Efficient Models for Cognitive Diagnosis With Continuous and Mixed-Type Latent Variables

Applied Psychological Measurement
2015, Vol. 39(1) 31–43
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621614524981
apm.sagepub.com

Hyokyoung Hong¹, Chun Wang², Youn Seon Lim³,
and Jeff Douglas³

Abstract

The issue of latent trait granularity in diagnostic models is considered, comparing and contrasting latent trait and latent class models used for diagnosis. Relationships between conjunctive cognitive diagnosis models (CDMs) with binary attributes and noncompensatory multidimensional item response models are explored, leading to a continuous generalization of the Noisy Input, Deterministic “And” Gate (NIDA) model. A model that combines continuous and discrete latent variables is proposed that includes a noncompensatory item response theory (IRT) term and a term following the discrete attribute Deterministic Input, Noisy “And” Gate (DINA) model in cognitive diagnosis. The Tatsuoaka fraction subtraction data are analyzed with the proposed models as well as with the DINA model, and classification results are compared. The applicability of the continuous latent trait model and the combined IRT and CDM is discussed, and arguments are given for development of simple models for complex cognitive structures.

Keywords

cognitive diagnosis, multidimensional item response model, noncompensatory item response model

Introduction

Cognitive diagnosis models (CDMs) are seen as alternatives to item response models, with a purpose of diagnosis through classification of skill mastery rather than locating examinees’ ability values on a continuum. These models tend to involve a vector of binary latent variables, as opposed to the continuous traits of item response theory (IRT) models or the vector of latent traits in multidimensional item response theory (MIRT) models. CDMs and MIRT models can be contrasted by the objective that is desired, whether it is classification or scoring. However,

¹Michigan State University, East Lansing, USA

²University of Minnesota, Minneapolis, USA

³University of Illinois, Champaign, USA

Corresponding Author:

Jeff Douglas, Department of Statistics, University of Illinois, 725 S Wright St., Champaign, IL 61820, USA.

Email: jeffdoug@illinois.edu

when the psychological constructs represented by the latent variables are generally understood as either fine-grained skills or broadly defined abilities, the granularity of the latent trait should be taken into consideration when selecting a modeling approach. When the number of latent traits is very small, CDMs can only be adequate for very simple domains, because of the simplicity of the score distributions implied by the model with so few latent classes. A MIRT model with broadly defined continuous ability variables may make more sense in such situations. However, a finer-grained parsing of the attributes would result in a greater number of latent classes, and CDMs may prove useful in this situation, when the dimensionality is too high for estimation of existing MIRT models.

The authors aim to introduce a MIRT model that is adequately constrained to be practical in high-dimensional problems, and also to study a combination of a CDM and MIRT model for situations in which fine-grained attributes have been identified, along with one or more broadly defined abilities. The proposed model for continuous latent traits may be seen as a continuous extension of the NIDA model (Junker & Sijtsma, 2001), and the model that combines continuous and binary latent variables has an item response function (IRF) of a noncompensatory item response model (NIRT) multiplied by the IRF of a CDM to result in a single IRF with a clear interpretation that involves latent variables of mixed type in a conjunctive manner. In both cases, the objective is to utilize and generalize some of the simpler existing models to arrive at models for more complex and challenging cognitive structures, such as fitting a conjunctive high-dimensional continuous model, or a model with mixed latent variable types.

The authors begin by reviewing CDM and MIRT models. Then, some previously proposed hybrid models, models that include latent variables of mixed type are discussed. This is followed by an introduction of a hybrid model and a continuous model. Several cases are considered in simulation and an analysis of fraction subtraction data is given, with comparisons with results of an eight-attribute CDM and a three-parameter logistic IRT fit. The authors conclude with a discussion of possible applications and future directions for research.

