Applied Psychological Measurement
2015, Vol. 39(1) 62–79
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621614561315
apm.sagepub.com
SAGE

A Family of Generalized Diagnostic Classification Models for Multiple Choice Option-Based Scoring

Louis V. DiBello[1], Robert A. Henson[2], and William F. Stout[1,3]

Abstract

This article proposes a new family of diagnostic classification models (DCM) called the Generalized Diagnostic Classification Models for Multiple Choice Option-Based Scoring (GDCM-MC). The GDCM-MC is created for multiple choice assessments with response options designed to attract particular kinds of student thinking and understanding, both desired (correct) thinking and problematic (incorrect or partially correct) thinking. Key features that combine to distinguish GDCM-MC are: (a) an expanded latent space that can include both desirable and problematic facets of thinking, (b) an expanded **Q** matrix that includes a row for each response option and that uses a three-valued coding scheme to specify which latent states are strongly attracted to that option, (c) a guessing component that responds to the forced choice aspect of multiple choice questions, and (d) a general modeling framework that can incorporate the diagnostic modeling functionality of almost any dichotomous DCM, such as deterministic input, noisy "and" gate (DINA), reparameterized unified model (RUM), loglinear cognitive diagnosis model (LCDM), or general diagnostic model (GDM). The article discusses these four components and presents the GDCM-MC model equation as a mixture of cognitive and guessing components. Two identifiability theorems are presented. A Bayesian Markov Chain Monte Carlo (MCMC) model estimation algorithm is discussed, and real and simulated data studies are reported.

Keywords

diagnostic testing, latent class models, psychometric theory

[1]University of Illinois at Chicago, USA
[2]The University of North Carolina at Greensboro, USA
[3]University of Illinois at Urbana–Champaign, USA

Corresponding Author:
Louis V. DiBello, Research Professor and Associate Director, Learning Sciences Research Institute, University of Illinois at Chicago, 1240 W. Harrison Street, Room 1570 S, Chicago, IL 60607-7137, USA.
Email: ldibello@uic.edu

Introduction

This article proposes a new family of diagnostic classification models (DCM) called the Generalized Diagnostic Classification Models for Multiple Choice (GDCM-MC) Option-Based Scoring. The GDCM-MC is created for multiple choice assessments with response options designed to attract particular kinds of student thinking and understanding, both desired (correct) thinking and problematic (incorrect or partially correct) thinking. The key distinguishing features of GDCM-MC are: (a) an expanded latent space that can include both desirable and problematic facets, (b) an expanded **Q** matrix with a row for each response option and with a three-valued coding scheme for specifying which latent states are strongly attracted to that option, (c) a guessing component that responds to the forced choice aspect of multiple choice questions, and (d) a general modeling framework that can incorporate the diagnostic modeling functionality of any dichotomous DCM, such as deterministic input, noisy ''and'' gate (DINA), reparameterized unified model (RUM), loglinear cognitive diagnosis model (LCDM), or general diagnostic model (GDM).

Concrete examples of option-linked diagnostic assessments include facet-based physics assessments (Minstrell, 1992, 2001), Diagnostic Algebra and Geometry Assessments (Masters, 2012, 2014; Masters & Chapman, 2011), and concept inventories (Miller et al., 2006; Pellegrino, DiBello, James, Jorion, & Schroeder, 2011; Pellegrino et al., 2013; Steif & Dantzler, 2005; also see http://ciHUB.org). These assessments acknowledge the formative assessment importance of diagnosing problematic thinking in addition to diagnosing desirable thinking, but they lack a psychometric, diagnostic foundation such as can be provided by the GDCM-MC.

A foundational introduction to the GDCM-MC modeling approach first reviews the DINA and RUM dichotomous DCMs and then presents the GDCM-MC components, the general GDCM-MC model equation, and two identifiability theorems. The model estimation algorithm is discussed, and real and simulated data studies are reported.

Background on DCMs for Dichotomous Item Scoring

The DINA and RUM models are reviewed to highlight similarities and differences between dichotomously scored DCMs and the GDCM-MC option-based scoring. The authors then show how to incorporate each model within the GDCM-MC framework.

DCMs are restricted latent class models (Haertel, 1989) designed to classify students into multidimensional skill mastery profiles $\underline{\alpha} = (\alpha_1, \ldots, \alpha_K)$ with $\alpha_k = 1/0$ for skill k mastery/nonmastery (Leighton & Gierl, 2007; Rupp, Templin, & Henson, 2010). An important component of DCM models is an $I \times K$ **Q** matrix, where I is the number of items and K the number of skills. For each item i, the ith row $\underline{q}_i = (q_{i1}, \ldots, q_{iK})$ of **Q** gives $q_{ik} = 1/0$ according to whether skill k is or is not required, respectively, for item i.

The DINA Model

The DINA model item response function (IRF; Junker & Sijtsma, 2001) involves two parameters for each item i, a ''slip'' parameter s_i and a ''guess'' parameter g_i:

$$P(X_i = 1|\underline{\alpha}) = (1 - s_i)^{\zeta_i} g_i^{(1-\zeta_i)}, \tag{1}$$

where $0 \leq s_i, g_i \leq 1$ and ζ_i is a function of both $\underline{q}_i$ and $\underline{\alpha}$ defined by $\zeta_i = 1$ if the student is a master of all skills required for item i, and $\zeta_i = 0$ if not. We see that the DINA probability of

$P(X_i = 1|\underline{\alpha}) = g_i$ is constant for *all* $\underline{\alpha}$ that lack one or more required skills for item i regardless of which or how many required skills are lacking. The usual item response local independence given latent skills vector $\underline{\alpha}$ is assumed for DINA and for all item response models herein.

The RUM

The (Reduced) RUM (DiBello, Roussos, & Stout, 2007; DiBello, Stout, & Roussos, 1995; Roussos et al., 2007) is also referred to in the literature as the Fusion Model. The RUM (the term *reduced* is dropped here) and Fusion models are equivalent. The RUM IRF is,

$$P(X_i = 1|\underline{\alpha}) = \pi_i \prod_{k=1}^{K} r_{ik}^{q_{ik}(1-\alpha_k)}. \tag{2}$$

As $r_{ik}^0 = 1,$ the RUM imposes the r_{ik} as multiplicative penalties on the probability of a correct response specifically depending on which of the required skills for item i are lacking. The RUM IRF involves $K_i + 1$ parameters, where K_i is the number of required skills for item i. For $K_i > 1$, this is more than the two DINA parameters (slip and guess) per item. The RUM's efficient use of $K_i + 1$ parameters to account for 2^{K_i} conditional probabilities of a correct answer over $\underline{\alpha}$ is useful when the assumption of homogeneity of probabilities across all item non-masters is not tenable, and when there are enough students and items to estimate more parameters per item.

The GDCM-MC Model Family

Next, the four key model features of GDCM-MC are discussed: (a) expanded latent space, (b) expanded **Q** matrix, (c) guessing component, and (d) fully general modeling framework that can incorporate almost any DCM functionality. This is followed with a discussion of model identifiability.

The Expanded Latent Space for the GDCM-MC

Motivated by current option-linked diagnostic assessments, the student latent space is expanded to include problematic (including misconceptions and partially correct thinking) as well as desirable (skills and conceptual understanding) facets of thinking. Let $k = 1, \ldots, K$, where each k is either a desirable or problematic facet. For simplicity, dichotomously coded facets that a student either possesses or lacks are assumed. Then, the latent space can be represented by the set of all 2^K vectors $\underline{\alpha} = (\alpha_1, \ldots, \alpha_K)$, where α_k = 1/0 means the student possesses/lacks, respectively, facet k. Note the shift in language from ''master/non-master of a skill'' to ''possession/lack of a facet.'' Whether α_k = 1 is advantageous or not for a student depends on whether k is a desirable or problematic facet of thinking.

