The Multicomponent Latent Trait Model for Diagnosis: Applications to Heterogeneous Test Domains

Susan E. Embretson

Applied Psychological Measurement published online 20 October 2014
DOI: 10.1177/0146621614552014

Applied Psychological Measurement
1–15

DOI: 10.1177/0146621614552014
apm.sagepub.com
SAGE

The Multicomponent Latent Trait Model for Diagnosis: Applications to Heterogeneous Test Domains

Susan E. Embretson[1]

Abstract

Heterogeneous item content is prevalent on psychological and educational tests that measure global traits or competencies, such as general intelligence and achievement tests. In such tests, the domain is broadly defined, so as to include many attributes and skills. Items often vary substantially in both the type and the number of attributes or skills that are involved in item solving. A hierarchical organization is often necessary to accommodate the heterogeneity of the test domain. For example, the mathematical achievement tests that are routinely administered in all U.S. states at the end of the school year typically have this hierarchical structure. The multicomponent latent trait model for diagnosis (MLTM-D) was developed for application to heterogeneous tests. MLTM-D is a confirmatory model that permits diagnosis at broad and more specific attribute or skill levels. In the current study, MLTM-D is applied to diagnose mastery at both the broad and specific skill levels in middle school mathematics. MLTM-D is presented, and methods involved in application are described.

Keywords

latent class models, cognitive psychology, psychometrics, diagnostic testing, multidimensional models

Tests designed to measure broad traits or competencies typically have heterogeneous item content to cover the scope of the domain. That is, many different types of attributes and skills are included on tests of general abilities, achievement, and proficiency. For such tests, representation of the various attributes or skills in the domain often is assured by organizing the items into subareas. For example, achievement tests in fundamental subjects such as mathematics, science, and reading are developed from specific blueprints that are hierarchically organized (e.g., National Governors Association for Best Practices, 2010). At the highest level, broad subareas are defined while at the lower levels, increasingly precise skills are defined. Proficiency tests that are widely used for certification in many professions and occupations often have similar structures. However, for many tests of broad traits or competencies, only an overall score is reported to examinees. For those examinees whose scores are near or below a proficiency

[1]Georgia Institute of Technology, Atlanta, USA

Corresponding Author:
Susan E. Embretson, School of Psychology, Georgia Institute of Technology, 654 Cherry St., Atlanta, GA 30332, USA.
Email: Susan.embretson@psych.gatech.edu

cutline, diagnostic information about the specific skills or attributes that are not mastered or possessed could be useful.

The year-end mathematical achievement tests that are routinely administered in all U.S. states exemplify the potential usefulness of diagnostic information. These tests have important consequences for students, teachers, schools, and states. Students with relatively low scores on these tests are not adequately prepared for the curriculum at the next grade level. However, the overall proficiency score is not helpful for diagnosing the specific skills that require remedial instruction. Diagnosing deficits in specific mathematical skills could help prescribe remedial instruction, particularly if coordinated with the year-end tests that are used for state accountability.

Developing models for diagnosing examinees' possession of attributes and skills has become an increasing prominent aspect of psychometrics (e.g., DiBello, Stout, & Roussos, 1995; Hensen, Templin, & Willse, 2009; von Davier, 2008). Unlike traditional item response theory (IRT) models, the diagnostic models provide assessments of skill patterns rather than global examinee scores on latent traits. In most diagnostic classification models (DCMs), latent classes define attribute or skill possession. Tests to which DCMs have been applied include both achievement and psychopathology tests (see Templin & Hensen, 2006).

Applications of diagnostic models require scores for the skills or attributes in individual test items. These scores are included in an item by attribute matrix (typically referred to as a Q-matrix), which is included in the diagnostic model. The Q-matrix represents a conceptual framework that defines the diagnostic estimates. The number of possible skill patterns increases geometrically with the number of skills in DCMs. For example, for a test with items that contain varying combinations of 15 skills, there are 2^{15} or 32,768 skill patterns. Thus, the assessment of patterns becomes computationally difficult for heterogeneous tests. Recent efforts have focused on developing more narrow diagnostic tests from a specified set of skills (Bradshaw, Izsák, Templin, & Jacobson, 2014). Thus, DCMs typically are not applied to heterogeneous tests, such as a year-end achievement tests.

The purpose of this article is to illustrate the application of a diagnostic model that is appropriate for heterogeneous tests. The test in the application is routinely administered at year-end to assess overall competency in mathematics for state accountability purposes. Furthermore, the test content covers a broad and heterogeneous domain, which is specified in a hierarchically organized blueprint. The multicomponent latent trait model for diagnosis (MLTM-D; Embretson & Yang, 2013) was developed to be applicable to heterogeneous and hierarchically organized domains. In the sections that follow, MLTM-D is applied to a recently administered form of a mathematical achievement test to diagnose mastery of both broad and narrow skills. Prior to presenting the study, background on MLTM-D along with methods for applying the model are presented.

The MLTM-D

This section includes a presentation of MLTM-D and the methods involved in its application, including specifying the model and parameter estimation, obtaining diagnostic categories, and assessing diagnostic reliability. A more detailed presentation of the theoretical development of MLTM-D, along with more extensive comparisons with other diagnostic models, is included in Embretson and Yang (2013).

