Journal of Classification 30 (2013)
DOI: 10.1007/s00357-013-9129-4

Measuring the Reliability of Diagnostic Classification Model Examinee Estimates

Jonathan Templin

The University of Georgia

Laine Bradshaw

The University of Georgia

Abstract: Over the past decade, diagnostic classification models (DCMs) have become an active area of psychometric research. Despite their use, the reliability of examinee estimates in DCM applications has seldom been reported. In this paper, a reliability measure for the categorical latent variables of DCMs is defined. Using theory- and simulation-based results, we show how DCMs uniformly provide greater examinee estimate reliability than IRT models for tests of the same length, a result that is a consequence of the smaller range of latent variable values examinee estimates can take in DCMs. We demonstrate this result by comparing DCM and IRT reliability for a series of models estimated with data from an end-of-grade test, culminating with a discussion of how DCMs can be used to change the character of large scale testing, either by shortening tests that measure examinees unidimensionally or by providing more reliable multidimensional measurement for tests of the same length.

Keywords: Diagnostic classification models; Cognitive diagnosis; Reliability; Classification; Psychometrics.

We would like to thank Terry Ackerman, Allan Cohen, Jeff Douglas, Robert Henson, John Poggio, and John Willse for their helpful comments and critiques of the concepts and text presented in this paper. Complete syntax for running all analyses herein and resulting program output are available at the first author's website.

This research was funded by National Science Foundation grants DRL-0822064; SES-0750859; and SES-1030337. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Corresponding Author's Address: Jonathan Templin, Department of Educational Psychology, The University of Georgia, 323 Aderhold Hall, Athens, GA 30602; email: jtemplin@uga.edu.

1. Measuring the Reliability of Diagnostic Model Examinee Estimates

Over the past decade, diagnostic models have become an active area of psychometric research. Diagnostic classification models (DCMs; e.g., Rupp, Templin, and Henson 2010), also known as models for cognitive diagnosis (e.g., Rupp and Templin 2008), are psychometric models that characterize examinee responses to test items through the use of categorical latent variables. These categorical latent variables, most frequently called attributes, represent the mastery status (e.g., master or non-master) of an examinee for each dimension the test purports to measure. DCMs differ from more traditional psychometric approaches such as item response theory (IRT) in that instead of providing examinees with an estimate of each latent variable that falls along a continuous scale (assumed to be on an interval level of measurement but with measurement error), DCMs give examinee estimates in the form of a classification based on the mastery status of each latent variable. Attribute mastery statuses represent ordered categorical states, yielding an ordinal level of measurement for examinees under DCMs. Although DCMs do not provide the ability to pinpoint an examinee's location on a scale as IRT models do, their use has been driven by the often unstated idea that ordinal, classification-based measurement would be more reliable and therefore could allow for more dimensions to be measured from a test.

Although DCMs have been the subject of numerous research studies, many psychometric concerns about their properties remain, giving pause to analysts considering the use of DCMs in their testing programs and research applications. Perhaps the most worrisome concern is that DCM applications seldom report the reliability at which examinee attributes are measured (Sinharay and Haberman 2009). It is our belief that this lack of reporting reliability stems from reliability not being well-defined for DCMs. As stated within the *Standards for Educational and Psychological Testing* (1999; Standard 2.1), reliability is of great importance to any psychometric model or application. Therefore, this paper derives a definition of reliability for the categorical attributes of a DCM designed to be comparable with reliability methods used in other psychometric models (e.g., IRT). Presenting the results of a theoretical evaluation, a simulation study, and an empirical application, we then show how DCM examinee estimates *uniformly* have higher reliability than analogous IRT model examinee estimates, meaning DCMs measure latent traits more precisely than analogous IRT models.

Throughout this paper we refer to the questions of a test as *items*, test takers as *examinees*, and the latent traits measured by the test as *attributes*. We restrict our discussion of DCMs and IRT models to cases where items are dichotomous (i.e., scored as correct or incorrect), although our results are not limited to dichotomous data. We first present a brief description of the psychometric characteristics of DCMs and their distributional assumptions for examinee attributes. Following that, we provide a section comparing DCMs to IRT models and showing how reliability is defined in an IRT context. We then define a comparable reliability metric for DCMs and show how it can be computed. Next we compare DCMs to IRT models with respect to reliability of examinee estimates by presenting a simulation study and an analysis of an end-of-grade large scale test from a Midwestern state. The paper concludes with a discussion of how the properties of reliability of examinee estimates in DCMs help to define when the use of DCMs is appropriate and how DCMs could be used to change the structure of large scale testing.

2. Diagnostic Classification Models

Diagnostic classification models, also known as *cognitive diagnosis models* (e.g., Leighton and Gierl 2007) or *multiple classification latent class models* (Maris 1999), are confirmatory latent class models that characterize the relationship of observed responses to a set of categorical latent variables, commonly called *attributes*. For an examinee *e*, the attributes (representing an attribute pattern $\boldsymbol{\alpha}_e = [\alpha_{e1},\alpha_{e2},\ldots,\alpha_{eA}]$) are binary indicators of the mastery of a set of A attributes representing multiple dimensions, combinations of which are thought to underlie an examinee's response to an item. To indicate the attributes measured by each item, an item-by-attribute *Q-matrix* is constructed, with $\boldsymbol{q}_i = [q_{i1}, q_{i2}, \ldots, q_{iA}]$ for item *i*. Similar to a factor pattern matrix in a confirmatory factor model, Q-matrix indicators are binary – either the item measures an attribute ($q_{ia} = 1$) or it does not ($q_{ia} = 0$). Although the DCMs we predominantly use for this paper are models that use two-category (binary) latent attributes, our results are not limited to the two-category case (e.g., Templin 2004; von Davier 2005).

Recent research on DCMs has focused on developing model parameterizations that specify how entries in the Q-matrix relate to item responses. General diagnostic models have been developed based on log-linear models with latent classes, with the Log-linear Cognitive Diagnosis Model (LCDM; Henson, Templin, and Willse 2009a) providing a general approach for diagnostic modeling based on the General Diagnostic Model (von Davier 2005). We focus our discussion on the LCDM because the model allows for both non-compensatory and compensatory links between attributes for items of the same test. Furthermore, the LCDM subsumes most commonly used DCMs, models such as the Deterministic Inputs Noisy And Gate model (or DINA; Haertel 1989; Junker and Sjitsma 2001; Macready and Dayton 1977), the Noisy Inputs Deterministic And Gate model (or NIDA; Maris 1999), the Reduced Reparameterized Unified Model (or RUM; Roussos, DiBello, Stout, Hartz, Henson, and Templin 2007), the Deterministic Inputs Noisy Or Gate model (or DINO; Templin and Henson 2006), the Noisy Inputs Deterministic Or Gate model (or NIDO; Templin 2006), and the Compensatory Reparameterized Unified Model (or C-RUM; Hartz 2002). We also note that the methods we develop for quantifying reliability apply to all DCMs, not just the LCDM.