Cognitive Diagnostic Models and Noncompensatory IRT Models

Latent class models for cognitive diagnosis are restricted to reflect some assumptions about the underlying process by which examinees respond to items. The authors review a few of the more popular models, among the class of conjunctive models with binary attributes. A feature in all the models considered is a $\mathbf{Q}$ matrix (K. Tatsuoaka, 1985). This matrix records which attributes or skills are required to correctly respond to each item. Suppose that there are N subjects and J items, with K attributes to classify each subject. Entry q_{jk} in the $\mathbf{J} \times \mathbf{K}$ matrix $\mathbf{Q}$ indicates whether item j requires the k th attribute. Let $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_N$ be random item response vectors of N subjects, where $\mathbf{Y}_n = (Y_{n1}, Y_{n2}, \dots, Y_{nJ})'$. Let $\boldsymbol{\alpha}_n$ denote the attribute pattern for subject n , where $\boldsymbol{\alpha}_n = (\alpha_{n1}, \alpha_{n2}, \dots, \alpha_{nK})'$ and each α_{nk} takes a value of either 0 or 1 for $k = 1, 2, \dots, K$. Specifically, α_{nk} is an indicator of whether the n th subject possesses the k th attribute.

Conjunctive models require that all attributes specified in $\mathbf{Q}$ must be possessed to attain the highest probability of answering the item correctly. An example is the DINA model (Junker & Sijtsma, 2001). Denote the ideal response pattern, the pattern that would arise if item responses were completely determined by skill mastery, $\boldsymbol{\eta}_n = (\eta_{n1}, \eta_{n2}, \dots, \eta_{nJ})'$, where $\eta_{nj} = \prod_{k=1}^K \alpha_{nk}^{q_{jk}}$. It denotes whether subject n has mastered all the attributes required by item j . The DINA model allows for random deviations from this pattern according to slipping parameters for each item, $s_j = P(Y_{nj} = 0 | \eta_{nj} = 1)$, and guessing parameters for each item, $g_j = P(Y_{nj} = 1 | \eta_{nj} = 0)$. The IRF of the DINA model is expressed by

$$P(Y_{nj} = 1 | \alpha_n) = (1 - s_j)^{\eta_{nj}} g_j^{1 - \eta_{nj}},$$

and a conditional likelihood function is constructed from these IRFs by utilizing an assumption of independence of the components of $\mathbf{Y}_n$ given the attribute vector α_n .

The NIDA model, introduced in Maris (1999), differs from the DINA model by treating the slips and guesses at the subtask level. A subtask response η_{nj} indicates whether subject n correctly applied attribute k to answer item j . The parameters of this model are specified for each attribute rather than each item. In particular, $s_k = P[\eta_{nj} = 0 | \alpha_{nk} = 1, q_{jk} = 1]$, $g_k = P[\eta_{nj} = 0 | \alpha_{nk} = 0, q_{jk} = 1]$, and $Y_{nj} = 1$, when all subtasks are correctly completed. By convention, $P[\eta_{nj} = 1 | \alpha_{nk}, q_{jk} = 0] = 1$, no matter the value of α_{nk} . The IRF of the NIDA model is

$$P(Y_{nj} = 1 | \alpha_n, \mathbf{s}, \mathbf{g}) = \prod_{k=1}^K P(\eta_{nj} = 1 | \alpha_{nk}, s_k, g_k) = \prod_{k=1}^K [(1 - s_k)^{\alpha_{nk}} g_k^{1 - \alpha_{nk}}]^{q_{jk}}.$$

An obvious shortcoming of the NIDA model is that items requiring the same set of attributes must have exactly the same IRFs. This implies unrealistically strict conditions on observed proportion correct values of the items. In particular, the model implies that any two items with the same entry in $\mathbf{Q}$ must have the same expected proportion correct. A generalization of the NIDA model that allows slipping and guessing probabilities to vary across items is a reduced version of the noncompensatory Reparameterized Unified Model (Hartz, Roussos, Henson, & Templin, 2005). In this model, the IRF is

$$P(Y_{nj} = 1 | \alpha_n) = \pi_j^* \prod_{k=1}^K r_{jk}^{*q_{jk}(1 - \alpha_{nk})}.$$

Parameter π_j^* is the probability that a subject who possesses all of the required attribute answers item j correctly, and r_{jk}^* can be viewed as a penalty parameter for not possessing the k th attribute, and is between 0 and 1.

More general cognitive diagnostic models have been developed that include common conjunctive, disjunctive, and compensatory models as special cases. For example, see the Generalized-DINA (G-DINA) framework (de la Torre, 2011), the log-linear cognitive diagnostic model (Henson, Templin, & Willse, 2009), and the general diagnostic model (von Davier, 2005).