The Model

MLTM-D is a confirmatory multidimensional latent trait model that specifies a noncompensatory relationship between the dimensions. The model is hierarchically organized and

provides diagnosis at two levels, the component level and the skill/attribute level, which is defined within components. Thus, similar to other diagnostic models, MLTM-D requires scores on individual items to represent content features that are relevant to components and to attributes or skills. However, unlike other diagnostic models, MLTM-D contains two different types of structures to represent two levels in a hierarchy of skills or attributes.

The first type of structure is a component structure matrix, $\mathbf{C}_{i\times m}$, that contains binary scores, c_{im}, to represent the involvement of component m in each item i, where $i = 1, \ldots, I$ and $m = 1, \ldots, M$. Components correspond to the highest level of a conceptual hierarchy that organizes a test domain. Some items may involve a single component whereas other items may involve two or more components, depending on content. The latent dimensions in MLTM-D are defined as components that combine multiplicatively. The second type of structure matrix, $\mathbf{Q}_m$, defines the attributes or skills in the lower level of the hierarchy. That is, given component m, $\mathbf{Q}_m$ is an $I \times K$ matrix, where $i = 1, \ldots, I_m$ and $k = 1, \ldots, K_m$ matrix for attributes, with I_m equal to the number of items involving component m and K_m equal to the number of attributes for component m. The Q_m are scored within components to represent the involvement of more specific attributes, skills, or other features in an item that may impact item difficulty. The scores may be either binary or continuous. Thus, Q_m contains scores on the K_m attributes that impact item difficulty for the items involving component m (i.e., $c_{im} = 1$). Similar to research on item difficulty modeling (see Gorin, 2007) for which Q-matrices are scored, item difficulty within the components is modeled as an additive combination of q_{ik}.

For MLTM-D, the probability that examinee j solves item i, $P(X_{ij} = 1)$ may be written as follows:

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

and

$$P_{ijm} = P\left(X_{ijm} = 1 | \theta_{jm}, \underline{q}_{im}, \underline{\eta}_m\right) = \frac{\exp\left(1.7\alpha_m\left(\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} + \eta_{m0}\right)\right)}{1 + \exp\left(1.7\alpha_m\left(\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} + \eta_{m0}\right)\right)}, \tag{2}$$

where c_{im} is a binary variable for the involvement of component m in item i, P_{ijm} is the probability that examinee j solves component m in item i, 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 α_m is a constant item discrimination for component m. Equation 1 specifies a conjunctive combination of the dimensions, where the probability of item solving depends on the probabilities of the components postulated for the items. If component m is not involved in item i ($c_{im} = 0$), then $P_{ijm} = 1$. Equation 2 specifies MLTM-D within components in the normal metric with the constant 1.7 and a single parameter for item discrimination, α_m. The q_{imk} in Equation 2 are scores for items that, when multiplied by estimated weights η_{mk}, model item difficulty. MLTM-D can be reformulated as a Rasch-family model because within components the items differ only in difficulty.

Special cases of MLTM-D should be noted. One special case occurs when $\mathbf{Q}_m$ is a matrix of binary variables that represent each item (i.e., dummy variables) when $K_m = I_m$. Hence, item difficulty for item i within component m equals η_{mk}, so that $\beta_{im} = \eta_{mk}$ and $\eta_{m0} = 0$. Thus, within component m, a one-parameter logistic (1PL) model is defined, such that each item obtains a parameter estimate, and Equation 2 is simplified as follows:

$$P_{ijm} = P\left(X_{ijm} = 1 | \theta_{jm}, \beta_{im}\right) = \frac{\exp\left(1.7\alpha_m\left(\theta_{jm} - \beta_{im}\right)\right)}{1 + \exp\left(1.7\alpha_m\left(\theta_{jm} - \beta_{im}\right)\right)}. \quad (3)$$

Formulating MLTM-D as 1PL model within components as in Equation 3 can provide a useful comparison with MLTM-D with scored attributes. In this case, MLTM-D would be a saturated model that could be compared with a restricted MLTM-D with attributes or skills that are postulated to predict item difficulty. Such a model is also useful for diagnosis when there is no attribute model that adequately predicts item difficulty. That is, if an empirically plausible model is not available at the attribute level, diagnosis would be available at the component level (see the following discussion). Another special case of MLTM-D is the linear logistic test model (LLTM; Fischer, 1973). That is, if only one component is specified, then Equation 2 for MLTM-D is a 1PL model variant of LLTM. Other special cases of MLTM-D are the earlier multicomponent latent trait model (Whitely, 1980) and the general latent trait model (GLTM; Embretson, 1984). Embretson and Yang (2013) presented details for these special cases and showed that MLTM-D is a generalization of the earlier models.

Specifying the Model

As MLTM-D is a confirmatory diagnostic model, successful applications require specifying components and attributes that are theoretically or conceptually meaningful and empirically plausible. The validity of the diagnosis depends on the previously obtained support for the theory or conceptual framework. Thus, to implement MLTM-D, $C_{i\times m}$ and the associated Q_m must be specified to operationalize a theoretical or conceptual framework for diagnosis. Furthermore, the scores in $\mathbf{C}_{i\times m}$ and the $\mathbf{Q}_m$ matrices must reflect a hierarchical organization, with the highest level providing scores for $\mathbf{C}_{i\times m}$ and the nested lower levels providing scores for $\mathbf{Q}_m$.