To help describe the process by which examinee mastery status is mapped onto item responses, we now describe the characteristics of the LCDM. The LCDM specifies the conditional probability that examinee *e* with attribute pattern $\boldsymbol{\alpha}_e$ provides a correct response to item *i* as:

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

where $\boldsymbol{q}_i$ is the set of Q-matrix entries for item *i*. The intercept parameter, $\lambda_{i,0}$, represents the log-odds of a correct response to item *i* for an examinee who has not mastered any of the Q-matrix indicated attributes for the item. $\boldsymbol{\lambda}_i$ represents a (2^A-1) x 1 sized vector of LCDM parameters for item *i*, with $\boldsymbol{h}(\boldsymbol{q}_i, \boldsymbol{\alpha}_e)$ being a (2^A-1) x 1 sized vector of indicators as to whether or not a parameter is present for an examinee on an item. The LCDM kernel function is:

$$\begin{aligned}\boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{q}_i, \boldsymbol{\alpha}_e) = &\sum_{a=1}^{A} \lambda_{i,1,(a)} \alpha_{ea} q_{ia} \\ &+ \sum_{a=1}^{A-1} \sum_{b>a} \lambda_{i,2,(a,b)} \alpha_{ea} \alpha_{eb}\, q_{ia} q_{ib} + \cdots.\end{aligned} \tag{2}$$

For an item *i*, this function includes all main effects ($\lambda_{i,1,(a)}$ for an attribute *a*) and interactions between attributes (e.g., $\lambda_{i,2,(a,b)}$ is a two-way interaction for an attribute *a* and an attribute *b*), where the level of the effect is denoted by the second subscript of the effect.

Therefore, for a Q-matrix with *A* attributes, the first *A* elements of $\boldsymbol{h}(\boldsymbol{q}_{i,}\ \boldsymbol{\alpha}_e)$ are indicators for the *A* main effect parameters: each attribute α_{ea} multiplied by its corresponding Q-matrix entry q_{ia}. For example, the first element corresponds to the main effect for Attribute 1 and is given by $\alpha_{e1} q_{i1.}$ This element is present on the response function only if Attribute 1 is measured by the item *and* examinee *e* has mastered the attribute. The second set of elements includes all two-way interactions (for items measuring and examinees mastering both of the attributes). Thus, these elements of $\boldsymbol{h}(\boldsymbol{q}_{i,}\ \boldsymbol{\alpha}_e)$ include the multiplication of two α values and two entries in the Q-matrix. For example, the two-way interaction between Attributes 1 and 2 is present when $\alpha_{e1}\alpha_{e2}q_{i1}q_{i2}$ is equal to one. The remaining linear combinations of $\boldsymbol{h}(\boldsymbol{q}_{i,}\ \boldsymbol{\alpha}_e)$ are defined as all possible three-way interactions (for items measuring three attributes) up to a final A-way interaction. Constraints are placed on the elements of $\boldsymbol{\lambda}_i$ so that the probability of a positive response increases as an examinee masters additional Q-matrix indicated attributes (for more details see Henson et al. 2009).

To demonstrate how the LCDM is specified, consider an item written to measure two attributes, resulting in Q-matrix entries for the item of $q_{i1} = 1$ and $q_{i2} = 1$. Conditional on an examinee's attribute pattern $\boldsymbol{\alpha}_e = [\alpha_{e1}, \alpha_{e2}]$, the LCDM provides the following item response function for the item (where the Q-matrix indicators are removed from notation as they each have a value of one):

$$P(X_{ei} = 1|\boldsymbol{\alpha}_e) = \frac{\exp\left(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1} + \lambda_{i,1,(2)}\alpha_{e2} + \lambda_{i,2,(1,2)}\alpha_{e1}\alpha_{e2}\right)}{1 + \exp\left(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{e1} + \lambda_{i,1,(2)}\alpha_{e2} + \lambda_{i,2,(1,2)}\alpha_{e1}\alpha_{e2}\right)}. \qquad (3)$$

The LCDM item parameters are analogous to the differing levels of effects found in an analysis of variance (ANOVA) where attributes are dummy coded (0/1) with a logistic link for dichotomous data. The attribute pattern, $\boldsymbol{\alpha}_e = [\alpha_{e1}, \alpha_{e2}]$, uses dummy-coding to indicate whether examinee e has mastered attribute a ($\alpha_{ea} = 1$) or has not mastered attribute a ($\alpha_{ea} = 0$). The intercept for the item ($\lambda_{i,0}$) represents the log-odds of a correct response for an examinee in the reference group who has not mastered either attribute. The main effects ($\lambda_{i,1,(1)}$ and $\lambda_{i,1,(2)}$) increase the log-odds of a correct response given mastery of each respective attribute. Finally, the two-way interaction between the two attributes ($\lambda_{i,2,(1,2)}$) allows the log-odds of a correct response to change given an examinee's mastery of both attributes.

3. Attribute Distributional Assumptions in Diagnostic Classification Models

Whereas the item response function of the LCDM specifies the link between the latent attributes an examinee has mastered and his or her performance on an item, also of importance is the assumed distribution of the attributes in the DCM. Specifically, when a test measures multiple latent attributes, the lower moments of the distribution of attributes characterize the association between pairs of attributes and the marginal frequency with which an attribute is mastered in the population of examinees. Because attributes in DCMs are categorical latent variables, their assumed distribution represents the probability an examinee from a population has a given attribute pattern (or $P(\boldsymbol{\alpha}_e = \boldsymbol{\alpha}_p)$). A general categorical distribution that maps onto the binary attributes commonly used by DCMs is that of the multivariate Bernoulli distribution (or MVB; e.g., Maydeu-Olivares and Joe 2005). For DCMs with more than two-category attributes, the analogous extension is the multivariate multinomial distribution. For a test measuring *A* binary attributes, there are a total of 2^A possible attribute patterns. The MVB distribution provides the probability an examinee at large has a given attribute pattern p, or π_p:

$$\begin{gathered}\boldsymbol{\alpha}_p = [\alpha_{p1}, \dots, \alpha_{pA}], \qquad \alpha_{pa} \in \{0,1\}, \qquad p = 1, \dots, 2^A \\ \pi_p = P(\boldsymbol{\alpha}_p)\,. \end{gathered} \tag{4}$$

Of importance in reporting an analysis of a test with a DCM is the marginal proportion of examinees mastering an attribute in a population, or p_a, which can be found by:

$$\begin{bmatrix} p_1 \\ \vdots \\ p_A \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_{2^A} \end{bmatrix}^T \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_{2^A} \end{bmatrix} = \begin{bmatrix} [\alpha_{11}, \alpha_{12}, \dots, \alpha_{1A}] \\ \vdots \\ [\alpha_{2^A1}, \alpha_{2^A2}, \dots, \alpha_{2^AA}] \end{bmatrix}^T \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_{2^A} \end{bmatrix}. \tag{5}$$

When considered marginally, an attribute *a* is said to have a Bernoulli distribution with probability of mastery p_a, or $\alpha_a \sim B(p_a)$. The distributional assumptions of the latent attributes are what distinguish DCMs from IRT models.

4. Diagnostic Classification Model Examinee Estimates

Examinee estimates resulting from DCMs come in the form of a (posterior) probability of an examinee having each possible attribute pattern, which we denote $\hat{\boldsymbol{\alpha}}_{ep}$. Because there are 2^A possible attribute patterns, $\hat{\boldsymbol{\alpha}}_{ep}$ is of size 2^A x 1, with the elements summing to one. $\hat{\boldsymbol{\alpha}}_{ep}$ is found by the combination of item responses and by the distributional parameters of the attributes from the MVB distribution. For $\hat{\alpha}_{ep}$, a given element of $\hat{\boldsymbol{\alpha}}_{ep}$ representing the probability an examinee *e* has attribute pattern p, the examinee attribute probability estimate is found by:

$$\begin{aligned}&\hat{\alpha}_{ep} \\ &= \frac{\pi_p \prod_{i=1}^{I}\left(P(X_{ei} = 1|\boldsymbol{\alpha}_{ep})^{X_{ei}}\right)\left(1 - P(X_{ei} = 1|\boldsymbol{\alpha}_{ep})^{1-X_{ei}}\right)}{\sum_{c=1}^{2^A}\pi_c \prod_{i=1}^{I}(P(X_{ei} = 1|\boldsymbol{\alpha}_c)^{X_{ei}})(1 - P(X_{ei} = 1|\boldsymbol{\alpha}_c)^{1-X_{ei}})},\end{aligned} \tag{6}$$

where $P(X_{ei} = 1|\boldsymbol{\alpha}_{ep})$ is given by the DCM item response function as provided for the LCDM in Equations (1) and (2).

