be made for these examinees based on the component patterns.

The component scores were linked to more specific attribute mastery for the Number/Computation component through the attributes that underlie its difficulty, q_{imk}. Figure 5 presents the indicators of Number/Computation with their estimated difficulty on the dimension. The difficulty level of each attribute shown in Figure 5 is determined by the value of q_{imk} and the weight, η_{mk} as in Equation (14). The lowest level is knowledge of the order of operations while the highest level is knowledge of different number systems. The probabilities of attribute mastery for examinees are given by linking their component score estimates to the common measurement scale. Shown on the left of Figure 5 are the competency levels of three examinees. To illustrate interpretations of more specific skill mastery, consider the examinee that scored below

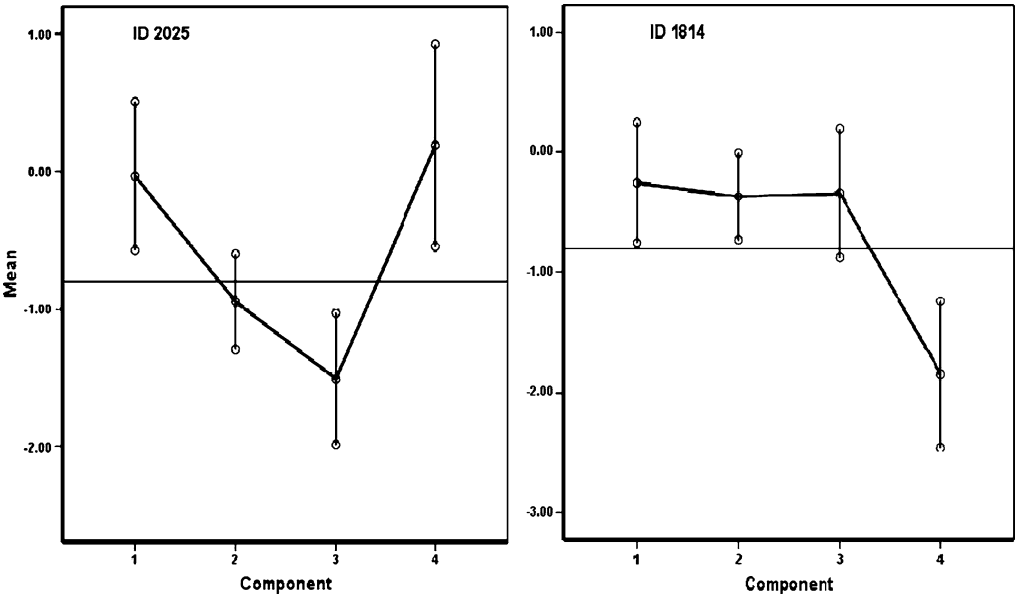


FIGURE 4.
Component thetas and confidence intervals for students near the cutline.

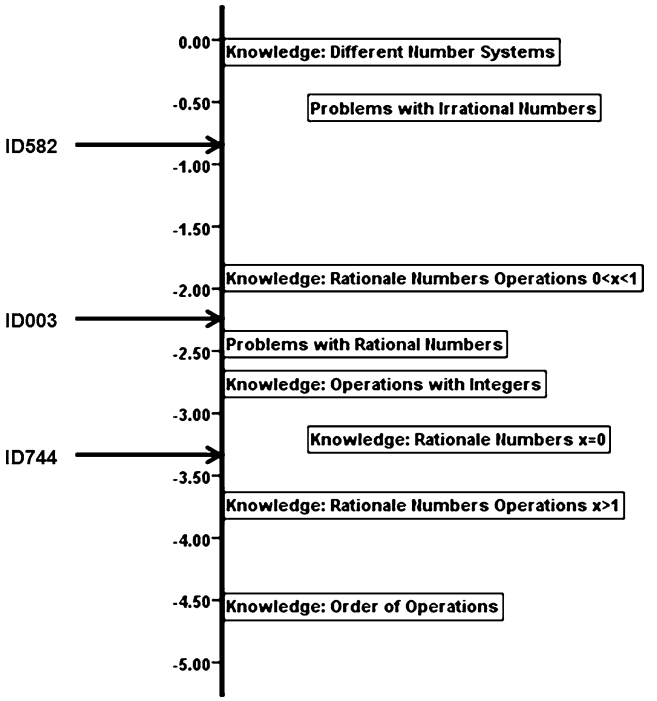


FIGURE 5.
Attribute and examinee locations on the number sense component.

the proficiency cutline on Component 1 (i.e., ID582). The probability that this examinee solves items with irrational numbers or with alternative number systems is less than 0.50, since competency level falls below these attributes. Thus, these skills are clearly not mastered. However, the probability of solving items with rational numbers is quite high as the examinee's estimated competency level of −0.95 is far above the location of these item attributes. Thus, these skills can be regarded as mastered.