Blueprints for achievement and certification tests typically have the theoretical framework and hierarchical organization needed for MLTM-D applications. While items are typically written to correspond to a single lower level category, it is likely that complex items involve more than one indicator or component. Thus, one strategy to develop $\mathbf{C}_{i\times m}$ and the $\mathbf{Q}_m$ matrices is to have experts evaluate each item for possible involvement of lower level skills within each component. Then, for any item i for which $q_{imk} > 0$, then $c_{im} = 1$ for component m. For example, mathematical achievement tests used in Grades 3 to 8 in most states have a hierarchical conceptual framework that includes four areas: Number Sense, Algebra, Geometry, and Data. The framework represents a consensus of educators, mathematicians, and other experts. Score for $C_{i\times m}$ would reflect the involvement of these areas in a particular item. Within the four areas, specific definitions of skills are included which would define Q_m.

As noted earlier, MLTM-D may be applied when no attributes are available to meet the criteria of a linear ordering and empirically plausibility. As shown in Equation 3, a special case of MLTM-D occurs when attributes are specified as dummy variables in Q_{m_m} to define a 1PL model.

Finally, it should be noted that model identification depends on the structure of both $C_{i\times m}$ between components and Q_m within each component. If item blocks are formed such that each block b contains items with similar patterns of involvement of M components, then the matrix of component involvement in the blocks, $\mathbf{C}_{b\times m}$, must contain an $\mathbf{M} \times \mathbf{M}$ submatrix that is of full rank. For example, if three components are scored, 2^3 or eight patterns of component involvement are possible for items. Suppose, however, that only three unique patterns are found in a particular set of items. That is, if the three patterns are shown on the rows of $C_{b\times m}$, and if

$C_{b\times m} = \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}$ or $C_{b\times m} = \begin{matrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{matrix}$, the model will be identified. In the first case, as each item block involves only one component, it is obvious that model identification will be achieved. In the second case, the items in Blocks 2 and 3 involve multiple components, but the matrix is of full rank.

Finally, within components, Q_m must contain $\mathbf{K}_m$ independent vectors of item scores. That is, $K_m \leq I_m$, where I_m is the number of items and none of the vectors are dependent on the other vectors.

Parameter Estimates

Marginal maximum likelihood (MML) parameter estimation is feasible for MLTM-D item parameters. As MML involves multidimensional quadrature, the number of components must not be large to reduce computational burden. As for other multidimensional IRT models, computational burden becomes excessive with more than five components. The first and second derivatives of the data likelihood for MML estimation are given in Embretson and Yang (2013). Bayesian estimates of MLTM-D item parameters are feasible. For example, Bolt and Lall (2003) used Bayesian estimates for MLTM. However, the Bayesian estimates are likely to involve substantially longer computational time and hence may not be practical for many applications.

Similar to other multidimensional IRT models, the measurement scale for MLTM-D must be fixed for model identification. Two alternative methods can be readily applied. The measurement scale may be anchored to a distribution of person estimates. That is, the means and variances of the person estimates may be fixed (e.g., $\theta \sim MVN(\mathbf{0}, \sum)$), where the diagonal of $\mathbf{\Sigma}$ equals 1. Alternatively, the measurement scale may be anchored to items, by fixing the mean item difficulty (e.g., $M_b = 0$) and setting item discrimination to 1 for each component.

Person estimates may be obtained by several methods. Embretson and Yang (2013) used multidimensional expected a posteriori (EAP) for their analyses. EAP is easily implemented as the method involves only direct computations of likelihoods, not iterative searches. However, as EAP involves multidimensional quadrature, computational burden may be excessive if many quadrature points are used per dimension. Also, EAP requires specifying a prior distribution for theta, which may not be known or may not correspond well to standard specifications, such as a normal distribution. The modal a posteriori (MAP) and the maximum likelihood (ML_B) with semi-Bayesian estimates (Jannarone, Yu, & Laughlin, 1990) for the extreme scores both involve iterative searches for estimates. For both methods, algorithms must be adapted for MLTM-D to accommodate its conjunctive combination of components. This research is in progress. In general, however, MAP requires specifying a prior distribution and has the same limitations as EAP in that regard. ML_B, however, although not requiring a prior, may not be as efficient as using a prior that is appropriate.

Diagnosis With MLTM-D

Diagnosis with MLTM-D is feasible at both the component level and the attribute or skill level, contingent on the empirical plausibility of the model. Examinee estimates of mastery at the component level can be obtained from θ_m using mastery thresholds. Define y as a specified probability for mastery and P_{jm} as the probability that the average item involving component m will be solved by an examinee with θ_m. Then, a cutline for component m, φ_m, can be specified

such that if $\theta_m \geq \varphi_m$, then $P_{jm} \geq y$ for solving the average item on the test that involved component m.