The most widely used multidimensional IRT model is the compensatory model, in which examinees with high abilities on one dimension can compensate for low abilities on the other dimensions, because the latent variables enter into the IRF as a linear combination. However, this compensatory rule might not hold for some cognitive-based tests. In particular, it may not hold for exams with items constructed so that examinees must master all the required attributes to go through a series of steps leading to a correct answer. The NIRT is more analogous to the conjunctive CDMs discussed above. The IRF results from a product of unidimensional IRFs over each required dimension, similar to how the IRF of a conjunctive CDM might result from a product of subtask indicators, or an intersection of attribute mastery requirements. Forming the IRF in this manner has a conjunctive interpretation, which is the intersection of the conditionally independent events that all required subtasks are performed correctly. A specific example of this is the NIRT model of Whitley (1980), which has IRF,

$$P(Y_{nj} = 1 | \theta_{n1}, \dots, \theta_{nM}) = \prod_{m=1}^M \frac{\exp(\theta_{nm} - b_{jm})}{1 + \exp(\theta_{nm} - b_{jm})}, \quad (1)$$

where, $b_{j1}, \dots, b_{jM}$ are the component-specific difficulty parameters for the j th item.

This NIRT model is embedded in the Embretson (1984, 1997) Multicomponent Latent Trait Model (MLTM). The MLTM differs by including two parameters, a and g , which are very similar to the slip and guess parameters of the DINA model. Compared with CDMs, an advantage of the NIRT model and the MLTM is that they allow finer distinctions among examinees in skill mastery, due to the continuous latent trait. However, a disadvantage is that they are more difficult to fit in high dimensions, requiring some form of numerical integration.

Hybrid models consisting of discrete and continuous latent variables have appeared in cognitive diagnosis and IRT for various purposes. The Unified model of DiBello, Stout, and Roussos (1995) extends the reduced form given above by multiplying a one-parameter logistic IRF to the term involving α to result in the IRF $P(Y_{nj} = 1 | \alpha_n, \theta_n) = \pi_j^* \prod_{k=1}^K r_{jk}^{*q_{jk}(1-\alpha_{nk})} P_j(\theta_n)$. The role of θ is not to represent a particular well-defined psychological construct, but rather to capture the residual dependence of the items if the identified attributes are incomplete. Another notable model that blends continuous and discrete latent variables is the HYBRID model of Yamamoto (1989). Rather than having discrete and continuous traits operating simultaneously and conjunctively to determine the response, the HYBRID model is a mixture of a unidimensional IRT model and a latent class model. Examinees might switch strategies and abandon the IRT model, to begin guessing or some other behavior according to a latent class model. An application is for speededness, when examinees are no longer making the best use of their abilities according to the IRT model. Bradshaw and Templin (2013) have developed a hybrid model that estimates general ability while diagnosing misconceptions. In this work, the discrete attributes do not refer to skills, but rather indicators of whether an examinee holds a misconception.

The Hybrid DINA-NIRT Model

Here, the authors discuss a model that blends a combination of binary and continuous latent traits, used conjunctively, to result in an IRF that is a product of a DINA IRF and a NIRT IRF. An important special case considered throughout is when there is a single continuous latent trait in the model. This pertains to the situation in which diagnosis of a handful of binary traits is of interest, but with a recognition that they only account for a portion of the variation in the responses, and a general ability component exists. That component might be the primary ability of interest, and α might be a vector of few dimensions modeled for the purpose of diagnosing misconceptions, or mastery of a handful of skills. The authors also consider when the NIRT term in the model is multidimensional, and study it in simulations. Both the DINA and NIRT terms are relatively efficient in their parameterization. The DINA term only requires two item parameters, and each component of the NIRT term follows a one-parameter logistic model. Though there may be some resulting bias when restricting models to have few parameters, it may be necessary to reliably fit models in complicated and high-dimensional situations.