Expanded GDCM-MC **Q** Matrix

The GDCM-MC **Q** matrix is expanded in two ways. First, **Q** has a row per response option, instead of just one row per item as in the dichotomous scoring case. So the **Q** matrix for a test with 30 items, each with 4 options, will have $30 \times 4 = 120$ rows. Second, each **Q** matrix entry can be any of three values 0/1/N as follows. The ***link vector*** for option h of item i is the (i, h) row $\underline{q}_{ih} = (q_{ih1}, \ldots, q_{ihK})$ of **Q**, where each $q_{ihk} = 0/1/N$, and $\underline{q}_{ih}$ specifies that the latent states $\underline{\alpha} = (\alpha_1, \ldots, \alpha_K)$ that are *cognitively most strongly attracted* to option h must satisfy:

$$\text{for each } k = 1, \ldots, K \text{ for which } q_{ihk} \neq N, \ \alpha_k \text{ must match } q_{ihk}.$$

In other words, $\underline{\alpha}$ is cognitively most strongly attracted to option h if $\underline{\alpha}$ satisfies for each k:

- $\alpha_k = 0$, that is, the student *lacks* facet k, if $q_{ihk} = 0$;
- $\alpha_k = 1$, that is the student *possesses* facet k, if $q_{ink} = 1$; and
- the value of α_k does not directly affect the strength of attraction to option h if $q_{ihk} = N$.

For GDCM-MC, the usual dichotomous DCM **Q** matrix notation is extended to include these two major differences. The three labels $0/1/N$ allow the **Q** matrix to express a variety of cognitive conditions for attraction to a given response option. Sometimes, $q_{ihk} = 0$ may be used when lacking a skill enhances the attractiveness of an incorrect option or when lacking a misconception makes a correct option more attractive. We can also use $q_{ihk} = 1$ to indicate that possessing a certain misconception makes a specific incorrect option attractive.

Note that the value $q_{ihk} = 0$ for GDCM-MC (*must lack* facet k) has quite a different meaning from that of $q_{ik} = 0$ in the dichotomous DCM case. For example, for GDCM, if facet k is a skill, then the condition $q_{ihk} = 0$ for an incorrect response option h means that lacking that skill makes it more likely a student will select that response, whereas $q_{ihk} = N$ means that for response option h, whether a student possesses or lacks skill k does not cognitively affect her attraction to option h. Thus, the GDCM-MC value $q_{ihk} = N$ has the same meaning that 0 has in the dichotomous DCM case: Neither possessing nor lacking facet k matters cognitively for option h. Finally, note that some options may be *cognitively neutral*, with $\underline{q}_{ih} = (N, \ldots, N)$.

The GDCM-MC's Mixture Modeling of Cognitive and Guessing Response Behaviors

Given the forced choice imposed by the standard multiple choice question format, it is posited that a typical student strategy for responding to a particular item can be categorized as one of three types: cognitive, guessing, or a hybrid of the two. A ''cognitive strategy'' uses the latent facets of thinking $\underline{\alpha}$ (the pattern of desirable and problematic facets that are possessed or lacked) to either select one answer or eliminate all answers but one. A ''guessing strategy'' selects randomly from available options with equal probability. A ''hybrid strategy'' can be modeled as an initial cognitive step that eliminates some options, followed by a random guess from the remaining options. Discussion of the hybrid strategy is omitted for simplicity.

A student's item response probability is modeled as a mixture of cognitive and guessing strategies, with mixing probabilities that depend on the student's latent state $\underline{\alpha}$. Dropping item i subscripts for convenience, let the event $C_i = C$ denote the use of a cognitive strategy on item i, and $G_i = G = \sim C$ the use of the complementary guessing strategy on item i. The GDCM-MC mixture model is defined for item i as follows:

$$\begin{aligned} P_i(h|\underline{\alpha}) &= P_i(h, C|\underline{\alpha}) + P_i(h, \sim C|\underline{\alpha}) = P_i(h|C, \underline{\alpha})P_i(C|\underline{\alpha}) + P_i(h|G, \underline{\alpha})P_i(G|\underline{\alpha}) \\ &= P_i(h|C, \underline{\alpha})\omega_{i,\underline{\alpha}} + \frac{1}{H_i}\left(1 - \omega_{i,\underline{\alpha}}\right), \end{aligned} \tag{3}$$

where the probability of applying a cognitive strategy to item i is denoted by,

$$\omega_{i,\underline{\alpha}} = P_i(C|\underline{\alpha}). \tag{4}$$

So, $1 - \omega_{i,\underline{\alpha}} = P_i(G|\underline{\alpha})$ is the probability of guessing, and guessing is modeled as $P_i(h|G, \underline{\alpha}) = 1/H_i$. Thus to define a GDCM-MC model, it is seen from (3) that two things must be specified for each i and h: (a) the *cognitive portions* $P_i(h|C, \underline{\alpha})$ and (b) the *mixing probabilities* $\omega_{i,\underline{\alpha}} = P_i(C|\underline{\alpha})$.

The GDCM-MC Cognitive Kernel Functions

Concrete functions must be selected for the cognitive portions $P_i(h|C, \underline{\alpha})$. For example, the dichotomous DINA model functionality could be imported for option h modeling, as is illustrated below. Suppose $F_{ih}(\underline{\alpha}, \underline{\beta}_{ih})$ is a function selected for each row h that brings the desired modeling functionality, where $\underline{\beta}_{ih}$ is whatever vector of parameters may be required by F_{ih}. As $P_i(h|C, \underline{\alpha})$ must be a probability over h, the following is set—suppressing $\underline{\beta}_{ih}$ for simplicity:

$$P_i(h|C, \underline{\alpha}) = \frac{F_{ih}(\underline{\alpha})}{\sum_{h'=1}^{H_i} F_{ih'}(\underline{\alpha})} = \frac{F_{ih}(\underline{\alpha})}{S_{i,\underline{\alpha}}}, \tag{5}$$

$$\text{where } S_{i,\underline{\alpha}} \equiv \sum_{h=1}^{H_i} F_{ih}(\underline{\alpha}). \tag{6}$$

The term *cognitive kernel modeling functions* is used for the functions $F_{ih}(\underline{\alpha})$. Other than capturing the desired modeling functionality, and imposing a monotonicity condition, as is discussed below, the only constraint on the cognitive modeling kernels is that they be non-negative:

$$F_{ih}(\underline{\alpha}) \geq 0. \tag{7}$$

Notice that

$$P_i(h, C|\underline{\alpha}) = P_i(h|C, \underline{\alpha}) P_i(C|\underline{\alpha}) = P_i(h|C, \underline{\alpha}) \omega_{i,\underline{\alpha}} = \frac{F_{ih}(\underline{\alpha}) \omega_{i,\underline{\alpha}}}{S_{i,\underline{\alpha}}}. \tag{8}$$

Thus, to instantiate the GDCM-MC model for item i, one can make any choice of cognitive kernel functions $F_{ih}(\underline{\alpha}) \geq 0$, one for each option h, and the GDCM-MC mixture Equation 3 can be written as,

$$P_i(h|\underline{\alpha}) = P_i(h|C, \underline{\alpha}) \omega_{i,\underline{\alpha}} + \frac{1}{H_i}\left(1 - \omega_{i,\underline{\alpha}}\right) = \frac{F_{ih}(\underline{\alpha}) \omega_{i,\underline{\alpha}}}{S_{i,\underline{\alpha}}} + \frac{1}{H_i}\left(1 - \omega_{i,\underline{\alpha}}\right), \tag{9}$$

with cognitive portion $P_i(h|C, \underline{\alpha}) = F_{ih}(\underline{\alpha}, \underline{\beta}_{ih}) / S_{i,\underline{\alpha}}$, guessing portion $P_i(h|G, \underline{\alpha}) = 1/H_i$, and mixing probabilities $\omega_{i,\underline{\alpha}}$ and $1 - \omega_{i,\underline{\alpha}}$. This model building is illustrated below with DINA and RUM.