6.3. Discussion

The mathematical achievement test data was well described by the hierarchical nature of MLTM-D. Parameters for four components were estimated using MML under both the full component model and the restricted component model. The item parameter estimates obtained for the full component model appeared reasonable, given the overall difficulty level of the test. Good fit at the item level was indicated by goodness of fit tests and the prediction of the item proportions correct from the model parameters was high. Further, the estimated correlations between the component competencies were moderate in level and did not indicate that the components were redundant.

The overall fit of MLTM-D to the mathematics data was also compared to other models. DINA was fit to the data using the same attributes that defined the MLTM-D components. An overall fit index (AIC) indicated that DINA did not fit as well as MLTM-D. Further, the DINA parameter estimates for guessing were too extreme to be plausible. Similarly, MLTM-D also fit the data better than an LCDM with both noncompensatory and compensatory effects of attributes. These overall fit indices, although substantially different between models, should be interpreted cautiously due to different estimation procedures. Finally, MLTM-D fit substantially better than a unidimensional 3PL model, which has many more parameters, including lower asymptotes to accommodate guessing.

The diagnostic potential of MLTM-D also was illustrated by the results. Examinees scoring just below the overall proficiency cutline were identified and analyzed for component levels. Within this group, substantial variability in component levels was observed and three of the four components had bimodal distributions, indicating differences in proficiency in regard to Number/Computation, Algebra, Geometry, and Data. Further, distinct patterns of component competencies for different examinees could be identified. Thus, remedial instruction could be differentiated at the level of the components. For the Number/Computation component, competency levels were linked to more specific attributes, for which a more concisely focused instructional intervention could be targeted.

7. Summary and Conclusion

The purpose of this paper was to present a multidimensional IRT model for cognitive diagnosis, the multicomponent latent trait model for diagnosis (MLTM-D). MLTM-D is a noncompensatory model with multiple latent dimensions that can be linked to diagnosis at two levels. That is, MLTM-D has a hierarchical structure, with components at the highest level and more specific attributes nested within components. At both levels, MLTM-D is a confirmatory model, as two specification matrices, a C-matrix and a Q-matrix, define the involvement of components and attributes in items, respectively. MLTM-D is a generalization of two earlier noncompensatory models, MLTM (Whitely, 1980) and GLTM (Embretson, 1984) that permits item level differences in component involvement.

MML estimators of the item parameters and conditions for model identification were presented, and parameter recovery was demonstrated with a simulation study. Then, MLTM-D was

applied to a high-stakes test of mathematical achievement to illustrate potential for diagnosis. Items were scored for the involvement of four components, defined by the test standards for mathematical achievement. Good fit was obtained for MLTM-D as compared to other diagnostic models. Diagnosis was illustrated at both the component level and the more specific attribute level. Thus, it was concluded that MLTM-D had potential for this type of achievement data.

MLTM-D can be compared and contrasted to the family of diagnostic classification models (DCMs) on several features, including, but not limited to: (1) the number of parameters to be estimated, (2) the conceptualization about the nature of competency (discrete versus continuous), and (3) the type of test for which the models are most appropriate. For all diagnostic models, fit will depend on both the appropriateness of the parameter structure of the model for the data and on the number of parameters to be estimated.

For both MLTM-D and DCMs, the number of parameters to be estimated can vary broadly, depending on the specific variant of the model that is applied. For MLTM-D, application at the component level alone leads to the largest number of parameter estimates. That is, at least one parameter, but no more than the number of components, is estimated for each item. Applying MLTM-D with attribute structures within components reduces the number of parameter estimates because item difficulty is replaced by linear combinations of attributes and control variables. Thus, MLTM-D can be applied to estimate only an intercept and a single weight for one ordered set of attributes within each component or it can be applied with multiple weights to estimate the impact of several attributes within components. Further, MLTM-D can be applied with mixtures of full versus restricted models within components.

Similarly, the number of estimated parameters varies greatly across different DCMs. For GDM (von Davier, 2008) and LCDM (Henson et al., 2009), models can be specified with large numbers of parameters to represent both compensatory and noncompensatory effects. Or, very restrictive models, with just a few parameters, also can be specified. Thus, the number of parameters is not a distinguishing feature between MLMT-D and the family of diagnostic latent class models.