As before, let $\mathbf{Q}$ be a $\mathbf{J} \times \mathbf{K}$ matrix with entry q_{jk} denoting whether the k th binary attribute is required for the j th item. Similarly, $\mathbf{R}$ is a $\mathbf{J} \times \mathbf{M}$ matrix with entry r_{jm} indicating whether the m th coordinate of the vector of continuous latent traits θ is required for the j th item. The probability of answering Y_{nj} correctly is modeled as the probability of utilizing all of the binary traits required correctly, according to a DINA model, multiplied by a product of Rasch IRFs for those components of θ required for the item. The portion of the IRF that is a product of Rasch terms is essentially the NIRT model discussed above, but with a matrix $\mathbf{R}$ used for identifiability, and for a clearer interpretation of the construct in a similar manner that the factor structure matrix plays in confirmatory factor analysis. The relatively complex structure involving both latent continuous latent variables and discrete latent variables raises difficulties in computation, motivating the use of somewhat simple terms for the various components. The DINA model is perhaps the simplest of stochastic conjunctive CDM models, with just two parameters per items.

Similarly, the Rasch model only requires one parameter per item for each dimension. Though simple models can suffer from some bias due to underfitting, they can provide a means for addressing difficult analyses when more elaborate models may not be practical or even possible.

Specifically, the IRF of this DINA-NIRT model is $P_j(\alpha_n, \theta_n) = D_j(\alpha_n)H_j(\theta_n)$. For ideal response $\eta_{nj} = \prod_{k=1}^K \alpha_{nk}^{q_{jk}}$, the DINA term in the IRF is $D_j(\alpha_n) = (1 - s_j)^{\eta_{nj}} g_j^{1-\eta_{nj}}$. The NIRT term in the IRF is $H_j(\theta_n) = \prod_{m=1}^M H_{jm}(\theta_{nm})^{r_{jm}}$, where $H_{jm}(\theta_{nm}) = \frac{\exp[a_m(\theta_{nm}-b_m)]}{1 + \exp[a_m(\theta_{nm}-b_m)]}$. Parameter a_m may be set to 1 if the variance of θ_m is treated as unknown. However, the variance of each component of θ to 1 is fixed, and the discrimination parameter must be estimated. The IRF given above is the conditional model, and the conditional likelihood for the entire response vector $\mathbf{Y}_n = (Y_{n1}, Y_{n2}, \dots, Y_{nJ})'$, is obtained using the assumptions that the components of $\mathbf{Y}_n$ are independent, conditional on θ_n and α_n . Furthermore, the usual assumption that $\mathbf{Y}_n$ and $\mathbf{Y}_{n'}$ are independent for $n \neq n'$ is made.

This model can be viewed as a special case of a very general model that combines both discrete and continuous latent variables introduced by Henson, Templin, and Irwin (2009). Their model adds a linear combination of continuous traits to the logit of the log-linear cognitive diagnostic model of Henson, Templin, and Willse (2009). The resulting model affords the chance to fit a general class of models that includes compensatory, conjunctive, and quite general non-compensatory models. Here, the authors focus on an important special case that trims away parameters for efficiency, which is especially useful when either the number of discrete attributes is large or when the continuous latent variable is multidimensional.

A final aspect of a latent trait model is the model for the joint distribution of the latent traits. Many options are available for modeling the joint distribution of θ and α . Because binary and continuous attributes will likely be associated with one another, a standard approach to model the joint distribution of α and θ , is a threshold multivariate normal model. Let $\omega = (\theta, \alpha^*)'$ be a multivariate normal random vector of dimension $K + M$ with mean vector 0. For identifiability in the model, let $\text{var}(\omega_l) = 1$ for $l = 1, 2, \dots, M + K$. Denote the correlation coefficient between ω_l and $\omega_{l'}$ by $\rho_{ll'}$ for $l \neq l'$. The binary components of α arise by estimating threshold parameters that determine the proportion of the population who possess the attributes, $\alpha_k = I[\alpha_k^* > \lambda_k]$. Other options would be threshold models, but with higher order latent traits to structure the correlation matrix of ω , and reduce the number of parameters to be estimated.