Selecting cognitive modeling kernel functions $F_{ih}(\underline{\boldsymbol{\alpha}})$. As building blocks for the cognitive portion $F_{ih}(\underline{\alpha})/S_{i,\underline{\alpha}}$ of the model, a cognitive kernel $F_{ih}(\underline{\alpha})$ must be selected for each option h, to be *monotone increasing* as a function of $\underline{\alpha}$ matching, that is, $F_{ih}(\underline{\alpha})$ should be larger when $\underline{\alpha}$ matches $\underline{q}_{i,h}$ and smaller as more mismatches occur between $\underline{\alpha}$ and $q_{ihk} \neq N$. In fact, there is a generic procedure for importing almost *any* dichotomous DCM to serve as a kernel $F_{ih}(\underline{\alpha})$. Most DCMs are based on mastery of required skills. Any such DCM's mastery of required skills can be converted to matching the GDCM-MC response option link vector's required possession or lack of particular facets. This modeling approach is named the GDCM-MC *penalty-for-mismatch heuristic* for $F_h(\underline{\alpha})$, and is demonstrated next with three concrete examples.

Defining the extended RUM-multiple choice (ERUM-MC) model. The authors define an instance of GDCM-MC, called the *ERUM-MC*, that uses the RUM as cognitive kernel $F_{ih}(\underline{\alpha})$. Recall the

RUM IRF from Equation 2: $F_{RUM,i}(\underline{\alpha}) \equiv P(X_i = 1|\underline{\alpha}) = \pi_i \prod_{k=1}^{K} r_{ik}^{q_{ik}(1-\alpha_k)}$. For option h of item i, with link $\underline{q}_{ih} = (q_{ih1}, \ldots, q_{ihK})$, the ERUM-MC cognitive kernel is defined as follows:

$$F_{ERUM,ih}(\underline{\alpha}) = \pi_{ih} \prod_{k \text{ such that } q_{ihk} \neq N} r_{ihk}^{|q_{ihk}-\alpha_k|}. \tag{10}$$

A penalty-for-mismatch via multiplication by r_{ihk} is applied for every k for which $q_{ihk} \neq N$ and $\alpha_k \neq q_{ihk}$. Notice that such a mismatch can occur in two ways: $(q_{ihk}, \alpha_k) = (1, 0)$ or $(0, 1)$.

For each option h of item i, the ERUM-MC parameters are a π_{ih} parameter and, for each k for which $q_{ihk} \neq N$, an r_{ihk} parameter, as in the RUM. As desired, $F_{ERUM,ih}(\underline{\alpha})$ is nonnegative and is monotonically sensitive to mismatches between elements $q_{ihk} \neq N$ and corresponding α_k. Thus, the ERUM-MC kernel $F_{ERUM,ih}(\underline{\alpha})$ imports the RUM modeling functionality of penalizing separately for different mismatches between non-N q_{ihk} and corresponding elements of α_k.

Defining the extended DINA–multiple choice (EDINA-MC) model. The same penalty-for-mismatch heuristic can be used to import the DINA model of Equation 1 as the GDCM-MC cognitive modeling function $F_{EDINA,ih}(\underline{\alpha})$. First, the DINA is rewritten in an equivalent form as follows:

$$P_{\mathrm{DINA}}(X_i = 1|\underline{\alpha}) = \begin{cases} 1 - s_i \text{ if } \alpha_k = 1 \text{ for all } k \text{ such that } q_{ik} = 1 \\ \\ g_i \text{ if there exists } k \text{ such that } q_{ik} = 1 \text{ and } \alpha_k = 0 \end{cases}. \tag{11}$$

The DINA mastery of required skills is converted to matching all non-N elements of $\underline{q}_{ih}$:

$$F_{\mathrm{EDINA},ih}(\underline{\alpha}) = \begin{cases} 1 - s_{ih} \text{ if } \alpha_k = q_{ihk} \text{ for all } k \text{ such that } q_{ihk} \neq N \\ \\ g_{ih} \text{ if there exists } k \text{ such that } q_{ihk} \neq N \text{ and } \alpha_k \neq q_{ihk} \end{cases}. \tag{12}$$

This non-negative kernel function $F_{\mathrm{EDINA},ih}(\underline{\alpha})$ results in an ‘‘all-or-nothing’’ matching of all non-N elements for EDINA-MC that is analogous to DINA's mastery of required skills.

Extended loglinear cognitive diagnosis model–multiple choice (ELCDM-MC)—The full generality of the GDCM-MC modeling family. The penalty-for-mismatch conversion above of the RUM and DINA works for most dichotomous DCMs. This generality is shown by importing the LCDM model (Henson, Templin, & Willse, 2009). The fundamental quantity used to build the LCDM IRF for item j (see Equation 12, Henson, Templin, & Willse, 2009, p. 198) is,

$$\boldsymbol{\lambda}_j^T \boldsymbol{h}\left(\boldsymbol{\alpha}_i, \boldsymbol{q}_j\right) = \sum_{u=1}^{K} \lambda_{ju}\left(\alpha_u q_{ju}\right) + \sum_{u=1}^{K} \sum_{v>u} \lambda_{juv}\left(\alpha_u \alpha_v q_{ju} q_{jv}\right) + \ldots .$$

The sums are over all one- and two-way interactions, followed by all k-way interaction sums for each k. Each summation is over sets of skills $u = 1, \ldots, K, v > u,$ and so on. Specific parametric coefficients $\lambda_{ju}, \lambda_{juv},$ and so on, provide the particular modeling functionality desired for a given instantiation of the LCDM such as DINA, deterministic input, noisy ‘‘or’’ gate (DINO; Henson & Templin, 2006), RUM, GDM (von Davier, 2008), and so on. Whether a given summand is non-zero for a particular model depends on whether the terms in parentheses, such as $\alpha_u q_{ju}$ and $\alpha_u \alpha_v q_{ju} q_{jv},$ are 1 or 0, that is, only when the $\alpha_u = 1$ match the $q_{ju} = 1$ in conjunctive patterns according to the particular k-way interaction. A given LCDM model can be imported as a GDCM-MC cognitive kernel $F_{ih}(\underline{\alpha})$ by inserting for option h of item j the term

$(1-|\alpha_u - q_{jhu}|)$ for $\alpha_u q_{ju}$, and so on, and using $[(1-|\alpha_u - q_{jhu}|)\cdot(1-|\alpha_v - q_{jhv}|)]$ for a two-way interaction $\alpha_u\alpha_v q_{ju}q_{jv}$, and so on. The sums are over ranges of u or v for which the corresponding $q_{jhu} \neq N, q_{jhv} \neq N$, and so on. Details are omitted. Thus, GDCM-MC can include any DCM that can be realized within LDCM, and that includes all commonly used DCMs. This example supports the full generality of the GDCM family in the following sense. A specific instance of an LCDM is defined in general by imposing particular constraints and conditions on all model coefficients for all one-, two-way (e.g., λ_{juv}), and so on, terms. The above transformation performed on each term in any such specific LCDM instantiation replaces the condition ''mastery of required skills'' by a *matching* condition on all non-N terms in the **Q** matrix. The transformed LCDM instance can then be used as a cognitive kernel function $F_{ih}(\underline{\alpha})$.