Examinees are more often provided with marginal probabilities of attribute mastery (which we denote as $\hat{p}_{ea}$), rather than the whole-pattern analog (or $\hat{\boldsymbol{\alpha}}_{ep}$). The marginal probability of attribute mastery can be found using Equation (5) and exchanging the population-level probability of attribute pattern possession with that for the examinee:

$$\begin{bmatrix}\hat{p}_1\\ \vdots \\ \hat{p}_A\end{bmatrix} = \begin{bmatrix}\boldsymbol{\alpha}_1\\ \vdots\\ \boldsymbol{\alpha}_{2^A}\end{bmatrix}^T \begin{bmatrix}\hat{\alpha}_{e1}\\ \vdots \\ \hat{\alpha}_{e2^A}\end{bmatrix} = \begin{bmatrix}[\alpha_{11},\alpha_{12},\dots,\alpha_{1A}]\\ \vdots\\ [\alpha_{2^A1},\alpha_{2^A2},\dots,\alpha_{2^AA}]\end{bmatrix}^T \begin{bmatrix}\hat{\alpha}_{e1}\\ \vdots\\ \hat{\alpha}_{e2^A}\end{bmatrix}. \tag{7}$$

To compute our measure of reliability of an attribute in a DCM, the examinee-level marginal probability of attribute mastery, $\hat{p}_{ea}$, is needed.

5. Latent Attribute Reliability in Diagnostic Classification Models

The initial concept of reliability comes from classical test theory where the reliability of a test refers to the squared correlation between an examinee's true score on a test and an examinee's observed score on a test. Reliability can also be thought of as the ratio of true-score variance to total variance (true-score variance plus error variance):

$$\rho = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}. \tag{8}$$

Latent variable models attempt to estimate the "true score" of an examinee, but on a different metric than the test score (the total score from classical test theory). As such, to derive our notions of reliability, we adopt the premise of Lord (1980), that "*true score ξ and ability θ are the same thing expressed on different scales of measurement*" (p. 46). Further, in DCMs, latent attributes are categorical, meaning the variance of the latent attribute and the variance of error are not independent. This causes difficulty in constructing reliability from a traditional point-of-view as direct estimates of these variances are not obtainable.

At the heart of reliability is the assumption that as the error variance decreases relative to the variance of the latent attribute, the precision of measurement is higher. Although feasibly impossible due to examinee memory and test carry over, if an examinee were to be given the same test a second time, reliability would characterize the relationship between the examinee estimates of the two tests. In this hypothetical situation, a perfectly reliable test (one with zero error variance) would yield identical estimates for the examinee. This is analogous to basic measurement of physical objects. For instance, measuring the length of a pencil with the same ruler under the same conditions (e.g., room temperature) will yield the same length as the ruler is perfectly reliable for measuring length. Using this conceptualization of reliability, our measure of reliability seeks to capture how consistent an examinee's estimate from a DCM will be over hypothetically repeated observations. As such, the calculation of the DCM reliability measure is enabled by simulating repeated testing occasions through repeated draws from an examinee's posterior distribution.

For DCMs, the distribution of an attribute a is Bernoulli with probability of mastery p_a (i.e., $P(\alpha_{ea} = 1) = p_a$). For an examinee e, the mean of the estimated distribution is $\hat{p}_{ea}$, the estimated marginal probability that examinee e has mastered attribute a as found by Equation (7). Similarly, the variance of the estimate is the resulting variance of a Bernoulli variable, or $(1 - \hat{p}_{ea})\,\hat{p}_{ea}$. Analogous to the physical measurement scenario of measuring the length of a pencil with a ruler twice, given two hypothetical administrations of the same test an examinee e's estimate of $\hat{p}_{ea}$ would not change for an attribute a. As such, we can calculate the probability that an examinee would be given the same mastery status estimate of the latent attribute (the same score under classical test theory) under these two administrations. This probability is based on the estimate of $\hat{p}_{ea}$, which is central to our estimate of reliability for DCMs.

Imagine a scenario where an examinee takes the same test twice, but with no memory of the test after the first administration, making the tests independent administrations. For a binary attribute, there are four possible mastery status combinations an examinee can be given on the two tests. Because of the assumption of repeated administrations of the same test, the marginal probability of an attribute a being mastered on either test is equal, or $P(\alpha_{ea_1} = 1) = P(\alpha_{ea_2} = 1) = \hat{p}_{ea}$ (where the subscript for attribute a represents the test). As the tests were assumed to be given independently, the probability of observing any *combination* of attribute mastery status for the pair of tests is found by the product of the marginal probabilities. Specifically, $P(\alpha_{ea_1} = 1;\ \alpha_{ea_2} = 1) = \hat{p}_{ea}\hat{p}_{ea}$, $P(\alpha_{ea_1} = 1;\ \alpha_{ea_2} = 0) = \hat{p}_{ea}(1 - \hat{p}_{ea})$, $P(\alpha_{ea_1} = 0;\ \alpha_{ea_2} = 1) = (1 - \hat{p}_{ea})\hat{p}_{ea}$, and $P(\alpha_{ea_1} = 0;\ \alpha_{ea_2} = 0) = (1 - \hat{p}_{ea})(1 - \hat{p}_{ea})$. Because of the independence of the two hypothetical administrations, for any given examinee, the correlation between α_{ea_1} and α_{ea_2} is zero. Across examinees in a sample, however, the correlation is non-zero, and represents an estimate of reliability of the DCM attribute. Therefore, we use this aggregated contingency table as the core of our reliability measure for DCMs.

6. Diagnostic Classification Model Reliability

To calculate reliability for an attribute under a DCM, a general approach using the contingency table of a hypothetical second test administration is used. Although shown for binary attributes, the following three-step process can be used for binary attributes and attributes with more than two mastery statuses, relying upon the correlation of mastery statuses between two hypothetical independent administrations of the same test:

1. *For each attribute a and examinee e, calculate the estimated attribute mastery $\hat{p}_{ea}$ probability using Equation (7).*

   Estimation of any latent-class based DCM will work in this context, however, the LCDM is used in our paper.

2. *Create the replication contingency table (size 2 x 2 for binary attributes):*

   Across a sample of *N* examinees, the replication contingency table comes from the sum of each individual examinee's probabilities for each possible mastery status. For binary attributes this table has four cells (where the subscript on *a* represents the test administration number and *N* is the number of examinees in the sample):

$$P(\alpha_{.a_1} = 1;\ \alpha_{.a_2} = 1) = \frac{\sum_{e=1}^{N} \hat{p}_{ea}\hat{p}_{ea}}{N}, \tag{9}$$

$$P(\alpha_{.a_1} = 1;\ \alpha_{.a_2} = 0) = \frac{\sum_{e=1}^{N} \hat{p}_{ea}(1-\hat{p}_{ea})}{N}, \tag{10}$$

$$P(\alpha_{.a_1} = 0;\ \alpha_{.a_2} = 1) = \frac{\sum_{e=1}^{N} (1-\hat{p}_{ea})\hat{p}_{ea}}{N}, \tag{11}$$

$$P(\alpha_{.a_1} = 0;\ \alpha_{.a_2} = 0) = \frac{\sum_{e=1}^{N} (1-\hat{p}_{ea})(1-\hat{p}_{ea})}{N}. \tag{12}$$

3. *Calculate the attribute reliability using the tetrachoric correlation of $\alpha_{.a_1}$ and $\alpha_{.a_2}$ (or polychoric correlation for attributes with more than two categories).*

   Because the Pearson correlation coefficient is bounded above -1 and below 1 for two-by-two contingency tables where the marginal proportions are not both 0.5 (a scenario that is likely for most applications of DCMs), the tetrachoric (or, in the case of attributes with more than two categories, polychoric) correlation coefficient is used as the metric of test-retest reliability.