In the simulations to follow, the authors use the multivariate normal threshold model for data generation, but a streamlined approach for estimation. Rather than estimate the distribution of α , each α_n is treated as a parameter to be estimated, much like the joint maximum-likelihood approach for θ in IRT, but in an Markov chain Monte Carlo (MCMC) framework that utilizes flat prior distributions on all parameters. This is problematic when estimating continuous traits, because something must be done to fix the scale. However, for binary traits the scale is solidly pinned down between the two possible values in the parameter space, 0 and 1. This results in a more streamlined model without any parameters needed to describe the joint distribution of α , which yields simpler Markov chains and consistent results, as shown in a later section on numerical studies.

The Continuous Conjunctive Model (CCM)

In this section, a simple conjunctive model with continuous latent variables is considered, which will be called the CCM. The aim is to trim the model parameters as much as possible, so that it is practical to use with a very large number of dimensions. This model is motivated by the NIDA for binary latent variables, but has continuous variables, and has no item parameters. Let the latent variable θ_{nm} be the probability that the n th examinee applies the m th skill correctly.

The IRF derives from the conjunctive assumption that an examinee must independently apply all of the required skills correctly, and has the form,

$$P(Y_{nj} = 1 | \theta_n) = \prod_{m=1}^M \theta_{nm}^{r_{jm}}, \quad (2)$$

where r_{jm} is the (j, m) element of a matrix $\mathbf{R}$ that indicates which skills are needed for which items. The likelihood function for each examinee becomes

$$L(\theta_{n1}, \theta_{n2}, \dots, \theta_{nM}) = \prod_{j=1}^J \left[\left(\prod_{m=1}^M \theta_{nm}^{r_{jm}} \right)^{Y_{nj}} \left(1 - \prod_{m=1}^M \theta_{nm}^{r_{jm}} \right)^{1-Y_{nj}} \right]. \quad (3)$$

This model assumes conjunctive relations among the latent abilities, as many of the diagnostic classification models do, but the difference is that there are no item-level parameters, and the stochastic components are absorbed into the examinees' ability profiles θ_{nm} with $m = 1, 2, \dots, M$. Without specifying any further parameters, this model has a direct interpretation and requires no numerical integration for fitting. The simplicity of the model raises questions of fit in low dimensions, but affords the chance to fit models in very high dimensions.

Each examinee is estimated exactly from his or her own response pattern, and the ability profile is not be affected by the estimation error of item parameters. One apparent conceptual advantage of this model is that the examinee's mastery level is continuous instead of dichotomous. This is beneficial in practice because sometimes it does not make sense for certain attributes to be simply classified as present or absent, but instead assessed along a continuum.

This simple continuous diagnostic model is very similar in structure to the NIDA model in which both the slipping and guessing parameters are defined at attribute level. There are no distinct item-level parameters involved, and therefore, for items with the same $\mathbf{Q}$ vector, they will have the same psychometric properties. For example, for a fixed examinee, items with the same $\mathbf{Q}$ vector will end up with exactly the same IRF. The NIDA and the CCM are the same in this respect, but because of the continuous θ of the CCM in contrast with the discrete α of the NIDA, a greater variety of values are achieved by the IRF, which guarantees the CCM will fit at least as well as the NIDA. In this regard, the CCM is more flexible than the NIDA, but the drawback is that the population cannot be conveniently partitioned into 2^K subpopulations.

Numerical Studies

A simulation study was carried out to evaluate estimation of the DINA-NIRT using an MCMC algorithm with flat prior distributions for all parameters and treating α as a structural parameter rather than integrating it over some estimated distribution. In the first condition, data were generated from the DINA-NIRT with a single binary attribute and a single continuous latent trait. This simulates the situation when a dominant continuous ability is meant to be scored simultaneously with diagnosis of a particular skill. Four cases were considered: two test lengths (short = 20, long = 40) and two examinee sample sizes (small = 500, large = 2,000). The results were reported based on 25 simulation runs. The $\mathbf{Q}$ matrix was constructed such that each attribute was measured by about half of the items in the test. The structural parameters were sampled from $s_j \sim \text{Unif}(0, .25)$ and $g_j \sim \text{Unif}(0, .25)$, and the incidental parameters $\omega_n = (\theta_n, \alpha_n^*)$ from the bivariate Normal distribution with mean vector $\mu = (0, 0)$ and covariance matrix Σ with all 1's along the diagonal, and 0.3 being the off-diagonal elements. Binary traits were constructed by $\alpha_n = I[\alpha_n^* > \lambda]$, where λ was set so that $P[\alpha = 1] = 0.7$. The b_j parameters were generated

Table 1. DINA-NIRT With $K = M = 1$: MSE for the s_p , g_p , a , b_p , and θ_n (More Than 25 Replications).