The conceptualization of competency, continuous versus discrete latent variables, varies between MLTM-D and the DCMs. In MLTM-D, global competencies are specified as continuous dimensions. Mastery can be assessed from global competencies by establishing cutlines that represent some minimum probability of item solving, such as a content standard or a norm-based standard. Within component dimensions, in turn, item difficulty is impacted by ordered combinations of attributes. In this case, further diagnosis of competency involves linking through the common scale measurement of persons and attributes by identifying attributes that the examinee has a high probability to solve (e.g., $P \geq 0.70$). The accuracy of diagnosing specific attributes depends on a number of conditions, including model fit within components, relative distances between attributes and measurement error in global competency estimates. For components in which attributes are not sufficiently predictive of item difficulty, MLTM-D may still be applied to assess global component competency level. Importantly, however, the addition of narrow attributes to MLTM-D does not increase the number of diagnostic entities.

In DCMs, in contrast, the latent variables are discrete classes that represent different patterns of attribute mastery. The attributes in these models are treated as independent such that the various combinations of attribute mastery, in theory, can be estimated. The number of latent classes can be quite large when more than a few attributes are scored; hence, the number of diagnostic entities increases dramatically with the addition of narrow attributes. Thus, diagnosis is usually based on the probability of attribute mastery, which is estimated from the best fitting latent classes for an examinee. Like MLTM-D, a cutline for mastery is required (e.g., $P \geq 0.50$) on a continuous estimate of attribute possession probability. The diagnostic entities based on DCM attribute mastery patterns, however, is still greater than the diagnostic entities in MLTM-D. Accuracy of diagnosis in latent class diagnostic models depends on several conditions, including

model fit, level of slipping and guessing (for some models), differentiation between classes and linkages of attributes in the Q-matrix.

Although on the surface DCMs differ from MLTM-D, applications of the models may yield classification structures that represent the assumptions of the other type of model. That is, with DCMs, it is possible that the latent classes with substantial frequency can be stochastically ordered on one or two dimensions, as in some models (e.g., de la Torre & Douglas, 2004). Or, in MLTM-D, latent trait estimates with bimodal distributions may be obtained, seemingly representing mastery versus nonmastery. In any case, simulation studies are needed to examine the impact of using latent trait versus latent class models to determine the extent to which the models give similar diagnoses of individual competencies.

Both MLTM-D and DCMs have potentially broad scopes of application. MLTM-D is probably most applicable to relatively long tests with heterogeneous items that vary in the numbers and types of cognitive operations or skills involved in solution. Long tests and item heterogeneity at a global level are necessary to reliably measure the multiple dimensions. In achievement measurement, there are many tests with these characteristics, including each state's accountability testing of mathematics, reading, and science in Grades 3 to Grade 8, end-of-course tests in high school and high school graduation tests. Additionally, professional licensure tests and technical certification tests for many occupations also are long and contain heterogeneous items. In either case, components and attributes can be based on hierarchically organized test blueprints, which are typical for these tests, or on a hierarchical structure defined by a cognitive theory or principled structure analysis. For example, currently in progress is a principled structure analysis of textbook examples to define components and attributes of items for diagnosing difficulties in solving complex algebra problems (Yang, 2010).

Applications of models such as MLTM-D that are appropriate for heterogeneous tests could potentially define a new role for high-stakes tests in diagnosing areas for remediation. For example, for the state mathematics achievement test analyzed above, online tutorials are available for each indicator. Since this state is 100 % computerized in test administration and student records that are available to teachers, MLTM-D results could be used to define differential remediation for students who are near the overall competency cutline. A research project (Embretson, 2010) is currently underway to determine if the year-end test results alone, or followed by a very short diagnostic test, are adequate to prescribe remedial instruction. With the increasing computerization of testing, greater use of educational accountability tests for diagnosis will be possible.

The DCMs, such as the many variants of GDM (von Davier, 2008) and LCDM (Henson et al., 2009), also have a potentially broad scope of application, but often to different types of tests than MLTM-D. That is, since the potential number of diagnostic entities increases with the attributes, the DCMs are probably most applicable to shorter tests with more homogeneous items. Thus, for example, tests that are coordinated with specific units of instruction, such as classroom teaching, online tutorials and so forth, generally contain items that vary in fewer attributes. There are large numbers of such tests and research to apply diagnostic models in this context is in progress (e.g., Pellegrino, Goldman, DiBello, Gomez, & Stout, 2011).

Finally, it should be noted that MLTM-D requires powerful estimation methods, especially since the model is most effectively applied to long tests. The application to mathematics achievement data reported above employed MML with nonadaptive multidimensional quadrature. However, multidimensional quadrature becomes computationally difficult with more than five dimensions. Further, adding component error terms and lower asymptotes to MLTM-D, which would extend the usefulness of the model, will most likely increase computational burden substantially. Thus, future research should be devoted to developing or adapting estimation methods to be suitable for MLTM-D. For example, it may be possible to adapt a recent algorithm developed for confirmatory MIRT (Cai, 2010) to the noncompensatory structure of MLTM-D. Other possible estimation algorithms that could be explored include stochastic approximation (von Davier &

Sinharay, 2010) and a method with alternating imputation posteriors with adaptive quadrature (Cho & Rabe-Hesketh, 2011).