As a Rasch-family model, cutlines for MLTM-D can be determined with $\beta_{.m}$, the mean item difficulty on component m. The log odds for an average item on component m for an examinee θ_m is

$$\ln\left(\frac{P_{jm}}{1-P_{jm}}\right) = 1.7\alpha_m(\theta_m - \beta_{.m}). \tag{4}$$

Then, the mastery cutline $\theta_m = \varphi_m$ may be given as follows, substituting $\theta_m = \varphi_m$ and setting $P_{jm} = y$:

$$\ln\left(\frac{P_{jm}}{1-P_{jm}}\right) = 1.7\alpha_m(\varphi_m - \beta_{.m}),$$

$$\varphi_m = \frac{\ln\left(\frac{P_{jm}}{1-P_{jm}}\right)}{1.7\alpha_m} + \beta_{.m}. \tag{5}$$

The cutlines established by applying Equation 5 for each component can then be compared with component theta estimates θ_m to assess component mastery for each examinee j and then combined to establish mastery patterns. For four components, for example, 2^4 or 16 mastery patterns are possible.

It should be noted that the value of y for the test as a whole is often established by agreement of panel of subject matter experts, standard-setting committees, or other considerations. The value of y can be applied to the test as a whole, to the components or within components to specify skills or attributes.

At the skill or attribute level, diagnosis depends on skill categories, q_{imk}, having an empirically strong linear relationship to item difficulty within each component m. In this case, examinees and attributes can be meaningfully located on a common measurement scale on each component because, as noted earlier, MLTM-D within components can be parameterized as the LLTM (Fischer, 1973), which belongs to the Rasch model family.

To obtain diagnosis for specific skills, boundary locations on the common measurement scale are needed for each skill. The boundary locations, γ_{km}, depend on (a) the predicted position of the skill categories from the MLTM-D restricted model, as in Equation 2, and (b) the probability that solving an item involving the skill exceeds a specified level, y. Similar to components, mastery for the k attributes or skills can be determined by comparing θ_{jm} to the mastery location γ_{km} of the attribute within a component. Thus, for an examinee with $\theta_{jm} \geq \gamma_{km}$, indicator k within component m is mastered.

Impact of Measurement Error on Diagnostic Reliability

As person estimates are based on dimensions in MLTM-D, individual standard errors of measurement are available. At issue here, however, is how measurement error impacts diagnosis, which depends not only on individual measurement errors but also on the mastery cutlines.

One approach to examining the impact of measurement error on diagnosis is to perform an uncertainty analysis on the examinee scores using imputation of plausible values. To implement the analysis, distributions of plausible thetas θ^*_{jm} are defined for each person j on each component m, based on the estimated thetas, θ_{jm}, and the standard errors of theta, $\sigma_{\theta_{jm}}$, from the person estimates for MLTM-D. Thus, the distribution of plausible thetas can be defined as normal, such

that $\theta^*_{jm} \sim N(\theta_{jm}, \sigma_{\theta_{jm}})$. For each person, values for θ^*_{jm} can be imputed as a random draw from the distribution, and component mastery can be determined by comparing each θ^*_{jm} to the established mastery cutlines. Diagnostic reliability would be assessed by the percentages of θ^*_{jm} that are above (or below) the cutline. Similarly, skill mastery for θ^*_{jm} can be determined using the locations and the procedures described earlier for skills and attributes. The uncertainty analysis can be applied to establish individual values for each examinee using multiple imputations. For a large population, values using a single imputation may be appropriate to determine overall impact on diagnosis.

Comparison With Other Diagnostic Models

As noted earlier, diagnostic models require scores on items to represent attribute or skill involvement. Thus, a Q-matrix, such as elaborated earlier, is typically formulated in the models to specify parameters to be estimated. MLTM-D differs from DCMs in both the nature of the attributes that are appropriate and the type of person estimates that are obtained. For person estimates, DCMs (e.g., DiBello et al., 1995; Hensen et al., 2009; von Davier, 2008) are based on latent classes of examinees, which represent varying patterns of attribute possession. In contrast, MLTM-D is multidimensional and the person estimates represented in the model are continuous, not discrete. However, this distinction while appropriate at the model level, in practice, both MLTM-D and DCM typically include both continuous and discrete person estimates. That is, for DCMs, continuous estimates for persons are often obtained as the probabilities of possessing each attribute. MLTM-D, however, can provide mastery classifications, using the cutline method described previously.

Perhaps the most salient difference between MLTM-D and the DCMs is the organization of the attribute specifications. MLTM-D is a hierarchically organized model in the sense that the latent dimensions (i.e., the components) are defined at the highest level and nested within the components are attributes and skills at the lower level. In contrast, while hierarchical organizations of skills are specifiable in DCMs, the nested relationships are not necessary for estimation of person parameters. In contrast, in MLTM-D, the latent person dimensions are estimated only at the highest level to model component probabilities. To determine skill mastery, MLTM-D requires an empirically plausible model of skill clusters, and other item stimulus features that underlie component difficulty. Thus, reliable estimates from MLTM-D are most likely when applied to heterogeneous item domains in which broad differences in person competencies and skill difficulties are observed. The DCMs, in contrast, can be reliably applied to more narrowly defined domains with fewer skills. Finally, while MLTM-D can be estimated by more computationally intense Bayesian methods utilized for many DCMs, MML is more feasible.

Application of MLTM-D for Diagnosing Skills in Grade 7 Mathematics

The application of MLTM-D in this study is part of a larger research project (Embretson, 2010) that is being conducted in a participating state. Typical of state accountability tests, as described above, the conceptual framework for the assessed skills is hierarchically organized with the standards (Number, Algebra, Geometry, and Data) at the highest level and increasingly specific skills defined within standards. Online tutorials are available to correspond to the skills within the mathematical standards. Thus, diagnosing the skill deficits from the year-end test has potential to target remedial instruction.