Modeling the Cognitive Mixing Probabilities $\omega_{\underline{\alpha}} = P(C|\underline{\alpha})$

There are too many $\omega_{i,\underline{\alpha}}$ to estimate as separate parameters. Instead the $\omega_{i,\underline{\alpha}}$ are chosen to be strongly positively associated with the sizes of the $S_{i,\underline{\alpha}} = \sum_{h=1}^{H_i} F_{ih}(\underline{\alpha})$. Since $S_{i,\underline{\alpha}}$ large means the combined cognitive attractions of $\underline{\alpha}$ across all options h is large, the probability of guessing should be low, and $\omega_{i,\underline{\alpha}}$ should be high. This is accomplished by defining $\omega_{i,\underline{\alpha}}$ as,

$$\omega_{i,\underline{\alpha}} = \min\{1, S_{i,\underline{\alpha}}\}. \tag{13}$$

Setting the $\omega_{i,\underline{\alpha}}$ in this way links the probability of guessing for $\underline{\alpha}$ piecewise linearly to the size of $S_{i,\underline{\alpha}}$. In particular, $S_{i,\underline{\alpha}} \geq 1$ implies that $P(G|\underline{\alpha}) = 1 - \omega_{i,\underline{\alpha}} = 0$, no guessing for $\underline{\alpha}$. This piecewise linear reduction seems a reasonable modeling simplification.

In summary, the GDCM-MC modeling equation for option h (suppressing index i), is,

$$P(h|\underline{\alpha}) = \frac{F_h(\underline{\alpha})}{S_{\underline{\alpha}}}\omega_{\underline{\alpha}} + \frac{1}{H}(1-\omega_{\underline{\alpha}}) = \left\{\begin{array}{ll} F_h(\underline{\alpha}) + (1 - S_{\underline{\alpha}})\frac{1}{H} & \text{if } S_{\underline{\alpha}} < 1 \\ \frac{F_h(\underline{\alpha})}{S_{\underline{\alpha}}} & \text{if } S_{\underline{\alpha}} \geq 1 \end{array}\right\}. \tag{14}$$

GDCM-MC Model Identifiability

Recall that a probability model $P_{\underline{\beta}}$ parameterized by a parameter vector $\underline{\beta}$ such that $P_{\underline{\beta}} = P_{\underline{\beta}'}$ for some $\underline{\beta} \neq \underline{\beta}'$ is said to be non-identifiable. Successful estimation and meaningful interpretation of model parameters require identifiability, and that will depend upon the parametric forms of the selected kernel functions $F_h(\underline{\alpha})$. As with most IRF models, a particular GDCM-MC can be made identifiable by introducing certain constraints that do not interfere with the intended modeling functionality. This is illustrated briefly with the EDINA-MC and ERUM-MC.

Identifiability requirements for the EDINA-MC. Recall that $S_{i,\underline{\alpha}} = \sum_{h'=1}^{H} F_{ih}(\underline{\alpha})$ for the general GDCM-MC of Equation 14. An identifiability issue that involves the scaling of the $S_{i,\underline{\alpha}}$ magnitudes needs to be addressed. Two conditions are necessary for identifiability of the EDINA-MC item-option parameters. It is conjectured that these conditions also are sufficient.

Theorem 1. For the EDINA-MC, the following conditions are necessary for identifiability of the item-option parameters:

$$\min_{\underline{\alpha}} S_{i,\underline{\alpha}} \leq 1 \text{ and } \max_{\underline{\alpha}} S_{i,\underline{\alpha}} \geq 1. \tag{15}$$

Proof. Suppose Equation 15 fails. There are two cases: (i) Assume that $m \equiv \min_{\underline{\alpha}} S_{i,\underline{\alpha}} > 1$. Then, by Equation 13 $\omega_{i,\underline{\alpha}} = \min\{1, S_{i,\underline{\alpha}}\} = 1$ for all α, that is, there is no guessing on item i for any $\underline{\alpha}$. Suppressing item i subscripts, the parameters of the EDINA modeling function $F_h(\underline{\alpha})$ are s_h and g_h for each option h of item i. Define $1 - s'_h = 1 - s_h/m$ and $g' = g_h/m$. Then, $0 \le g'_h, s'_h \le 1$ follows from $0 \le g_h, s_h \le 1$. Let $F'_h(\underline{\alpha})$ denote the EDINA-MC kernels with g'_h and s'_h inserted. Then, $F'_h(\underline{\alpha}) = F_h(\underline{\alpha})/m$ so $S'_{\underline{\alpha}} = S_{\underline{\alpha}}/m$. Also $\min_{\underline{\alpha}} S'_{\underline{\alpha}} = 1$, implying that $\omega'_{\underline{\alpha}} = 1$, and hence by Equation 14, $P'(h|\underline{\alpha}) = F'_h(\underline{\alpha})/S'_{\underline{\alpha}} = F_h(\underline{\alpha})/S_{\underline{\alpha}} = P(h|\underline{\alpha})$. But $s'_h \ne s_h$ and $g'_h \ne g_h$, so identifiability fails.

(ii) Suppose instead that $M \equiv \max_{\underline{\alpha}} S_{\underline{\alpha}} < 1$. Then guessing occurs for every $\underline{\alpha}$. First assume that none of the options are cognitively neutral. It is easy to show that there exists a positive number c that satisfies $c < 1 - g_h$ and $c < s_h$ for every h, and that also satisfies $c < 1 - M/H$. For each h define $g'_h = g_h + c$ and $s'_h = s_h - c$. Define the EDINA kernels $F'_h(\underline{\alpha})$ with parameters g'_h and s'_h, then $F'_h(\underline{\alpha}) = F_h(\underline{\alpha}) + c$, so $S'_{\underline{\alpha}} = S_{\underline{\alpha}} + c \cdot H$. Then,

$$P'(h|\underline{\alpha}) = F'_h(\underline{\alpha}) + \left(1 - S'_{\underline{\alpha}}\right)\frac{1}{H} = F'_h(\underline{\alpha}) + c + \left(1 - S_{\underline{\alpha}} - c \cdot H\right)\frac{1}{H} =$$
$$F_h(\underline{\alpha}) + \left(1 - S_{\underline{\alpha}}\right)\frac{1}{H} = P(h|\underline{\alpha}).$$

That implies that the model is not identified. The case with one or more options unlinked has corresponding $F_h(\underline{\alpha}) = \pi_h$—only one parameter. The same proof goes through. ***QED***

As noted above, the authors conjecture that the necessary conditions (15) are also sufficient to guarantee item-option parameter identifiability for EDINA-MC. It is assumed that the definition and model estimation for EDINA-MC includes the conditions (15) of Theorem 1.

Identifiability requirements for the ERUM-MC. Next, necessary conditions for ERUM-MC item-option parameter identifiability are stated, which the authors conjecture are also sufficient.

Definition. For a GDCM-MC model **Q**, let ***Q_i*** denote the consecutive rows in **Q** pertaining to the options of item *i*. *Simple structure* holds for *i* if each row of ***Q_i*** has at most one non-*N* entry. In other words, each response option is cognitively influenced by at most one facet.