Case	s	g	a	b	θ
$n = 500, j = 20$	0.014	0.003	0.117	0.281	0.445
$n = 500, j = 40$	0.009	0.001	0.040	0.152	0.244
$n = 2,000, j = 20$	0.013	0.003	0.111	0.454	0.471
$n = 2,000, j = 40$	0.010	0.001	0.032	0.075	0.243

Note. NIRT = noncompensatory item response model; MSE = mean square error.

Table 2. DINA-NIRT With $K = M = 1$: Percentage of Correct Attribute Classification (More Than 25 Replications).

Case	α_1
$n = 500, j = 20$	0.918
$n = 500, j = 40$	0.975
$n = 2,000, j = 20$	0.928
$n = 2,000, j = 40$	0.978

Note. NIRT = noncompensatory item response model.

Table 3. DINA-NIRT With $K = 2$ and $M = 1$: MSE for the s_p , g_p , a , b_p , and θ_n (More Than 25 Replications).

Case	s	g	a	b	θ
$n = 500, j = 20$	0.007	0.009	0.254	0.325	0.701
$n = 500, j = 40$	0.005	0.004	0.048	0.317	0.633
$n = 2,000, j = 20$	0.008	0.016	0.347	0.382	0.732
$n = 2,000, j = 40$	0.005	0.005	0.054	0.302	0.653

Note. NIRT = noncompensatory item response model; MSE = mean square error.

from the normal distribution $N(-.5, 1)$. Simulation results are reported in Tables 1 and 2. As shown in Table 1, increasing the test length decreases the mean square error (MSE) of all parameters, and results are as expected. Likewise, in Table 2, it can be seen that correct classification rates are extremely high and get close to 1 for the longer test. Information for attributes is determined by the number of items that require the attribute. When fitting the correct model, one should expect the probability of a correct classification to approach 1 as more and more items are used, provided the model parameters are calibrated accurately, which is dependent upon sample size.

The next condition considered was to add a second attribute. All parameter distributions remained the same. $\mathbf{Q}$ was designed so that a quarter of the items required the first attribute, another quarter required the second attribute, and half required both. Table 3 shows that the MSE for estimation of the IRT terms increase, as can be expected when more latent variables are responsible for the response, and it becomes relatively less influential. DINA parameters s and g are estimated quite accurately, but the slope parameter of the IRT term suffers in the cases of a shorter test length. Table 4 indicates that classification remains strong when an additional binary attribute is included. Each attribute is classified with accuracy nearly over 90%, and the entire vector is correctly classified around 92% in the long test cases, and more than 81% in all cases.

Table 4. DINA-NIRT With $K = 2$ and $M = 1$: Percentage of Correct Attribute Classification (More Than 25 Replications).

Case	α_1	α_2	Both
$n = 500, j = 20$	0.908	0.909	0.840
$n = 500, j = 40$	0.954	0.964	0.928
$n = 2,000, j = 20$	0.892	0.897	0.814
$n = 2,000, j = 40$	0.960	0.955	0.924

Note. NIRT = noncompensatory item response model.

Table 5. MSE for the θ Estimates.

	$M = 3$	$M = 5$	$M = 7$
25	0.018	0.035	0.049
50	0.008	0.018	0.028
100	0.004	0.009	0.015

Note. MSE = mean square error.