In the current study, MLTM-D is applied to diagnose skill patterns using results from the state accountability test administered at the end of the school year. In the larger project,

estimates from the year-end test are combined with a short adaptive test to prescribe remedial instruction for the next grade.

Method

Test and procedures. Items were developed to correspond to the lowest level in the hierarchy, which consists of indicators. Figure 1 presents examples of items that are similar to items for the Geometry standard. The test administered in Spring 2012 included 70 operational items. The number of items designated for each standard was as follows: (a) Number, 15 items; (b) Algebra, 23 items; (c) Geometry, 23 items; and (d) Data, 9 items, with 31 subindicators assessed. All tests were computer administered in three parts on 3 consecutive days.

Examinees. The examinees were 32,724 Grade 7 students taking the year-end achievement test. Three operational test forms were administered, containing the same 70 operational test items, but in scrambled orders to minimize cheating. A total of 14 subforms were nested within the operational test forms, containing different sets of tryout items. Thus, 84 items were administered; 70 operational items and 14 tryout items.

Item scores. The 70 operational items were scored for the involvement of the indicators of the four standards by a panel of educators, an educational psychologist, and a mathematician. Consider the three items in Figure 1. These items were written for three separate subindicators on Geometry. Thus, the presence of the indicators in the items was reflected by scores in Q_m. As each item involved an indicator of the geometry standard, component m for Geometry could be scored ''1'' in $C_{i\times m}$. However, the item shown at the bottom of Figure 1 was also scored as involving one of Number indicators; hence, it could be scored for the appropriate indicator in Q_m under Number and also scored ''1'' in a $C_{i\times m}$ for involvement of the Number component.

Using the standards to define components as described previously, eight different patterns of component involvement were observed in the items as shown in Table 1. It can be seen that approximately 80% of the items involved a single component (i.e., patterns 0001, 0010, 0100, 1000) and that approximately 20% of the items involved multiple components. A $\mathbf{C}_{i\times m}$ matrix was prepared to specify the component involvement in each item. Also, within each component, $\mathbf{Q}_m$ matrices were prepared to specify the indicators involved in each item.

Results

Descriptive statistics. The overall mean (M = 52.63) and standard deviation (SD = 12.401) across test forms indicated high levels of performance, as typical for a state achievement test in middle school. The test form means did not differ statistically (F = .108; df = 2, 32721; p = .956). The mean of p values for item difficulty was high ($M_{p\text{value}}$ = .752), and the mean correlation of p values across forms was .974. Thus, scrambling of item order between forms had little impact on performance. A random sample of 8,585 students taking the same test form was used for the estimation of MLTM-D parameters.

MLTM-D models. MML estimates of the item parameters were obtained, using a SAS nonlinear mixed modeling procedure, where items were specified as fixed variables and examinees as random variables. For computational convenience with multidimensional quadrature, MLTM-D was specified in the normal metric with a constant item discrimination parameter estimated for each component, as presented in Equations 2 and 3. Examinees were specified as random variables from a standardized multivariate normal distribution ($\theta \sim \text{MVN}(0,\Sigma)$), where Σ was

Item for Indicator "Figure Properties-Quadralaterals" K3c

Refer to the image below. Which statement about a parallelogram is always true?

A) X Two pairs of sides are parallel, and opposite sides are congruent.
B) The sum of the interior angles is 270 degrees.
C) No sides are congruent.
D) No sides are parallel.

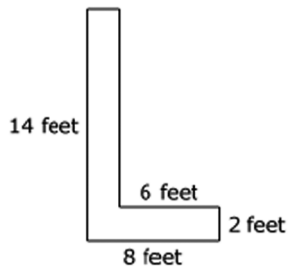
Item for Indicator "2d Area Formulas" K4

The radius of a circle is 4 meters (m). What is the area of this circle in square meters (m^2)? (Use $\pi = 3.14$)

A) 200.96 m^2
B) X 50.24 m^2
C) 25.12 m^2
D) 12.56 m^2

Item for Indicator "Real Geometry Problems" A1c

A letter L that is going to be painted on the roadway is shown below.



What is the total area that must be painted?

A) 12 ft^2 B) 18 ft^2
C) X 40 ft^2 D) 120 ft^2

Figure 1. Examples of geometry items for Grade 7.

specified with 1.0 for the diagonal and six free parameters to represent trait covariances. Integration across the random variables was achieved using Gauss–Hermite quadrature.

Successful diagnosis of performance at the lowest level in MLTM-D depends on both the multidimensionality of the test and the empirical plausibility of predicted item difficulties from

Table 1. Final Component Patterns on 70 Operational Items.