7. Item Response Models

Parametrically, IRT models have many variants: differing parametric forms (discrimination/difficulty or slope/intercept), link functions (normal ogive or logistic), and inclusion or omission of asymptotic parameters as limiting regions of the item characteristic curve (i.e., 2-parameter logistic/3-parameter logistic models). For didactic reasons, we restrict our discussion of IRT models to those that are formulated as slope/intercept models with logistic link functions and without any asymptotic parameters. Such models provide a direct link between IRT models and DCMs. We note, however, that the reliability concepts discussed in the next section emanate from and apply to all types of IRT models, not just those presented here.

IRT models relate a set of continuously-valued latent variables (as denoted by $\boldsymbol{\theta}_e$ for examinee *e*) to the item responses made by an examinee. To show the communalities of IRT models with DCMs, we use the same notation for both, differing only for the latent variable (the term ability for $\boldsymbol{\theta}_e$ in IRT for the term attribute for $\boldsymbol{\alpha}_e$ in DCMs). The distributional assumptions of the latent variables distinguish IRT models from DCMs. In IRT applications, $\boldsymbol{\theta}_e$ is assumed to follow a multivariate normal distribution (MVN) with a zero mean vector and covariance matrix $\boldsymbol{\Sigma}_\theta$. For model identification when measurement of the latent variables is the purpose, the diagonal elements of $\boldsymbol{\Sigma}_\theta$ are often set to one, making $\boldsymbol{\Sigma}_\theta$ a correlation matrix between latent variables. Often in educational measurement, a single dimension is most commonly estimated, where, analogously, the variance of the latent variable is set to one. When assuming $\boldsymbol{\theta}_e$ is MVN, and substituting $\boldsymbol{\theta}_e$ for $\boldsymbol{\alpha}_e$, the LCDM from Equations (1) and (2) thus describes a confirmatory multidimensional item response model (MIRT) with all possible item-level interactions between latent variables. The Q-matrix, a label not commonly associated with MIRT models, provides the confirmatory structure of the model, allowing items to measure specific sets of the latent variables represented by $\boldsymbol{\theta}_e$. In common applications of confirmatory MIRT models, the latent variable interaction terms are not present, leaving $\boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{q}_i, \boldsymbol{\theta}_e)$ to only be the linear combination of latent variables with the main-effect item parameters of the model:

$$\boldsymbol{\lambda}_i^T \boldsymbol{h}(\boldsymbol{q}_i, \boldsymbol{\theta}_e) = \sum_{a=1}^{A} \lambda_{i,1,(a)} \theta_{ea} q_{ia} \,. \tag{13}$$

Such confirmatory MIRT models are said to have a compensatory or additive structure for $\boldsymbol{\theta}_e$, in that having low levels of one or more θ_{ea} (for a latent variable *a*) would not necessarily lead to a low item response probability as high levels of other measured latent variables could compensate for such a deficiency.

Estimation of examinee latent variables in IRT typically uses marginal maximum likelihood. Similar to estimation of latent attributes in DCMs, IRT latent variable estimation uses information from the test (the joint likelihood of the items, expressed as a product) and information from the structural components of the multivariate normal distribution (mean vector and covariance matrix). The resulting point estimates of the latent variables are found through iterative algorithms that maximize the joint likelihood function, as no closed form solution exists. Asymptotically, the resulting estimates of the latent variables are assumed to follow a normal distribution marginally, with the mean being the location of the point estimate and the standard deviation being the estimated standard error of the estimate, as provided by the inverse of the test information function for a given latent variable location (e.g., Birnbaum 1968). We capitalize on this property of IRT latent variable estimates in forming our estimate of reliability for an IRT model across a test.

8. Latent Variable Reliability in Item Response Models

As with the categorical latent attributes estimated in DCMs, we will consider IRT reliability from the latent variable metric, θ. Further, because we are interested in comparing the reliability for the continuous latent variables in IRT and the categorical latent attributes in DCMs, we develop a measure of reliability analogous to test-retest reliability. Although the distributions of the latent variables/attributes differ between the two models, our measure of reliability provides roughly comparable reliability estimates for the differing models, as it can be used to summarize how precisely a test measures an examinee's latent variable(s)/attribute(s) regardless of the type of distribution assumed by the latent trait. As with our reliability metric in DCMs, we will again rely upon the notion of how reliability would function if an examinee would take two independent administrations of the same test.

One of the desirable properties of IRT models has been that the standard error attributed to the latent variable measured by a test is conditional on the point estimate of the latent variable. As such, tests can be constructed to provide more precise measurement for regions of the scale determined to be important. This uncertainty is reflected in the standard error of an examinee's latent variable estimate, $\sigma_{\hat{\theta}_e}$. In IRT, from a single test, a latent variable, $\hat{\theta}_e$, and standard error, $\sigma_{\hat{\theta}_e}$ are estimated for each examinee. The normal distribution of an examinee's latent variable is provided by these two parameters: $\hat{\theta}_e$ is the mean of the distribution and $\sigma_{\hat{\theta}_e}$ is the standard deviation. Under our two independent test scenario, for the second test, the distribution of an examinee's ability should remain the same, namely $N\left(\hat{\theta}_e, \sigma_{\hat{\theta}_e}\right)$.

In a real-world testing situation, each examinee's responses to the items of a test are known (observed), which in turn provide the examinee's estimated latent variable ($\hat{\theta}_e$) using the IRT model. The standard error for this latent variable is calculated from the test information function (e.g., de Ayala 2009). However, the distribution of θ across a population of examinees is unknown. As such, we must use a simulation-based resampling approach to arrive at an analogous IRT reliability estimate:

1. *For each examinee e and each continuous latent variable a, determine* $\hat{\theta}_{ea}$ *and* $\sigma_{\hat{\theta}_{ea}}$.

2. *Sample at random from the set of examinees who took the test by drawing an examinee's estimated latent variable* ($\hat{\theta}^*_{ea}$) *and its associated standard error* ($\sigma_{\hat{\theta}^*_{ea}}$).

    This process is similar to that of a bootstrap and is used to generate a distribution of θ_e from the sample of examinees taking the test. The drawn values are referred to as $\hat{\theta}^*_{ea}$ and $\sigma_{\hat{\theta}^*_{ea}}$.

3. *Draw two independent values* $(\theta_{a_1}, \theta_{a_2})$ *from* $N(\hat{\theta}^*_{ea}, \sigma_{\hat{\theta}^*_{ea}})$.

    These represent two plausible values for θ_e, drawn from the posterior distribution of the latent variable. They will represent the independent estimates from taking the test twice and are analogous to the 2 x 2 contingency table formed in DCM reliability.

4. *Repeat steps 2 and 3, with replacement.*

    Repeated draws from the empirical posterior distribution of θ_e allows us to approximate the joint distribution of $(\theta_{a_1}, \theta_{a_2})$, provided the number of draws is sufficiently large. For our results, we chose 100,000 repeated draws.

5. *Calculate IRT reliability by computing the Pearson correlation between the simulated set of* $(\theta_{a_1}, \theta_{a_2})$.

    If a large sample is drawn, this correlation will converge to the test-retest reliability.

If an examinee's latent variable estimate has a large standard error, the two simulated draws in step 3 will likely be very different, causing a low correlation indicating a low reliability. In contrast, if the standard error is small, the draws will be similar, causing the correlation/reliability to be high. Under the extreme case where examinees are measured without error (i.e., the test has a perfect reliability), the two draws will be identical across all replications, yielding a correlation of 1.0, indicating perfect reliability. Thus, this correlation provides a measure of how reliable or consistent the score is; a high positive correlation indicates that examinees are expected to score the same on the two tests, meaning the score is reliable. We note that if the standard error $\sigma_{\hat{\theta}_e}$ was constant across the distribution of the latent attribute (as is the case in latent variable models for continuous, normally distributed data, such as confirmatory factor models), the correlation would represent an intraclass correlation coefficient, and we would not need the resampling procedure to produce an estimate of reliability.