References

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, *46*, 443–445.

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

Cai, L. (2010). Metropolis-Hastings Robbins-Monroe algorithm for confirmatory factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307–335.

Cai, L., du Toit, S.H.C., & Thissen, D. (2012). *IRTPRO: Flexible, multidimensional, multiple category IRT modeling*. Chicago: Scientific Software International. Computer software.

Cho, S.-J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics & Data Analysis*, *55*, 12–25.

Daniel, R.C., & Embretson, S.E. (2010). Designing cognitive complexity in mathematical problem solving items. *Applied Psychological Measurement*, 348–364.

DeCarlo, L.T. (2011). The analysis of fraction subtraction data: the DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8–26.

de la Torre, J., & Chiu, C.-Y. (2009). Q-matrix validation under the generalized DINA model framework. In *The international meeting of the psychometric society*, Cambridge, St John's College, July 20th–24th.

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

DiBello, L.V., Stout, W.F., & Roussos, L.A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P.D. Nichols, S.F. Chipman, & R.L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Erlbaum: Hillsdale.

Embretson, S.E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika*, *49*, 175–186.

Embretson, S.E. (1995). Working memory capacity versus general central processes in intelligence. *Intelligence*, *20*, 169–189.

Embretson, S.E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika*, *64*, 407–433.

Embretson, S.E. (2010). *An adaptive testing system for diagnosing sources of mathematics difficulties*. Project R305A100234. Washington, Institute of Educational Sciences.

Embretson, S.E., & Daniel, R.C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science*, *50*, 328–344.

Embretson, S.E., & McCollam, K.M. (2000). A multicomponent Rasch model for covert processes. In M. Wilson & G. Engelhard (Eds.), *Objective measurement V*. Norwood: Ablex.

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

Gorin, J. (2007). Test design with cognition in mind. *Educational Measurement, Issues and Practice*, *4*, 21–35.

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

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

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

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.

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

Mislevy, R.J., & Bock, R.D. (1990). *BILOG 3: item analysis and test scoring with binary logistic models*. Chicago: Scientific Software International.

National Mathematics Advisory Panel (2008). *Foundations for success: final report*. U.S. Department of Education.

Pellegrino, J., Goldman, S., DiBello, L., Gomez, K., & Stout, W. (2011). *Evaluating the cognitive, psychometric and instructional affordances of curriculum-embedded assessments: a comprehensive validity-based approach*. Chicago: Learning Sciences Institute, University of Illinois.

Poggio, J.P. (2011). *Indicators, benchmarks and standards reflected in items for mathematical achievement tests in middle school* (Technical Report KU-10012011). Institute of Educational Science Mathematics Diagnosis Project, Lawrence, Kansas.

Roussos, L., DiBello, L., Henson, R., Jang, E., & Templin, J. (2010). Skills diagnosis for education and psychology with IRT-based parametric latent class models. In S.E. Embretson (Ed.), *Measuring psychological constructs: advances in model-based approaches*. Washington: American Psychological Association.

Rupp, A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*, 78–96.

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

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

Tatsuoka, K.K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*, *12*, 55–73.

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

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report RR-05-16). Princeton: ETS.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical & Statistical Psychology*, *61*, 287–307.

von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement: Interdisciplinary Research and Perspectives*, *7*(1), 67–74.

von Davier, M., & Sinharay, S. (2010). Stochastic approximation for latent regression item response models. *Journal of Educational and Behavioral Statistics*, *35*, 174–193.

von Davier, M., & Yamamoto, K. (2004). *A class of models for cognitive diagnosis*. Paper presented at the 4th Spearman conference. Educational testing service: the inn at Penn. Philadelphia, October.

von Davier, M., & Yamamoto, K. (2007). Mixture distribution Rasch models and hybrid Rasch models. In M. Davier & C.H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models*. New York: Springer.

Webb, N.L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Madison, National Institute for Science Education University of Wisconsin-Madison.

Whitely, S.E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*, 479–494.

Yang, X. (2010). *Construct theory-driven cognitive diagnostic testing in the domain of algebra story problems*. Shanghai: National Science Foundation of China.

Yang, X., & Embretson, S.E. (2007). Construct validity and cognitive diagnostic assessment. In J.P. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education*. New York: Cambridge University Press.

Manuscript Received: 30 APR 2010
Final Version Received: 23 FEB 2012