Standard NAGD	Frequency	%	Valid %	Cumulative %
0001	4	5.7	5.7	5.7
0010	20	28.6	28.6	34.3
0100	16	22.9	22.9	57.1
0101	4	5.7	5.7	62.9
0110	3	4.3	4.3	67.1
1000	16	22.9	22.9	90.0
1001	1	1.4	1.4	91.4
1100	6	8.6	8.6	100.0
Total	70	100.0	100.0	

Note. N = Number; A = Algebra; G = Geometry; D = Data.

more specific skills or attributes. Thus, assessing fit involves model comparisons. Thus, three variants of MLTM-D were specified. For all models, component involvement in items is specified by C_{ixm}. For the saturated MLTM-D, separate item difficulties were estimated for each item within a component with $\mathbf{Q}_m$ as an identity matrix to specify a 1PL model within each component. For the restricted hierarchical MLTM-D, $\mathbf{Q}_m$ was a matrix of binary variables of scores for the indicators or subindicators involved in each item. Finally, for the null MLTM-D, $\mathbf{Q}_m$ was a unity vector, such that a constant item difficulty is specified within components. The null model is a comparison model with constant item difficulty. The three variants of MLTM-D were estimated; a saturated model (−2lnL = 539,469; Akaike information criterion [AIC] = 539,657; #parameters = 94), a restricted model (−2lnL = 558,793; AIC = 558,901; #parameters = 54), and a null model (−2lnL = 615,156; AIC = 615,184; #parameters = 14). Thus, the models include varying numbers of parameters for item difficulty plus four-item discrimination parameter estimates (one for each component) and six trait covariances. As expected, because there are no constraints on item difficulty, the best-fitting MLTM-D is the saturated model, as it has lowest values for both AIC and BIC (Bayesian information criterion).

To examine the multidimensionality of the data, a unidimensional 1PL model (−2lnL = 544,266; AIC = 544,408; #parameters = 71) was also fit to the data. Compared with the full MLTM-D, fit was significantly worse for the 1PL model ($\Delta\chi^2 = 4{,}872$; $df = 23$; $p < .001$) and AIC was higher, which supports multidimensionality.

The empirical plausibility of the restricted within-component model in MLTM-D was examined in two ways. First, a likelihood ratio fit statistic (Embretson, 1997), a generalization of a fit statistic used in structural equation modeling, was computed as follows:

$$\Delta^2 = \frac{(\ln L_{\text{null}} - \ln L_{\text{restricted}})}{(\ln L_{\text{null}} - \ln L_{\text{saturated}})}. \tag{6}$$

In unidimensional applications of explanatory IRT models, such as LLTM (Fischer, 1973), the value of Δ is similar in magnitude to a multiple correlation of predictors with item difficulty. Using the likelihoods for the three MLTM-Ds earlier, the fit index ($\Delta = .862$) indicated overall strong prediction of item difficulty from the indicators. Second, a 1PL variant of the LLTM (Fischer, 1973) was also fit to the data (−2lnL = 566,493; AIC = 566,577; #parameters = 42). As compared with the restricted MLTM-D model from earlier, fit was significantly worse ($\Delta\chi^2 = 7{,}700$; $df = 12$; $p < .001$) and AIC was higher than the restricted MLTM-D. Thus, based on the skills represented by the indicators, the restricted MLTM-D is the best explanatory model.

Item fit for MLTM-D was assessed by comparing expected and observed frequencies of item responses based on grouping examinees with similar expectations for item success. Item probabilities were predicted for the full sample of 8,585 examinees and 70 items (i.e., 600,950 responses) using the item parameters from the MLTM-D saturated model. As MLTM-D is multidimensional and conjunctive, categorizations for all items based on a single dimension cannot be justified. Thus, separate categorizations of examinees were made for each item to define similar expectations, with examinees placed into categories based on their expected probability of item solving, P_{ij}, for each item. Fourteen intervals were defined to classify examinees. Categories with fewer than 10 observations were collapsed into the next higher category, with the mean observed number of categories at 12.

The predicted and observed frequency of examinees with correct responses was obtained for each category within each item. Likelihood ratio fit statistics, such as the Bock and Mislevy (1990) G^2, could be computed for each item. However, due to the very large sample size, power was calculated to be in the .90s for very small differences between the expected and observed probabilities. Thus, two alternative methods were used to assess fit. First, standardized residuals, SR_i, were computed by the difference between observed and expected frequencies, f_{ic}^o and f_{ic}^e, respectively, in category c summed across categories as follows:

$$SR_i = \left(\frac{1}{N_i^c}\right)\left[\frac{\sum_c \left(f_{ic}^o - f_{ic}^e\right)}{\sqrt{f_{ic}^e}}\right], \tag{7}$$

where N_i^c is the number of categories within item i. One very easy item, with only two observed categories, had a very large standardized residual (SR_i = 17.762). However, for the remaining items, an approximately normal distribution of standardized residuals was obtained (M = 0.130, SD = 1.123), thus indicating good fit. Second, plots were prepared to examine the relationship of MLTM-D predicted item success within categories, P_{ic}^e, to observed probabilities, P_{ic}^o. The summary plot across all items and categories indicated a very strong relationship (r = .945). Thus, the item fit data by both standardized residuals and plots supported the MLTM-D saturated model.

Table 2 presents descriptive statistics on the item difficulty estimates from the MLTM-D saturated model. It can be seen that the mean item difficulty was low, consistent with the high p values typical for an achievement test. The four constant item discrimination estimates were as follows: Number Sense, .559; Algebra, .672; Geometry, .701; and Data, .610.