Theorem 2. For **ERUM**-MC for item *i*, the following conditions must hold for identifiability of the item-option parameters:

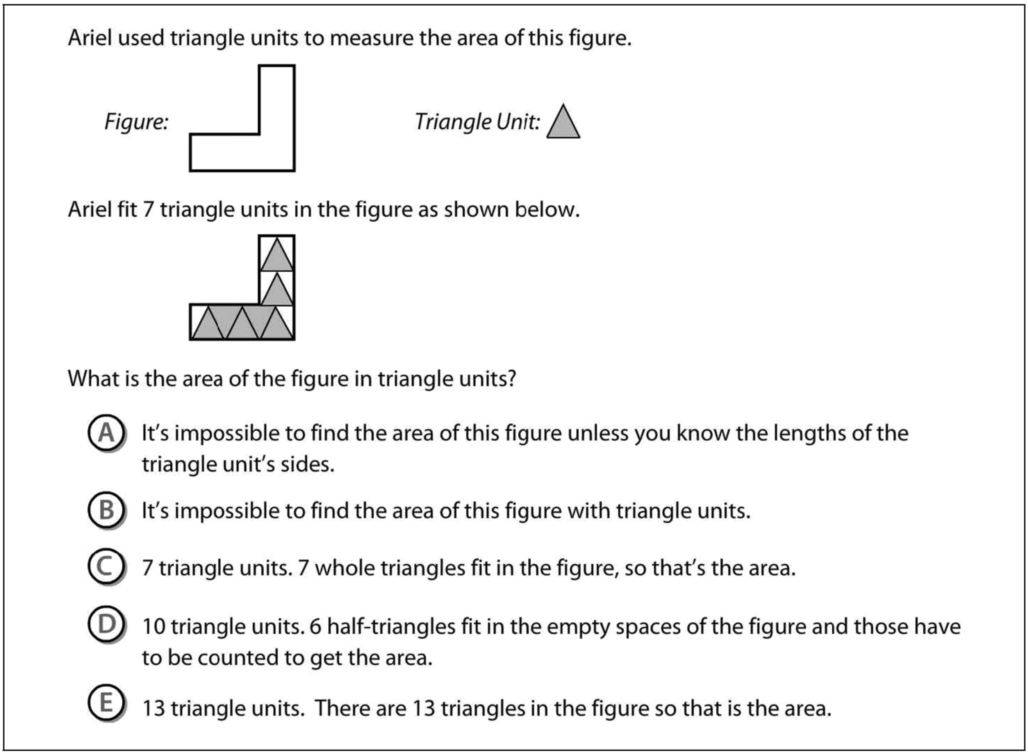
$$\text{(i) } \min_{\underline{\alpha}} S_{i,\underline{\alpha}} \le 1 \text{ and (ii) either simple structure does not hold or } \max_{\underline{\alpha}} S_{i,\underline{\alpha}} \ge 1. \tag{16}$$

Proof. The proof for ERUM-MC is similar to that of Theorem 1 above. Details are omitted.

The definition of an ERUM-MC model is assumed to include the conditions (16) of Theorem 2. These conditions for EDINA-MC and ERUM-MC are noted to be easy to check for a given model and are included as required in model estimation algorithms and software.

Other Approaches to Option-Based Scoring in the Research Literature

Other approaches that link multiple choice options to latent facets of thinking include the Ordered Multiple Choice Model (Briggs, Alonzo, Schwab, & Wilson, 2006), the multiple choice–deterministic input, noisy ‘‘and’’ gate (MC-DINA) model (de la Torre, 2009), and the Scaling Individuals and Classifying Misconceptions (SICM) model (Bradshaw & Templin,



Ariel used triangle units to measure the area of this figure.

Figure:
Triangle Unit:

Ariel fit 7 triangle units in the figure as shown below.

What is the area of the figure in triangle units?

Ⓐ It's impossible to find the area of this figure unless you know the lengths of the triangle unit's sides.

Ⓑ It's impossible to find the area of this figure with triangle units.

Ⓒ 7 triangle units. 7 whole triangles fit in the figure, so that's the area.

Ⓓ 10 triangle units. 6 half-triangles fit in the empty spaces of the figure and those have to be counted to get the area.

Ⓔ 13 triangle units. There are 13 triangles in the figure so that is the area.

Figure 1. Example Item 856 from the Diagnostic Geometry Test (Masters, 2014).

2014). These models differ substantially from the GDCM-MC—for example, none has a guessing component. Future work will compare GDCM-MC to these models.

Description of a Real Data Application

An analysis is described of posttest data from the Diagnostic Geometry Assessment (DGA) section on Geometric Measurement (GM; Masters, 2012, 2014; Masters & Chapman, 2011). Figure 1 presents an example item from the Diagnostic Geometry Test. In this example, the correct answer (D) is linked to a GM skill, two incorrect responses (C, E) are linked to a misconception, and Options A and B are neutral, i.e. not linked.

The DGA GM posttest consisted of 10 multiple choice items, with 48 response options, and was designed to diagnose two facets: (a) a conceptual understanding of area measure that is called ''D'' for desired thinking and (b) a problematic facet of thinking about area called ''M'' for ''misconception.'' Briefly, D was understanding how the area of a plane figure is related to a tiling of the figure by congruent, non-overlapping tiles of unit area, and particularly how the area formula $L \times W$ for a rectangle is related to such tiling. M represented a partial understanding that area is related to tiling, together with specific, systematic misunderstandings such as incorrectly allowing area to be determined with overlaps or gaps among tiles, or non-congruence of tiles. The selection of D and M and the associated assessment design were based on the learning and cognitive literatures and were intended to be formatively useful for middle school teachers.

The GDCM-MC codings for the options of the above sample item are shown in Table 1. The codings for Options A and B were identical, as were those for C and E. Options with identical

Table 1. Portion of **Q** Matrix for Sample Item Given in Figure 1.

Original **Q** matrix			Compressed **Q** matrix		
Response	Skill	Misconception	Option	D	M
A	N	0	0	N	0
B	N	0	1	0	1
C	0	1	2	1	N
D	1	N			
E	0	1			

Table 2. **Q** Matrix for DGA, GM, Posttest.

Item	option	D	M
840	0	1	N
	1	0	1
	2	0	0
844	0	0	0
	1	0	1
	2	1	N
	3	N	N
848	0	N	N
	1	0	1
	2	1	N
	3	0	0
	4	0	N
852	0	0	N
	1	1	N
	2	0	1
856	0	N	0
	1	0	1
	2	1	N
860	0	0	1
	1	1	N
	2	N	0
864	0	1	N
	1	0	1
	2	0	0
868	0	0	1
	1	1	N
	2	0	0
872	0	0	N
	1	1	N
	2	0	0
880	0	0	0
	1	1	N
	2	0	1

Note. Correct answers are shaded. Options are numbered 0, 1, 2, and so on, instead of a, b, c, and so on. DGA = Diagnostic Geometry Assessment; GM = Geometric Measurement.

link vectors were thus combined to form a compressed **Q** as in Table 1 (see DiBello, Henson, & Stout, 2014a, 2014b; DiBello, Henson, Stout, & Roussos, 2014, for further details on the **Q** matrix coding). The compressed **Q** matrix used for analysis consisted of 33 options and two facets of thinking and is given in Table 2.

Model Performance on Real Data

Convergence of model parameter estimates. ERUM-MC analysis of Diagnostic Geometry Assessment (DGA) section on Geometric Measurement (GM) posttest data are reported from a national sample of 1,765 middle school students. Bayesian model expected a posteriori (EAP) model estimation is performed, described below, by running two Markov Chain Monte Carlo (MCMC) analyses, each of length 10,000 with burn-in 7,000. Parameter estimation convergence was satisfactory overall.