Finally, a numerical study was done to investigate whether CCM ability parameters could be estimated with sufficient accuracy using maximum-likelihood estimation with the Newton–Rahpson algorithm. Two factors were considered, length of the exam (small = 25, medium = 50, large = 100) and the number of latent variables M (small = 3, medium = 5, large = 7). Ability values θ_{nm} were generated according to $\Phi(\theta_{nm}^*)$, where $\Phi()$ is the cumulative distribution function of a standard normal random variable, and the vector θ_n^* was generated from a multidimensional normal distribution in which each component mean was 1.25, each component standard deviation was 1, and the correlation between all pairs of components was 0.3. Generating abilities in this way resulted in a population mean value for θ_{nm} of about 0.8 with a standard deviation of 0.2. The resulting IRFs yielded observed score distributions with realistic values and slight positive correlation of item pairs. The $\mathbf{R}$ matrix was constructed randomly from a uniform distribution on the $2^M - 1$ possible vectors, which excluded the case in which none of the abilities were needed for an item. For each condition, a set of 10,000 examinees were independently generated from the CCM. Note that the only factor in the accuracy of ability estimation is test length, and not sample size, because all examinees’ ability estimates are statistically independent of one another.

Results are given in Table 5, where MSEs are averaged over the M dimensions of θ . The expected result that MSE values increase with dimension of θ and decrease with test length can be seen. Not surprisingly, a very short exam is insufficient for a large number of dimensions, and it can be seen that the MSE for θ is about the same as the population variance of θ . However, the results improve steadily as the exam becomes longer. Estimation is quite accurate in all cases when the dimension of the latent ability vector is 3.

Analysis of Fraction Subtraction Data

As an illustration of the models with real data, several models were fitted to the fraction subtraction data of K. Tatsuoka (1990), which consist of responses to 20 items from 536 subjects. The eight-attribute $\mathbf{Q}$ matrix analyzed by C. Tatsuoka (2002), and de la Torre and Douglas (2004)

Table 6. Agreement of $\hat{\alpha}$ Estimation Between DINA-NIRT and DINA Model.

	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	$\hat{\alpha}_7$	$\hat{\alpha}_8$
Simple hybrid models	0.746	0.774	0.524	0.855	0.815	0.526	NA	0.774
Full hybrid model	0.856	0.866	0.670	0.856	0.840	0.701	0.925	0.800

Note. NIRT = noncompensatory item response model; NA = not applicable.

was used. According to these papers, the eight attributes required to answer these items are as follows: (a) Convert a whole number to a fraction, (b) separate a whole number from fraction, (c) simplify before subtracting, (d) find a common denominator, (e) borrow from whole number part, (f) column borrow to subtract the second numerator from the first, (g) subtract numerators, and (h) reduce answers to simplest form. Several researchers have studied the DINA model, and more generally the assumption of a conjunctive response process with these data and this $\mathbf{Q}$ matrix. De Carlo (2011) provides an in-depth analysis and gives an argument for systematic misclassification when using the DINA model. However, de la Torre and Douglas demonstrate excellent fit with the DINA model. Chiu (2013) studied refinement of the $\mathbf{Q}$ matrix, not using the DINA model, but using nonparametric methods of classification that assume a conjunctive underlying but unknown model. When applying a measure of fit after switching the 160 elements of $\mathbf{Q}$ one by one, it was found that the $\mathbf{Q}$ derived through expert opinion could only be improved by switching two entries. A feature of this data set that helps lead to a good fit by conjunctive models with binary attributes is that the skills are essentially steps in solving a problem, making it more plausible that they could be binary. The authors study the DINA-NIRT with these data and also see whether fit can be improved by allowing the attributes to take values in an interval, by applying the CCM.

In the present study, 12 models were fitted and compared. The first eight models were DINA-NIRT models in which a single attribute was fitted, together with a unidimensional continuous trait. Specifically, among these eight models, model k included only the k th column of $\mathbf{Q}$ to try and classify α_k together with a unidimensional θ to try and absorb the effects of the omitted attributes, or simply to estimate a general ability of fraction subtraction. The next model was a full DINA-NIRT model in which the entire matrix $\mathbf{Q}$ was used along with a unidimensional continuous trait. Then a DINA model with no continuous trait was fitted to the entire matrix $\mathbf{Q}$, and a three-parameter logistic IRT model was fitted. Finally, the CCM was fitted treating the eight attributes as continuous.