Person estimates and diagnosis. Multidimensional EAP estimates of the person component competencies were obtained using MLTM-D item parameters and an SPSS macro. Table 2 also presents the means, standard deviations, and empirical reliabilities (Bock & Zimowski, 2003) of θ_m for each component. It can be seen that the mean thetas are somewhat higher than zero and the standard deviations are somewhat larger than 1. The empirical reliabilities were in the .80s for Number, Algebra, and Geometry. However, the empirical reliability for Data was only .613.

Competency cutlines on the components, φ_m, are also shown in Table 2 for each component using the procedures described previously. For the cutlines shown in Table 2, y was specified as .70 for examinees' mean probability of solving an item P_{jm} on component m. Applying Equation 5, the cutlines shown in Table 2, φ_m, were obtained. Component mastery patterns were determined by comparing individual theta estimates to the cutline, if $\theta_m \geq \psi_m$, then ψ_m = 1, and ψ_m = 0, otherwise.

Table 3 presents the frequencies of component standard mastery for 8,585 students used to calibrate MLTM-D parameters. The patterns of competency shown in Table 3, in order, refer to

Table 2. Descriptive Statistics for MLTM-D Parameters for Primary Components.

		Item parameters			Person estimates			
	Items	*M*	*SD*	*M SE*	*M*	*SD*	Reliability	Cutline
Number	16	−1.3370	0.7280	0.0337	0.3143	1.2682	.817	−.4356
Algebra	23	−1.4019	0.8614	0.0502	0.1760	1.1995	.828	−.6579
Geometry	22	−1.3606	0.9644	0.0360	−0.0398	1.1966	.853	−.6473
Data	9	−1.6234	1.6838	0.0670	0.0577	1.1284	.610	−.8041
	70	−1.4026	0.9856	0.0438				

Note. MLTM-D = multicomponent latent trait model for diagnosis.

Table 3. Frequency and Percentages of Students in the 16 Mastery Patterns.

Pattern NAGD	Frequency	%	Pattern NAGD	Frequency	%
0000	856	10.0	1000	141	1.6
0001	527	6.1	1001	184	2.1
0010	77	0.9	1010	90	1.0
0011	136	1.6	1011	220	2.6
0100	242	2.8	1100	175	2.0
0101	289	3.4	1101	389	4.5
0110	126	1.5	1110	441	5.1
0111	320	3.7	1111	4,372	50.9

Note. N = Number; A = Algebra; G = Geometry; D = Data.

Number, Algebra, Geometry, and Data. All 16 possible patterns of standard mastery were observed in the sample, and 49.1% of the examinees were indicated as below mastery on one or more components. Two patterns had relatively low frequencies: Pattern 0010 (77 students) and Pattern 1010 (90 students). The patterns with a single standard not mastered (e.g., 0111) were observed for a total of 1,370 students, and patterns in which two standards were not mastered (e.g., 0011) were observed for 738 students.

Table 4 shows the percentage of students with mastery for each component standard. The percentage of students with mastery ranged from 67.4 to 75.0, with the lowest mastery level for Geometry.

A diagnosis of skill mastery was obtained by comparing person estimates on the four components to the locations of skills on each latent dimension. As for overall mastery, for each skill indicator, the cutlines γ_{km} were set with $y = .70$. Thus, for an examinee with $\theta_m \geq \gamma_{km}$, indicator k on component m is mastered. Table 4 presents descriptive statistics on the number of non-mastered skills for the four standards by students' mastery status, with the overall mastery cutlines applied to the composite of skills within the standards. Thus, for students with overall mastery, some more difficult skills may not be mastered. For example, that the mean number of non-mastered skills in Number is about 5 (4.7474) for students below mastery, but less than 1 (0.4168) for students with mastery. The other three standards have similar differences between masters and non-masters.

Diagnostic reliability is impacted by both the measurement error and the cutlines that are specified. To develop expectations for diagnostic reliability, an uncertainty analysis was performed. To implement the analyses, distributions of plausible thetas for each examinee, θ_m^*,

Table 4. Descriptive Statistics on Composite Mastery and the Number of Skills Not Mastered by Component Standard Mastery.

				Number of skills not mastered				
	Composite overall mastery			Non-masters		Masters		Skill
Standard	0	1	Consistency	*M*	*SD*	*M*	*SD*	RMSE
Number	30.0	70.0	89.6	4.7474	2.6628	0.4168	0.6836	1.267
Algebra	26.0	74.0	90.4	7.8826	0.6905	3.6193	1.9560	1.280
Geometry	32.6	67.4	90.3	10.0582	1.6839	2.2542	2.8984	1.981
Data	25.0	75.0	83.7	5.8953	0.3167	1.3603	1.4931	1.500

Note. RMSE = root mean square error.

were defined for each component m. The plausible distribution was based on the estimated thetas, θ_m, and the standard errors of theta, σ_{θ_m}, obtained from the EAP estimates. Thus, the distribution of plausible thetas for each person was defined as normal, with $\theta^*_m \sim N(\theta_m, \sigma_{\theta_m})$. A single θ^*_m for each person was obtained by randomly drawing from their Stage 1 distribution. Then, for each θ^*_m, component mastery and skill mastery cutlines were defined as described earlier. The results were summarized across the sample of 8,585.