Model-data fit. As a measure of fit, an index of predictive accuracy for each response option was calculated as follows. For each item and option (i, h), the ERUM-MC model probability $P_i(h|\underline{\hat{\alpha}}, \hat{\underline{\beta}}_{ih})$ of choosing option h given an examinee's estimated state $\underline{\hat{\alpha}}$ and estimated item-option parameters $\hat{\underline{\beta}}_{ih}$ was compared with the corresponding empirical probability $\mathrm{N}_{\underline{\hat{\alpha}},i}(h)/\mathrm{N}_{\underline{\hat{\alpha}}}$, where $\mathrm{N}_{\underline{\hat{\alpha}}}$ is the number of examinees classified as $\underline{\hat{\alpha}}$, and $\mathrm{N}_{\underline{\hat{\alpha}},i}(h)$ is the number of $\underline{\hat{\alpha}}$ examinees selecting option h for item i. Fit index $D_{i,h}$ was defined as an average discrepancy:

$$D_{i,h} = \frac{\sum_{\underline{\hat{\alpha}}} N_{\underline{\hat{\alpha}}} \left| P_i(h|\underline{\hat{\alpha}}) - \frac{N_{\underline{\hat{\alpha}},i}(h)}{N_{\underline{\hat{\alpha}}}} \right|}{N},$$

where N is the number of examinees in the sample, and for simplicity, it is assumed that each examinee answers every item. Table 3 shows the results for the GM posttest.

A small value of the fit index constitutes evidence of good fit for the associated response option. The fit of most options was quite good, with modest discrepancies for Options 0 and 1 of Item 4 and Options 0 and 2 of Item 5. Even those cases were well within reasonable limits. The fit indices provide useful information for test redesign, and for critiquing the **Q** matrix.

Model discriminability. An index of discriminability is defined for item i and facet h that used estimated parameter values from the real data analysis to determine for each facet k how well the model differentiated between $\alpha_k = 0$ and $\alpha_k = 1$:

$$d_{i,k} = \left(\frac{1}{2}\right) \max_{\underline{\beta}} \left\{ \sum_h \left| \hat{P}\left(X_i = h|\alpha_k = 1, \underline{\beta}\right) - \hat{P}\left(X_i = h|\alpha_k = 0, \underline{\beta}\right) \right| \right\}.$$

The max is over $\underline{\beta} = (\beta_1, \ldots, \beta_{k-1}, \beta_{k+1}, \ldots, \beta_K)$, all $K - 1$ element latent vectors except k, and $\hat{P}$ represents the IRF with item parameters estimated from the real data. It can be shown that each $d_{i,k} \leq 1$.

In Table 3, larger index values for a given facet indicate higher discrimination power. The table shows that each item does a better job of discriminating M than D, and half of the items had poor discrimination indices for D, namely, $d_{i,D} < 0.100$. It is important to note that the low discrimination values for D occurred within a context of reasonably high fit. The analyses provide useful information for evaluating diagnostic functioning of items and facets.

Parameter recovery mean absolute deviations (MADs) and correct classification rates (CCRs). The authors estimated πs, rs, and αs for each $k = D, M$, and parameter p_k, the population proportion of students with $\alpha_k = 1$. Simulated data were generated from the estimated ERUM-MC model.

Table 3. Fit Indices for Each Option of the DGA GM Posttest and Discrimination Power for Facets D and M for Each Item.

Item	Fit index for each response option					Discrimination power	
	$h = 0$	$h = 1$	$h = 2$	$h = 3$	$h = 4$	D	M
1	0.026	0.018	0.008			0.064	0.459
2	0.002	0.015	0.015	0.004		0.112	0.574
3	0.062	0.008	0.040	0.025	0.033	0.164	0.535
4	0.128	0.118	0.014			0.091	0.433
5	0.109	0.028	0.082			0.503	0.595
6	0.013	0.046	0.046			0.313	0.520
7	0.010	0.011	0.007			0.081	0.548
8	0.008	0.013	0.006			0.102	0.558
9	0.012	0.012	0.006			0.008	0.487
10	0.017	0.006	0.023			0.015	0.455

Note. DGA = Diagnostic Geometry Assessment; GM = Geometric Measurement.

Table 4. MADs for Model Parameters and CCRs for D and M.

MAD for πs	MAD for rs	MAD for p_ks	CCR for D	CCR for M
0.027	0.120	0.018	0.89	0.80

Note. MAD = mean absolute deviation; CCR = correct classification rate.

Assuming reasonably good fit, as shown above, the procedure's performance on this *simulated* data gives an indication of how well the procedure is performing on actual data. Model parameters and simulee $\hat{\underline{\alpha}}$ were estimated from the simulated data and compared with true values. Table 4 reports the MAD for each type of ERUM-MC parameter πs, rs, and p_ks, and CCRs for D and M. The MADs were quite good for the πs and p_ks and acceptable for the rs. The CCRs were excellent given the test length of 10.

The MCMC Estimation Algorithm for the GDCM-MC

A high-level description of the MCMC estimation algorithm is provided for ERUM-MC, an instantiation of the GDCM-MC. For other GDCM-MC model instantiations, such as EDINA-MC and ELCDM-MC, the estimation algorithm is similarly determined by incorporating the model's **Q** and the resulting kernel functions $F_h(\underline{\alpha})$ into the estimation algorithm.

As defined above, the ERUM-MC item parameters for option h of item i consisted of a single π_{ih} parameter and parameters r_{ihk} for all facets k for which $q_{ihk} \neq N$. As for all GDCM-MC models, the ERUM-MC is defined by specifying the kernel functions $F_h(\underline{\alpha})$ (see Equation 10). The MCMC estimation procedure used a Metropolis–Hastings within Gibbs (MHG) sampling approach (Gelman, Carlin, Stern, & Rubin, 2004). Recall, in broad strokes, when certain easily verifiable theoretical assumptions are valid MCMC provides an asymptotically consistent estimate of the posterior distribution given the data, for example, in the case including $\hat{P}(\beta|x)$. A simulation-based sampling procedure produces repeated values of β that when combined (after an initial burn-in period) provide an estimate of the posterior distribution of β (Gelman, Carlin, Stern, & Rubin, 2004). Next, the MHG sampling approach for the ERUM-MC item parameters and its examinee α profiles are briefly described.

MHG for Item Parameters

The MHG new candidate distribution and its associated candidate acceptance probability for each item parameter require specification of the parameter's prior distribution, the candidate sampling distribution, and the likelihood of the data. The generic notations $f(\beta)$ and $q(\beta^t|\beta^{t-1})$ are used for the prior and candidate distributions, respectively, of a single item parameter β^t in time step t. For the ERUM-MC, each β^t will be either π_{ih} or one of the r_{ihk}.

Item parameter priors—$f(\pi_{ih})$ and $f(r_{ihk})$—were assumed to be independent uniform distributions between 0 and 1: $\pi_{ih} \sim U(0,1)$ and $r_{ihk} \sim U(0,1)$. The moving window (Templin, 2004) approach was used as the candidate distribution $q(\beta^t|\beta^{t-1})$ for all item parameters. Specifically, given β^{t-1}, parameter β^t was selected from a uniform distribution $U(LB_{t-1}, UB_{t-1})$ with $LB_{t-1} = \max(0, \beta^{t-1} - \delta)$ and $UB_{t-1} = \min(1, \beta^{t-1} + \delta)$, where δ was a user specified value whose choice controls acceptance rates. A uniform candidate distribution was used:

$$q\left(\beta^t|\beta^{t-1}\right) = \frac{1}{UB^{t-1} - LB^{t-1}} \text{ for } LB^{t-1} \leq \beta^t \leq UB^{t-1}.$$

The likelihood function of data given item parameters was the usual locally independent product based on the ERUM-MC model presented in Equation 10.