9. Consequences of Diagnostic Model Reliability

To demonstrate how DCMs compare to IRT models in terms of the reliability of examinee estimates, we present four results. Throughout each, a common theme remains: using the metrics of reliability we define, the latent attributes from DCMs are more reliably estimated than the latent variables from IRT models. First, we report DCM and IRT model reliability estimates using theoretical properties of a common IRT model – the Rasch model. Second, a larger simulation study is described where multiple tests were generated varying the number of latent variables and number of items measuring each latent variable. The third and fourth examples demonstrate our reliability result with real-world data by describing the reliability of multiple models applied to the same data set (Henson, Templin, and Willse 2009b) and then determining how many items would be needed by the same test to achieve the same reliability in a unidimensional multicategory DCM as a unidimensional IRT model. For brevity of presentation, we include the methods of analysis and results under separate headings. We conclude the paper by discussing the ramifications for use of DCMs and IRT models and what types of information can be obtained from each.

10. Theoretical Results

As a first demonstration of reliability in DCMs, we compared the reliability of a unidimensional DCM with that of the reliability of the unidimensional IRT Rasch model for tests of a varying number of items. The Rasch model was chosen to provide a basic model for comparison; we expect our results will be similar for other item response models. Our aim was to determine reliability if estimation error was removed from an analysis. Therefore, we used known item parameters for varying sizes of tests. Furthermore, we wished to restrict our analysis to the unidimensional case to determine a baseline level of reliability for both IRT models and DCMs under tests of varying lengths. We label this section "theoretical" due to the lack of item parameter estimation—we are using "known" parameters to avoid estimation error.

Under the Rasch model, the probability of a correct response to item i by examinee e is a function of the value of the latent variable for examinee e, θ_e, and the difficulty for the item i, or b_i, as shown with the scaling constant 1.7 (e.g., Hambleton, Swaminathan, and Rogers 1991):

$$P(X_{ei} = 1|\theta_e) = \frac{\exp\left(1.7(\theta_e - b_i)\right)}{1 + \exp\left(1.7(\theta_e - b_i)\right)}. \tag{14}$$

For our study, we assumed θ_e followed a (standard) normal distribution with a zero mean and unit standard deviation.

In order to compare reliability in IRT models and DCMs, a DCM analogous to the Rasch model must be created. Because our comparison involves a two-category unidimensional DCM, we must first define a mapping of the categorical latent variable in the DCM onto the continuous latent variable of the Rasch model. Previous research dictates that the categorical latent variables in DCMs approximate the continuous latent variables in IRT models (e.g., Haberman, von Davier, and Lee 2008). Furthermore, a simple simulation study (not reported) using a test created with an IRT model but estimated with a two-category DCM revealed that the DCM approximately divided the θ distribution in half (half of the simulated examinees were considered masters and the other half were considered non-masters), with a value of zero being the location of where masters (α = 1 if $\theta > 0$) and non-masters ($\alpha = 0$ if $\theta \leq 0$) were set. Therefore, we set the proportion of masters for our example to be $p_a = 0.5$ and to correspond to having a mastery cut-point at $\theta = 0$.

Continuing with the DCM as an approximation to IRT approach, the analogous item parameters for the DCM were found by taking the logit (natural logarithm of the odds of a correct item response) theoretical item difficulty for masters and non-masters for each item. For non-masters this became:

$$\lambda_{i,0} = logit\left(\int_{-\infty}^{0} P(X_i = 1|\theta)\, d\theta\right), \tag{15}$$

and for masters this became:

$$\lambda_{i,0} + \lambda_{i,1,(1)} = logit\left(\int_{0}^{\infty} P(X_i = 1|\theta)\, d\theta\right), \tag{16}$$

with each DCM parameter $(\lambda_{i,0}, \lambda_{i,1,(1)})$ being found from these two probabilities.

Having defined the Rasch model and the DCM analog, we now discuss the method we used to compare reliability in both. The key determination for us was the number of items in a test, varying the number of items from 3 to 100. As the number of items measuring a latent attribute/variable grows, the reliability for a latent attribute/variable will increase.

In the development of the theoretical study, an additional concern was the shape of the information function for the continuous latent variable, θ. Because differing information functions will have a differential peak where reliability is highest (and $\sigma_{\hat{\theta}}$ is lowest), we chose two different information functions: one flat across θ (and therefore having the same reliability at all points) and one peaked at the mean of θ (and therefore having a higher reliability for the most frequently observed values of θ).

For the flat information function, we selected item difficulty values that equally divided the interval between -2 and 2. For the peaked information function, we chose item difficulties that were found using normal deviates corresponding to equal percentiles of a normal distribution. The information functions were then used to provide estimates of $\sigma_{\hat{\theta}}$.

Once the item difficulty values were set, we then assessed the reliability for measurement of θ under the Rasch model by drawing 100,000 random values from a standard normal distribution, giving an analog to $\hat{\theta}$ and, consequently, the associated standard error for each, $\sigma_{\hat{\theta}}$. From these quantities we then built our estimate of test-retest reliability, using the method described previously. We note that we did not simulate the item responses for each associated θ because, using the test information function, we were able to derive an associated $\sigma_{\hat{\theta}}$.

To replicate the Rasch model simulated test and examinees, for each test length and information function condition we generated a unidimensional DCM where the assumed proportion of examinees having the single attribute was set to 0.5 ($p_a = 0.5$). For each item difficulty generated for the Rasch model, we created analogous intercept and main effect parameters for the DCM using Equations (15) and (16). We then generated 100,000 simulated examinees. For each examinee, we generated a simulated test using the item parameters of the model, and then estimated the probability of attribute mastery using Equation (6). From the probabilities of attribute mastery, we estimated the DCM attribute reliability using the method described previously.

The results of our study show that, for all tests created (regardless of the type of information function or number of items), the reliability of the DCM was markedly higher than the reliability for the IRT model. A graphical depiction of the results for the simulation study can be found Figure 1. Furthermore, Table 1 lists the numbers of items needed to reach a set of reliability milestones. Specifically, the IRT model needed 50 items to reach a reliability of 0.85 and 78 items to reach a reliability of 0.9 (results consistent with Mislevy, Beaton, Kaplan, and Sheehan 1992). By comparison, the DCM needed seven items to reach 0.85 reliability and nine items to reach a reliability of 0.9.