Table 4 presents results on the impact of measurement error on component mastery and skill mastery for the sample as a whole based on imputed thetas. Consistency was approximately 90% for all areas except data (83.7%). Thus, the classification of mastery versus non-mastery was consistent based on plausible thetas. The impact of measurement error on skill mastery was explored by comparing the predicted number of skills not mastered based on the estimated θ_m to the predicted number of skills based on plausible thetas θ^*_m. The root mean square errors were computed and are shown in Table 4. The expected difference in number of skills mastered ranged from 1.267 to 1.981.

Discussion

Similar to other applications of diagnostic models to achievement tests (e.g., Bradshaw et al., 2014), a major goal of the current application of MLTM-D was to provide a basis for individualized instruction in mathematics. The results from this study generally supported this goal using a year-end mathematical achievement for Grade 7 students. Eight different patterns of the involvement of mathematical standards (areas) were observed in the items. Using these standards to define components, MLTM-D was estimated on a large sample of students. The fit of MLTM-D was supported through both model comparisons and analysis of item fit. Furthermore, the empirical reliabilities of student scores in the Number, Algebra, and Geometry components were moderately high. However, for Data, reliability was lower due to the smaller number of items. As diagnosis depends on mastery cutlines as well as empirical reliability, an uncertainty analysis was undertaken to further examine diagnostic reliability. Diagnostic reliability was high for Number, Algebra, and Geometry (approximately .90) but lower for Data (approximately .84). Furthermore, the difference in the diagnosed number of non-mastered skills was small, ranging from approximately 1 and 2 skills per area. While diagnostic reliability was generally supported, a second stage of testing may be needed to increase precision, especially for the Data area.

Analysis of the student mastery patterns generally supported differential diagnosis. At the component level of MLTM-D, a substantial percentage of students fell below mastery on one

or more mathematics standards. That is, even for students who had overall proficiency as established by the state cutlines on the whole test score, applying competency cutlines to the component estimates often indicated one or more standards below mastery level. These results can be expected as the overall test score is compensatory with respect to the standards. Frequencies in all possible patterns of mastery were found, thus supporting differential remedial instruction based on component-level estimates.

Diagnosis of more specific skills was also supported. A relatively high likelihood ratio fit index of the restricted MLTM-D, in which specific skills predict item difficulty, supported the model as empirically plausible. Thus, student competencies at the component level could be aligned with skill difficulty through common scale measurement. Substantial differences between students in the number of deficient skills were observed. Thus, these results suggest that for students with a non-mastered component, the full set of skills may not need remediation.

In summary, the results generally support individual patterns of mastery in the four standards and in the associated skills. These results are suggestive of the potential of individualized instruction to remediate deficiencies in mathematics. However, the results also suggest that while diagnostic reliability is high in three of four areas of mathematics, one area could benefit from increased reliability. Thus, the usefulness of a second stage of testing should be explored in future research.

Finally, it should be noted that MLTM-D, similar to other diagnostic models, has a potentially broad scope of application. MLTM-D (Embretson & Yang, 2013) was developed for application to heterogeneous tests in which item content represents a broad domain and items vary in the number and the nature of skills or attributes that are involved. While achievement and certification tests often have appropriate conceptual structures based on item content, theoretical frameworks may also be based on theories of cognitive processing stages or cognitive demand in items (e.g., Gorin, 2007) and thus be applicable to ability and aptitude tests. Multistage processing theories of item cognitive complexity have hierarchical structures because different aspects of item content impact different stages of processing. For example, Embretson and Daniel (2008) applied an empirically supported cognitive processing model of mathematical problem solving to the Quantitative Reasoning Test on the Graduate Record Examination. At the highest level, the theory postulates five processing stages (Translation, Integration, Solution Planning, Solution Execution, and Decision Processing) based on an empirically supported theory of mathematical problem solving. The items on the Quantitative Reasoning Test were heterogeneous with respect to involvement of the stages. For example, some items, with given equations and two or more unknowns, involved only the stages of Translation, Solution Planning, and Solution Execution. Other items with text, but no equations or unknowns, involved only Translation and Integration. Thus, the stages would be appropriate as components in MLTM-D because varying combinations are required for problem solution. Within stages, the model of mathematical problem solving specified different content features that impacted processing difficulty.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research in this report was partially supported by a Goal 5 (Measurement)

Grant from the Institute of Educational Science R305A100234 to Susan Embretson, Principal Investigator (Georgia Institute of Technology).

References

Bock, R. D., & Zimowski, M. (2003). *IRT from SSI: BILOG-MG* (M. du Toit, Ed.). Chicago, IL: Scientific Software International.

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.

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational number: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice*, *33*, 2-14.

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). Hillsdale, NJ: Erlbaum.

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

Embretson, S. E. (1997). Multicomponent latent trait models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305-322). New York, NY: Springer-Verlag.

Embretson, S. E. (2010). *An adaptive testing system for diagnosing sources of mathematics difficulties* (Project R305A100234). Washington, DC: 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., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika*, *78*, 14-36.

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.

Hensen, 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*, 1919-210.

Jannarone, R. J., Yu, K. F., & Laughlin, J. E. (1990). Easy Bayes estimation for Rasch-type models. *Psychometrika*, *55*, 449-460.

National Governors Association for Best Practices. (2010). *Common core standards*. Washington, DC: Council of Chief State School Officers.

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

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

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