MHG for Examinee Latent States

Candidate $\underline{\alpha}^t$ for all examinees is generated in time point t in several stages. First, time point t Bayesian hyperparameters (Gelman et al., 2004), $\boldsymbol{\rho}^t$ and $\underline{\tau}^t$, were generated. Each element of correlation matrix $\boldsymbol{\rho}^t$ was generated (via random candidate generation and probability of acceptance) by using a uniform prior and moving window approach centered around the same matrix element at time point $t-1$, while preserving positive definiteness of the matrix (for details, see Hartz, 2001). Let marginal prior $p_k = P(\alpha_k = 1)$ and p_k^t its accepted value at time t. Updated p_k^t values were produced using a uniform prior $p_k \sim U(0,1)$ and a moving window on the value p_k^{t-1} at time $t-1$. A continuous $\underset{\sim}{\alpha} \sim MVN_K(0, \boldsymbol{\rho}^t)$ was then drawn, and 0/1 α_k was defined by,

$$\alpha_k = \left\{ \begin{array}{c} 0, \tilde{\alpha}_k < \tau_k \\ 1, \tilde{\alpha}_k \geq \tau_k \end{array} \right\},$$

where the p_k^t values determine the associated threshold vector $\underline{\boldsymbol{\tau}}^t$ as shown in Equation 17 below. A given set of cutoff values $\underline{\tau}$ and correlation matrix $\boldsymbol{\rho}^t$ define a distribution on the discrete latent facet state with specified proportions p_k of 1 and 0s for each facet as follows:

$$p(\underline{\alpha}) = \int_{LB_1}^{UB_1} \cdots \int_{LB_K}^{UB_K} MVN(\underline{0}, \boldsymbol{\rho}; \underline{\tau}), \text{ where} \tag{17}$$

$$UB_k = \left\{ \begin{array}{c} \infty \text{ if } \alpha_k = 1 \\ \tau_k \text{ if } \alpha_k = 0 \end{array} \right\}, \text{ and} \tag{18}$$

$$LB_k = \left\{ \begin{array}{c} \tau_k \text{ if } \alpha_k = 1 \\ -\infty \text{ if } \alpha_k = 0 \end{array} \right\}. \tag{19}$$

At convergence, the EAP value for each parameter was taken as the parameter's estimate (see, for example, Hartz, 2001; Templin, 2004; Henson et al., 2009).

Table 5. Four Simulated Matrices for Simulated Data Analyses.

Q matrix	No. of items	No. of skills measured	No. of misconceptions
Q Matrix A	10	3	3
Q Matrix B	20	3	3
Q Matrix C	30	3	3
Q Matrix D	30	4	6

A Simulation Study of the ERUM-MC

A simulation study was performed to evaluate model performance when using relatively short tests with varying sample sizes, using the ERUM-MC. Four **Q** matrices were crossed with three sample sizes. Twenty replications were performed for each of the 12 conditions.

Four **Q** matrices were created to resemble realistic tests using four-option multiple choice items. The four **Q**s differed in number of items and facets, as given in Table 5. The four **Q**s crossed with three samples sizes (500; 1,000; 2,000) made 12 simulation conditions.

Given a **Q** matrix and a specified number of examinees, both item parameters and examinee facet patterns were generated, as follows: the $\pi_{ih}, r_{ihk},$ and p_k were drawn from uniform distributions, $\pi_{ih} \sim U(0.65, 0.95), r_{ihk} \sim U(0.10, 0.50),$ and $p_k \sim U(0.45, 0.65)$. The π_{ih} and r_{ihk} ranges were chosen to produce a ''medium'' level of discriminability similar to various real and simulated data analyses. The p_k determined the threshold vector $\underline{\boldsymbol{\tau}},$ as defined above. The correlation matrix was set to $\boldsymbol{\rho} = \boldsymbol{I}_{K \times K},$ the $K \times K$ identity matrix—no association between facets, making correct classification harder. Examinee facet profiles were drawn from the discretized normal distribution in Equation 17 using $\boldsymbol{\rho} = \boldsymbol{I}_{K \times K}$ and $\underline{\tau}$. Examinee response vectors were simulated from the ERUM-MC using the generated parameters.

For each Test A through D, the same **Q** was fixed across three sample sizes and across all 20 replications within each corresponding test-by-sample size condition. New item parameters, simulees, and simulated item response vectors were generated for each of the 240 replications.

The ERUM-MC was estimated using the algorithm described above. Model performance was evaluated using MADs and CCRs. Given a set of I true parameters β and corresponding estimated $\hat{\beta},$ the MAD was computed as,

$$\text{MAD} = \frac{\sum |\beta - \hat{\beta}|}{I}. \tag{20}$$

The MAD was used to summarize parameter recovery for π_{ih} parameters and r_{ihk} parameters. For examinees $j = 1, \ldots, J$ and facets $k = 1, \ldots, K,$ the CCR was the observed proportion $p(\alpha_{jk} = \hat{\alpha}_{jk})$ of facet state 1/0 estimates that agreed with truth. Table 6 summarizes the MADs and CCRs for the π and r parameter estimations across the 20 replications for each condition.

Table 6 shows that as the number of items or sample size increased, item parameter estimation and classification accuracy improved. That is, (a) the MAD's decreased (i.e., improved) within any column from Test A (10 items) to B (20 items) to C (30 items); (b) for fixed number of 30 items, the MAD's increased as complexity of **Q** increased from Test C (30 items, three skills, three misconceptions) to Test D (30 items, four skills, six misconceptions); (c) the CCR's increased (i.e., improved) with increasing test length from Test A, to B, to C; (d) CCR's decreased with increasing **Q** matrix complexity from Test C to Test D; (e) within any row with a fixed test length and **Q** matrix, the MADs improved as sample sizes increased from 500 to 1,000 to 2,000 examinees; and (f) within any row, the CCRs increased slightly or remained constant as sample size increased. Overall, parameter estimation and classification were successful.

Table 6. A Summary of the MADs for π and *r* Parameter Estimation and CCRs.

Test (No. of items/No. of skills/ No. of misconceptions)	π MADs			*r* MADs			CCRs		
	No. of examinees			No. of examinees			No. of examinees		
	500	1,000	2,000	500	1,000	2,000	500	1,000	2,000
Test A (10/3/3)	0.17	0.14	0.13	0.17	0.13	0.10	0.74	0.76	0.77
Test B (20/3/3)	0.13	0.10	0.09	0.11	0.08	0.05	0.89	0.91	0.91
Test C (30/3/3)	0.12	0.10	0.08	0.09	0.07	0.05	0.94	0.94	0.95
Test D (30/4/6)	0.14	0.11	0.10	0.12	0.10	0.07	0.84	0.85	0.85

Note. MAD = mean absolute deviation; CCR = correct classification rate.

Future Research

In ongoing research, the authors are investigating a variety of application issues and performing further real and simulated data studies to provide practical guidelines and limitations across a range of assessment contexts. For example, option-based scoring for well-designed tests should provide more cognitive information about students for a given test length than right-wrong scoring. A well-designed 20-item option scored test with four options per item might be as diagnostically informative as a dichotomously score test with 40 or 60 items. Research on this is underway.