11. Simulation Study

Although the theoretical analysis provides an insight into the nature of reliability in DCMs and how it compares to IRT models, it was incomplete in that it focused on a single dimension and did not incorporate errors into the measurement of item parameters as is often the case in practice. Therefore, we conducted a small simulation study varying two factors: the number of dimensions measured (one or three) and the number of items measuring each dimension (5, 10, or 15). For each condition, the data were

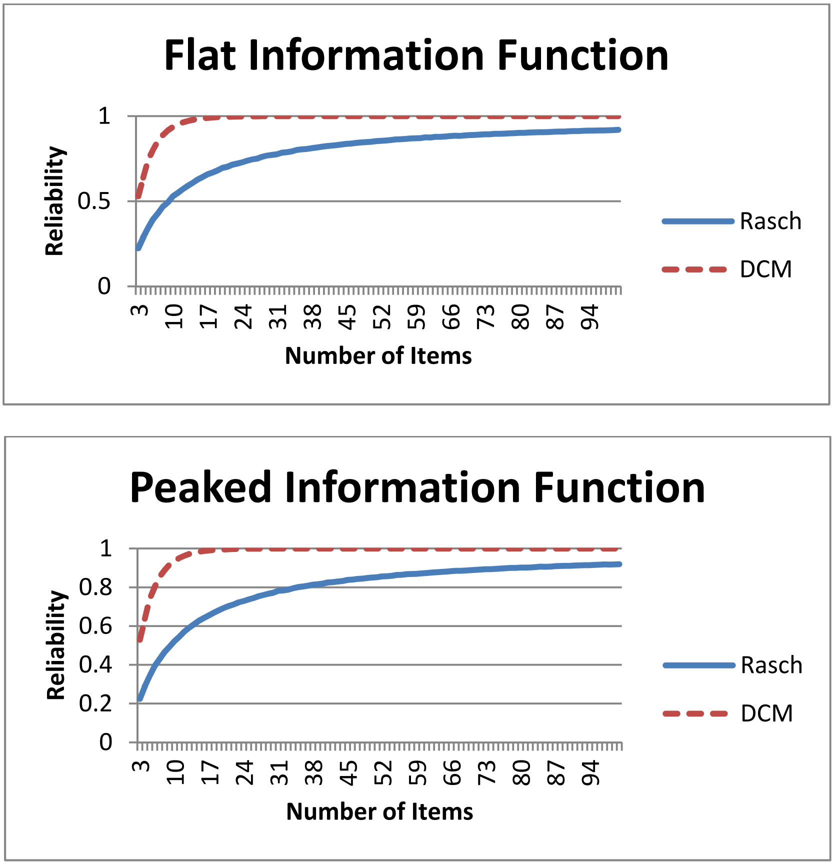

Figure 1. Theoretical Reliability for DCM and IRT Rasch Models Under Flat and Peaked Test Information

Table 1. Theoretical Number of Items Needed to Reach Reliability Milestone

	IRT – Rasch Model		DCM – Rasch Model Analog	
Reliability	Flat Information	Peaked Information	Flat Information	Peaked Information
0.50	10	10	3	3
0.60	14	14	4	4
0.70	21	21	5	5
0.80	35	34	6	6
0.85	50	49	7	7
0.90	78	78	9	9

generated using a continuous IRT model with 4,000 simulated examinees (meant to mimic current conditions of large scale testing programs). Multidimensional models were generated with items only measuring a single latent variable, yielding so-called simple structure. The distribution of the latent variables was multivariate normal with unit variances and correlations of .7 (for conditions with three latent attributes). A total of 100 tests were generated, each with a new set of item parameters. Intercept parameters were generated from uniform distributions ranging from -1 to 1 ($\lambda_{i,0} \sim U(-1,1)$), and slope parameters were generated from uniform distributions ranging from 0.5 to 1.5 ($\lambda_{i,1,(1)} \sim U(0.5,1.5)$). For each simulated data set, two models were estimated using the Mplus package (Muthén and Muthén 2010): a DCM (with two categories per dimension) and an IRT model.

The simulation study results showed a similar trend to those from the theoretical analysis: DCM reliability was uniformly higher than the IRT model analog. *In fact, there was never a case for any condition, simulated data set, or dimension where the IRT model had a higher reliability than the DCM.* Table 2 shows the mean reliabilities for each type of model and simulation condition, which is also provided graphically in Figure 2. For tests generated with a single latent attribute, the average reliability for DCMs was .7 for 5 items, .8 for 10 items, and .85 for 15 items. For the IRT model, these reliabilities were .58 for 5 items, .61 for 10 items, and .66 for 15 items. Similar results held for the three latent attribute conditions with DCMs having reliabilities of .76 for 5 items per dimension, .83 for 10 items per dimension, and .87 for 15 items per dimension. This is compared to the multidimensional IRT model with reliabilities of .60, .66, and .69, respectively. We note that the gain in reliabilities for the multidimensional conditions comes from the estimates of the correlation between attributes (in both DCMs and IRT models), which allows for shared information to help in the estimation of all traits. In sum, DCMs have higher latent variable reliability than comparable IRT models.

12. Empirical Data Analyses

Finally, to demonstrate the reliability result from a real-world test, we present the reliability estimates for a set of varying models applied to data from an active testing program—a large scale test from a Midwestern state. Specifically, the test used is an end-of-grade (EOG) test of reading ability and contains a total of 73 multiple choice items. Of these 73 items 55 items measured reading ability (i.e., understanding the meaning of words and phrases) and the remaining 18 items measured comprehension (i.e., understanding the characters and purpose of a passage). Data from a total of 2,318 students were used in the estimation of all models described

Table 2. Simulation Study Results: Reliability of Uni- and Multidimensional IRT and DCM for 100 Random Tests

		IRT			DCM		
Dimen-sions	Items Per Dimension	Mean (SD)	Min	Max	Mean(SD)	Min	Max
1	5	0.58 (0.01)	0.55	0.61	0.70 (0.03)	0.62	0.77
1	10	0.61 (0.01)	0.60	0.65	0.80 (0.02)	0.74	0.84
1	15	0.66 (0.01)	0.64	0.69	0.85 (0.02)	0.81	0.89
3	5	0.60 (0.02)	0.57	0.65	0.76 (0.02)	0.79	0.71
3	10	0.66 (0.01)	0.65	0.68	0.83 (0.01)	0.81	0.85
3	15	0.69 (0.01)	0.67	0.70	0.87 (0.01)	0.85	0.88

Note: Three dimensional results are average reliability across all three latent variables.

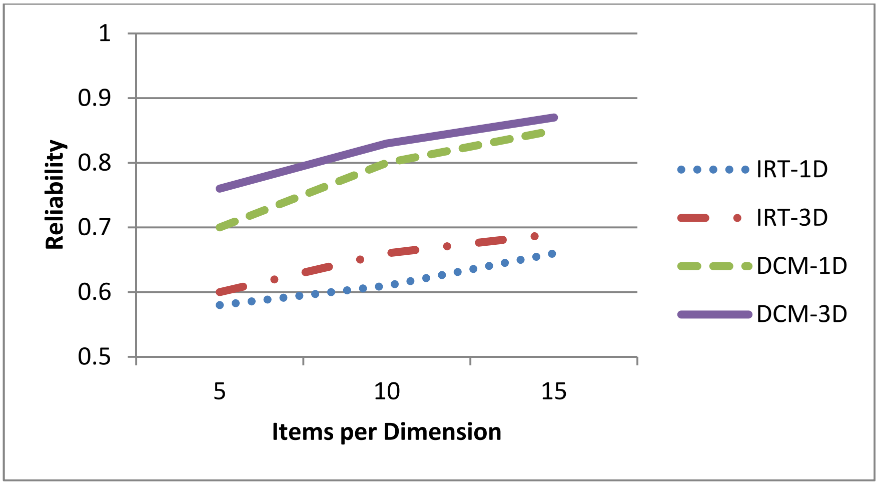

Figure 2. Simulation Study Results: Reliability of DCM and IRT Models with 100 Generated Tests

herein. The test was analyzed using 10 different models from Henson, Templin, and Willse (2009b).

13. Models Assessed

A total of 10 different psychometric models were fit to these data, six being unidimensional and four being multidimensional. The characterization of the reliability for each model is provided along with a description of the model. All models were estimated using the Mplus package, version 5.21 (Muthén and Muthén 2010). Table 3 summarizes the reliability estimates and the information criteria for each model. Because the version of Mplus used at the time did not provide estimates of standard errors for θ, standard errors were assessed via resampling from the posterior distribution of θ so as to provide a comparable metric across estimated models.

Unidimensional Models.

1. *Unidimensional Rasch model.* The first model fit was a unidimensional Rasch model where the variance of θ was estimated. The estimated reliability for θ was 0.841.