A second focus is developing practical information to guide design of assessments able to take full advantage of the GDCM-MC modeling capabilities to optimize diagnostic information and to support evaluation and improvement of assessments at the item-option level.

GDCM-MC for Short-Answer Open-Ended Questions

The GDCM-MC modeling approach can readily be applied to short-answer open-ended questions designed for ‘‘nominal rubric’’ scoring. In other words, items can be designed so that open-ended responses can be scored for evidence that indicates possession or lack of one or more targeted skills and/or misconceptions. Such open-ended questions effectively can be treated as if they were multiple choice questions with facet-linked response options. Current research is investigating data from a real test constructed with such purpose-designed open-ended questions.

Discussion and Summary

This article introduces the GDCM-MC and its four distinguishing features: (a) expanded latent space that includes both desirable and problematic facets of thinking, (b) expanded **Q** matrix with a row for each response option and three-valued *0/1/N* elements, (c) a modeled guessing component, and (d) a fully general modeling framework that can incorporate most dichotomous DCMs, including DINA, RUM, GDM, DINO, LCDM. For the EDINA-MC and ERUM-MC, necessary conditions were proven for item-option parameter identifiability, conditions that are conjectured to also be sufficient, as supported by small MADs for the π and *r* parameters in the ERUM-MC simulation study.

The article includes an analysis of real data from the DGA and a realistic simulation study of the ERUM instantiation of the GDCM-MC. The real data analysis demonstrated satisfactory MCMC convergence of model parameter estimates, a high degree of model-data fit, good parameter estimation accuracy, and acceptable discriminability for facet M. Although the discriminability index for facet D was quite low for 5 of the 10 items, the high degree of fit indicates that the low discriminability values likely were not related to the model but rather to the definitions of the facets and/or the structure of the **Q** matrix. In general, such findings are useful for test and **Q** matrix design, development, evaluation, and improvement.

The simulation studies showed that the MADs and CCRs varied as expected according to test length, sample size, and **Q** matrix complexity. Parameter recovery as measured by the MADs was quite good. Classification accuracy as measured by the CCRs was high except for test length 10. As classroom formative assessment uses will likely be ''low stakes,'' a CCR of 0.74 to 0.76, as shown in Table 6 for test length 10, should be useful for a teacher using a realistic short test length (such as 10) for classroom use.

An important formative assessment goal is to provide accurate finer grained diagnostic information, even for fairly short tests common in classroom settings. The GDCM-MC demonstrated a high enough level of performance for useful classroom formative assessment. The flexibility and breadth of GDCM-MC's option-based modeling of multiple choice testing and the reported successful real and simulated data results suggest that well-designed multiple choice tests can be very informative about student learning. In particular, the more complex testing demands resulting from new standards such as the Common Core State Standards in Mathematics (Achieve, n.d.; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) and the Next Generation Science Standards (see NGSS, n.d.) call for strengthened diagnostic testing capabilities, including a need for diagnostically informative multiple choice and short-answer open-ended questions, and the targeting of problematic as well as desired facets of thinking. The GDCM-MC approach promises to be applicable for such tests.

Author's Note

Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the funding agencies.

Acknowledgments

The authors thank Jessica Masters of Innovation Lab, Measured Progress, for access to the DGA data and for several conversations about the DGA. They specially thank Louis Roussos of Measured Progress for ongoing interest and conversations about the GDCM-MC model and for collaborating on the **Q** matrix coding and real data analysis. They also acknowledge comments from two anonymous reviewers.

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: This material is based on work partially supported by the National Science Foundation under Grants 0918552, 0920242, and 0815065, and by the Institute of Education Sciences Grants R305A100475 and R305D140023.

References

Achieve. (n.d.). *Next generation science standards*. Available from www.nextgenscience.org

Bradshaw, L., & Templin, L. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*, 403-425.

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

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

DiBello, L. V., Henson, R. A., & Stout, W. F. (2014a, July). *A family of generalized diagnostic classification models for multiple-choice option-scored assessments*. Paper presented at the annual meeting of the Psychometric Society, Madison, WI.

DiBello, L. V., Henson, R. A., & Stout, W. F. (2014b, April). *A new diagnostic model for multiple-choice option-based scoring with applications*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia, PA.

DiBello, L. V., Henson, R. A., Stout, W. F., & Roussos, L. A. (2014, April). Psychometric model for diagnostic classification for multiple-choice option-based scoring: Application to a diagnostic classroom assessment instrument. In J. Masters (Chair), *Diagnostic assessment: Recent advances from psychometric modeling to classroom applications*. Symposium conducted at the annual meeting of the American Education Research Association, Philadelphia, PA.

DiBello, L. V., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics 26* (pp. 979-1030). Amsterdam, The Netherlands: Elsevier.

DiBello, L. V., Stout, W., & Roussos, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-390). Hillsdale, NJ: Lawrence Erlbaum.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London, England: Chapman & Hall.

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

Hartz, S. M. (2001). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practicality* (Doctoral dissertation). University of Illinois at Urbana–Champaign.

Henson, R. A., & Templin, J. (2006, June). *The DINO: A disjunctive model for skills assessment*. Paper presented at the annual meeting of the Psychometric Society Meeting, Montreal, Quebec, Canada.

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

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

Masters, J. (2012, April). *The validity of concurrently measuring students' knowledge and misconception related to shape properties*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC.

Masters, J. (2014, April). *The diagnostic geometry assessment system: Results from a randomized controlled trial*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Masters, J., & Chapman, L. (2011, April). *Measuring geometric measurement ability and misconception with a single scale*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Miller, R. L., Streveler, R., Olds, B., Chi, M., Nelson, M., & Geist, M. (2006, June). *Misconceptions about rate processes: Preliminary evidence for the importance of emergent conceptual schemas in thermal and transport sciences*. Proceedings of the American Society for Engineering Education Annual Conference (electronic), Chicago, IL.

Minstrell, J. (1992). Facets of students' knowledge and relevant instruction. In R. Duit, F. Goldberg, & H. Niedderer (Eds.), *Research in physics learning: Theoretical issues and empirical studies* (pp. 110-128). Kiel, Germany: IPN.

Minstrell, J. (2001). Facets of students' thinking: Designing to cross the gap from research to standards-based practice. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications for professional, instructional, and everyday science* (pp. 369-394). Mahwah, NJ: Lawrence Erlbaum.

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

Next Generation Science Standards. (n.d.). Available from http://www.nextgenscience.org/

Pellegrino, J. W., DiBello, L. V., James, K., Jorion, N., & Schroeder, L. (2011, October). *Concept inventories as aids for instruction: A validity framework with examples of application*. Proceedings of Research in Engineering Education Symposium, Madrid, Spain.

Pellegrino, J. W., DiBello, L. V., Miller, R. L., Streveler, R. A., Schroeder, L., & Stout, W. F. (2013, April). Conceptual underpinnings of concept inventories. In J. W. Pellegrino (Chair), *Evaluating the validity of concept inventories as aids for STEM teaching and learning*. Symposium conducted at the annual meeting of American Educational Research Association, San Francisco, CA.

Roussos, L. A., DiBello, L. V., Stout, W. F., Hartz, S. M., Henson, R. A., & Templin, J. H. (2007). The fusion model skills diagnosis system. In J. Leighton & M. Gierl (Eds), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275-318). New York, NY: Cambridge University Press.

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

Steif, P. S., & Dantzler, J. A. (2005). A statics concept inventory: Development and psychometric analysis. *Journal of Engineering Education*, *94*, 363-371.

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

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