2. *Unidimensional two-parameter logistic model.* To provide a more complex modeling approach, the unidimensional two-parameter logistic model was fit (2PL), where the variance of θ was fixed to one for identification. The estimated reliability for θ was 0.842.

3. *Unidimensional two-category DCM.* A unidimensional two-category DCM was fit to the data, as is found in Equation (1). The model is the categorical analog to the 2PL model with two categories for α. The categories could represent proficient and not-proficient students. The estimated reliability for α was 0.992.

4. *Unidimensional three-category DCM.* A unidimensional three-category DCM was fit to the data, as is found in Equation (1) with α having three categories. The distribution of α was multinomial, with each category having an associated parameter for the proportion of examinees in the population having that level of α. The model is the categorical analog to the 2PL model with three categories for α. The categories could represent three proficiency levels of students. The estimated reliability for α was 0.975.

5. *Unidimensional four-category DCM.* A unidimensional four-category DCM was fit to the data, as is found in Equation (1) with α having four categories. Again, the distribution of α was multinomial, with each category having an associated parameter for the proportion of examinees in the population having that level of α. The model is the categorical analog to the 2PL model with four categories for α. The categories could represent four proficiency levels of students, as is used in states such as Georgia. The estimated reliability for α was 0.953.

6. *Unidimensional five-category DCM.* A unidimensional five-category DCM was fit to the data, as is found in Equation (1) with α having five categories. Again, the distribution of α was multinomial, with each category having an associated parameter for the proportion of examinees in the population having that level of α. The model is the categorical analog to

Table 3. Empirical Application: Reliability Estimates

Model	BIC	Number of Parameters	Latent Variable 1	Latent Variable 2	Latent Variable 3
IRT – 1PL	121,988.7	74	0.841	-	-
IRT – 2PL	119,302.5	146	0.842	-	-
IRT – MIRT	119,266.2	147	0.839	0.798	-
IRT – 3D MIRT	119,479.2	220	0.706	0.511	0.433
DCM – 1D 2C	121,797.0	147	0.992	-	-
DCM – 1D 3C	120,173.3	148	0.975	-	-
DCM – 1D 4C	119,645.7	149	0.953	-	-
DCM – 1D 5C	119,479.9	150	0.930	-	-
DCM – 2Dim	121,732.3	149	0.991	0.967	-
DCM – 3Dim	119,504.1	222	0.759	0.692	0.580

the 2PL model with five categories for α. The categories could represent five proficiency levels of students, as is used in states such as the Midwestern state where the test was used. The estimated reliability for α was 0.930.

Multidimensional Models.

7. *Two-dimensional IRT model.* A MIRT model with two dimensions (one for each sub-domain of the test) was estimated. The model estimated the correlation between latent variables but fixed each variable's variance to one, as reported by Ackerman (2009). The reliability for the first latent variable (Reading) was 0.839. The reliability for the second latent variable (Comprehension) was 0.798. The MIRT model had the lowest BIC index value, when compared with all 9 other models estimated.

8. *Two-dimensional two-category DCM.* A two-category DCM that was analogous to the two-dimensional MIRT model was estimated, as reported by Henson, Templin, and Willse (2009b). The model estimated the general association between attributes using an unstructured attribute model (see Rupp, Templin, & Henson, 2010). The reliability for the first latent variable (Reading) was 0.991. The reliability for the second latent variable (Comprehension) was 0.967.

9. *Three-dimensional MIRT model with continuous general factor independent from two correlated continuous subfactors.* A three-dimensional

MIRT model was estimated, as reported by Ackerman (2009). The model estimated a continuous general factor and two continuous subfactors (Reading and Comprehension). The model estimated the association between the sub-factors but left both sub-factors uncorrelated with the general factor. The reliability for the general factor was 0.706. The reliability for the first sub-factor (Reading) was 0.511. The reliability for the second sub-factor (Comprehension) was 0.433.

10. *Three-dimensional DCM with continuous general factor independent from two correlated categorical subfactors.* A three-dimensional DCM that was analogous to the MIRT model was estimated, as in Henson, Templin, and Willse (2009b). The general factor was continuous and the subfactors were categorical with two categories. The model estimated an association between the sub-factors (Reading and Comprehension) but left both sub-factors uncorrelated with the general factor. The reliability for the general factor was 0.759. The reliability for the first sub-factor (Reading) was 0.692. The reliability for the second sub-factor (Comprehension) was 0.580.

14. General Results

As shown in the results, for all models, the DCM analogs produced higher reliability for the latent attributes measured by the test and extracted by each model. Again, this fits with the theoretical and simulation studies indicating that the nature of the attributes allows for more precise measurement in DCMs than in IRT models with tests of the same length. Two key results are of note. First, the reliability for unidimensional DCMs far exceeds the unidimensional IRT model analogs. We will use this result shortly to show how, when *built for* and analyzed with a DCM, current empirical tests may be shortened to assess examinees with the same precision as the unidimensional IRT model. Second, the reliability for the two-dimensional DCM was high enough to extract useful information about both dimensions. This indicates that it is feasible to precisely estimate more than one latent trait with a large scale test, which opens up possibilities for aligning assessment with more complex standards that may greatly aid the processes of instruction, assessment, and curriculum.

We note that reliability is at best an upper bound when a model does not fit the data well. As such, we present these results more to compare reliability across different types of models than to use it as a metric to pick which model to use when scoring a test. We caution the reader from doing so as well and suggest that a test that is constructed and calibrated with a given model will likely fit that model best. The results presented here are to suggest that if the calibration model is a DCM, then a high degree of reliability will likely result (provided the test fits the model when data are collected).

Table 4. Items Needed to Achieve IRT Reliability Level Using DCM

Model	Items	% of Original Test Length
IRT – 2PL	73	100
DCM – 1D 2C	14	19
DCM – 1D 3C	25	34
DCM – 1D 4C	36	49
DCM – 1D 5C	44	60

Note: Reliability for all reported tests in table is 0.842.

15. Shortening End-of-Grade Test to Achieve IRT Reliability with DCMs

One feature emanating from the comparison of the unidimensional modeling results between DCMs and IRT models is the possibility of creating *shorter* tests with DCMs, measuring at the same level of precision as tests constructed with IRT models. Specifically, the unidimensional IRT model represents a standard for analysis and reporting of information in end-of-grade testing. The estimated reliability for θ using the 2PL for the 73 item test was 0.842. The unidimensional DCM provided analogous information at a higher reliability, which was 0.992 for two categories, 0.975 for three categories, 0.953 for four categories, and 0.930 for five categories. This result suggests two important findings: (1) DCMs are approximations to IRT models that have higher reliability because they attempt to locate examinees with fewer latent scale points, and (2) if only a few scale points were used at the outcome of an assessment (as is done ubiquitously with end-of-grade tests assessing the proficiency of examinees), shorter tests could be used to assess examinees at a similar level of reliability.

We now present the results of a final analysis that demonstrates how much shorter the empirical test could be to measure the categorical attribute of a unidimensional DCM with the reliability found under the 2PL model, or 0.842. Using the results of the 2, 3, 4, and 5 category DCMs, we created tests using the items of varying lengths (between 3 and 73 items) using the items with the most information (based on estimated parameters) for each respective model. We then reclassified examinees based on the shorter test and computed the reliability of our estimates. Table 4 lists the number of items needed for the DCM to result in a reliability of 0.842, equal to the 2PL reliability for the test. The two-category DCM needed only 14 items (less than one fifth as many) to reach this reliability. Such a model could classify examinees as proficient in reading or not proficient, meeting policy mandates for end-of-grade tests. The other models needed 25, 36, and 44, respectively. Each of these results suggests that precision can be increased by using a DCM, or test length can be greatly decreased. In a testing cycle, this could save two to three days for students taking the shorter end-of-grade tests.

16. Concluding Remarks

A common saying in introducing DCMs is that the purpose of their use is to be able to extract more information from a test. For instance, take Rupp and Templin (2008): "over the last 20 years there has been a renewed and widened psychometric interest in statistical models with latent variables that provide *multidimensional classifications of respondents* for the purpose of a fine-grained *diagnosis*" (p. 220). This paper sought to quantify the reliability at which such information has been extracted in DCMs. In so doing, we showed that DCMs provided more precise measurement of the latent variables modeled than did analogous IRT models, something implicit in the statements of the rationale for using DCMs cited by many authors in addition to Rupp and Templin (2008).

In defining DCM reliability, however, we now understand such statements to describe a paradox of DCMs. DCMs measure latent traits as categorical latent variables, which, when compared to IRT models' continuous latent variables, present a more coarse level of measurement. It is because of this ordinal level of measurement that DCMs are more precise at locating examinees than their IRT analogs. As a consequence, tests constructed for and analyzed with DCMs can extract more dimensions from the same number of items, precisely measuring multiple attributes with complex loading structures with test lengths similar to those commonly used in large scale testing for unidimensional latent variables today. Therefore, perhaps it is more correct to state that DCMs can extract *multiple course-grained* attributes from a test, providing a practical mechanism for assessing ability from a multidimensional perspective. The consequences of the properties of DCMs are only now being realized, but have the potential to vastly change—and hopefully improve— the large scale testing landscape by better understanding the nature of an examinee's knowledge.

We note a few remaining points about the results presented in this paper. First, we again caution that reliabilities reported on empirical data are upper bounds that are dependent on model fit. This is true for both IRT models and DCMs. We expect that as tests are created using DCMs as calibration models (i.e., screening items that do not fit DCMs, as is common with large testing programs and IRT models), the estimated reliabilities will be more accurate than the ones reported herein. More specifically, reliability is not a measure of model fit. We also note that the comparison between IRT and DCM reliabilities uses different metrics: the Pearson and tetrachoric correlation coefficients. As the types of correlations are different, the comparison is not necessarily equivalent. The design of the simulation study was built to provide a rough comparison and, given the results, we feel our claims of higher reliability for DCMs over IRT models are accurate. IRT models have differential reliability across the scale of the latent attribute, meaning they have more places where error can occur. By categorizing the latent variable, error is consolidated, increasing the reliability of the estimate. Quantifying the exact difference, however, is a difficult task given the nature of the two types of variables.

The empirical use of DCMs is where the methods will ultimately be tested. Multidimensional feedback from tests is not frequently given in educational measurement with the lack of reliability of the multiple traits being a primary reason. Furthermore, the reliability of IRT model examinee estimates is not commonly reported and not well understood with far too many unreliable estimates being used in practice. Since their inception, DCMs have been purported to provide more information than IRT models. Implicit in that claim is that the information they provide is reliable. This paper developed a metric of reliability for DCMs that attempted to bridge the gap in the research literature, providing researchers and practitioners with a means to compute reliability when using DCMs. Although the comparative results in this paper may be somewhat counter intuitive to measurement researchers, they open the door to future research with DCMs thereby further illuminating their role and usefulness as a psychometric method.

References

ACKERMAN, T. (2009), "Using Confirmatory MIRT Modeling to Provide Diagnostic Information in Large Scale Assessment", paper presented at the April 2009 meeting of the National Council for Measurement in Education, San Diego CA.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION, and NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (1999), *Standards for Educational and Psychological Testing*, Washington DC: Authors.

BIRNBAUM, A. (1968), "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability", in *Statistical Theories of Mental Test Scores*, eds. F.M. Lord and M.R. Novick, Reading MA: Addison-Wesley, pp. 397–479.

DE AYALA, R.J. (2009), *Theory and Practice of Item Response Theory*, New York: Guilford.

HABERMAN, S.J., VON DAVIER, M., and LEE, Y.-H. (2008), "Comparison of Multidimensional Item Response Models: Multivariate Normal Ability Distributions Versus Multivariate Polytomous Ability Distributions", Research Report 08-45, Princeton NJ: Educational Testing Service.

HAERTEL, E. (1989), "Using Restricted Latent Class Models to Map the Skill Structure of Achievement Items", *Journal of Educational Measurement, 26*, 333–352.

HAMBLETON, R.K., SWAMINATHAN, H., and ROGERS, H.J. (1991), *Fundamentals of Item Response Theory*, Newbury Park CA: Sage.

HARTZ, S.M. (2002), *A Bayesian Framework for The Unified Model for Assessing Cognitive Abilities: Blending Theory with Practicality*, unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

HENSON, R., TEMPLIN, J., and WILLSE, J. (2009a), "Defining a Family of Cognitive Diagnosis Models Using Log-Linear Models with Latent Variables", *Psychometrika, 74*, 191–210.

HENSON, R., TEMPLIN, J., and WILLSE, J. (2009b), "Ancillary Random Effects: A Way to Obtain Diagnostic Information from Existing Large Scale Tests", paper presented at the April 2009 meeting of the National Council for Measurement in Education, San Diego CA.

JUNKER, B.W., and 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., and GIERL, M.J. (Eds.) (2007), *Cognitive Diagnostic Assessment for Education: Theory and Practices*, Cambridge: Cambridge University Press.

LORD, F.M. (1980), *Applications of Item Response Theory to Practical Testing Problems",* Hillsdale NJ: Erlbaum.

MACREADY, G.B., and DAYTON, C.M. (1977), "The Use of Probabilistic Models in the Assessment of Mastery", *Journal of Educational Statistics, 2*, 99–120.

MARIS, E. (1999), "Estimating Multiple Classification Latent Class Models", *Psychometrika*, *64*, 197–212.

MAYDEU-OLIVARES, A., and JOE, H. (2005), "Limited- and Full-Information Estimation and Goodness-of-Fit Testing in 2^n Contingency Tables: A Unified Framework", *Journal of the American Statistical Association, 100*, 1009–1020.

MISLEVY, R.J., BEATON, A.E., KAPLAN, B., and SHEEHAN, K.M. (1992), "Estimating Population Characteristics from Sparse Matrix Samples of Item Responses", *Journal of Educational Measurement, 29,* 133–161.

MUTHÉN, L.K., and MUTHÉN, B.O. (2010), "Mplus User's Guide" (Version 5.21, Computer software and manual), Los Angeles CA: Muthén and Muthén.

ROUSSOS, L., DIBELLO, L., STOUT, W., HARTZ, S., HENSON, R., and TEMPLIN, J. (2007), "The Fusion Model Skills Diagnosis System", in *Cognitive Diagnostic Assessment in Education,* eds. J. Leighton and M. Gierl, New York NY: Cambridge University Press, pp. 275–318.

RUPP, A., and TEMPLIN, J. (2008), "Unique Characteristics of Diagnostic Models: A Review of the Current State-of-the-Art", *Measurement*, *6*, 219–262.

RUPP, A., TEMPLIN, J., and HENSON, R. (2010), *Diagnostic Measurement: Theory, Methods, and Applications,* New York: Guilford.

SINHARAY, S., and HABERMAN, S. J. (2009), "How Much Can We Reliably Know About What Examinees Know?", *Measurement, 7,* 49–53.

TEMPLIN, J. (2004), *Generalized Linear Mixed Proficiency Models for Cognitive Diagnosis,* unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

TEMPLIN, J. (2006), *CDM User's Guide,* Lawrence KS: University of Kansas.

TEMPLIN, J., and HENSON, R. (2006), "Measurement of Psychological Disorders Using Cognitive Diagnosis Models", *Psychological Methods, 11*, 287–305.

VON DAVIER, M. (2005), "A General Diagnostic Model Applied to Language Testing Data", ETS Research Report RR-05-16.