Table 6 shows the percentage of agreement of α classifications between the DINA-NIRT models and the standard DINA model. Not surprisingly, they are much higher for the full DINA-NIRT because both it and the DINA used the exact same matrix $\mathbf{Q}$. Nevertheless, there is substantial disagreement in these classifications, and an analysis of model fit was done to compare the models. In this analysis, models were compared on their ability to replicate the observed score distribution and to capture pairwise item associations measured by odds ratios.

The observed score distribution is the probability distribution for the total score on the exam. Consider a test with J dichotomous items, let x_j denotes the number of examinees answering exactly j items correctly, and let $\mathbf{X} = (x_0, x_1, \dots, x_J)'$. A reasonable model should predict this $\mathbf{X}$ vector well. A summary statistic for checking the discrepancy between the observed and model predicted score distribution is $\delta = \sum_{i=0}^J \frac{1}{N} |x_i - \hat{x}_i|$, where $\hat{x}_i$ is obtained empirically by averaging over the results obtained from 100 replications, where in each replication scores were generated conditional upon the estimated model parameters. This statistic for fit is similar to that used by Sinharay and Almond (2007) to assess fit of CDMs, and they compare observed values

Table 7. Discrepancy Between Observations and Model Predictions.

Discrepancy	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 8	Full model	3PL	DINA	CCM
δ	0.257	0.300	0.313	0.231	0.290	0.284	0.253	0.269	0.280	0.254	0.193
$PPP \geq .95$	13	12	30	8	0	16	13	2	39	1	76
$PPP \leq .05$	42	66	31	52	58	41	58	107	25	85	44

Note. 3PL = three-parameter logistic model; CCM = Continuous Conjunctive Model; PPP = posterior predictive p .

with the predictive distribution of the statistic. In the first row of Table 7, it can be seen that the DINA-NIRT model that minimizes δ is Model 4, the model that measures only the ability to find a common denominator in addition to a general ability trait. This may indicate a general ability for fraction subtraction, and a skill that separates those who have mastered finding a common denominator and those who have not. Overall, the model that minimizes δ is the CCM. This was fitted with all eight attributes, but may offer more flexibility in modeling than the competing DINA, because attributes are allowed to take any value in the interval [0, 1].

Odds ratios may be used to measure the association between item pairs. Consider an item pair with items i and j , the odds ratio is defined as $OR_{ij} = \frac{n_{11}n_{00}}{n_{10}n_{01}}$, where $n_{kk'}$ denotes the number of individuals getting a score of k on item i and score k' on item j . Investigating observed versus expected odds ratios can help detect whether the fitted model can adequately explain the associations among the test items and has been used to investigate goodness of fit in CDMs by de la Torre and Douglas (2004), and goodness of fit in multidimensional IRT models by Levy, Mislevy, and Sinharay (2009). Here, the odds ratios are studied through posterior predictive model checking. The posterior predictive p value (PPP value) was computed, and Table 7 shows how often they were unusually high or unusually low among the 190 item pairs. The PPP values indicate how many odds ratios calculated from posterior predictive samples are larger than the observed odds ratio. A PPP value is essentially a version of a p value in a fully Bayesian model, and is found by comparing the observed value of the statistic with the appropriate tail of its posterior predictive distribution. This idea has been applied to item response model in Sinharay, Johnson, and Stern (2006). All models considered had a tendency to underestimate item associations measured by odds ratios, especially the full DINA model and the full DINA-NIRT model. It is difficult to know precisely why, but one could speculate that conditioning on binary latent traits is too coarse, and possibly finer gradations found in continuous latent traits would be required to achieve conditional independence, which is needed for modeling item associations correctly. Despite showing the best overall fit to the score distribution, the CCM had several PPP values in the extreme. However, because the CCM is not a fully Bayesian model, predictive values were found by drawing data from the estimated parameters, rather than from models with parameters taken from the posterior distribution. This may have accounted for predictive distributions with tighter tails, leading to the observed statistics appearing more extreme.

As a final diagnostic, estimates of the CCM parameters were investigated, conditional on their corresponding classifications using the full DINA model. The $\mathbf{R}$ matrix of the CCM was precisely the eight-column $\mathbf{Q}$ matrix of the DINA. By looking dimension-by-dimension at the distribution of θ estimates corresponding to whether $\hat{\alpha} = 1$ or $\hat{\alpha} = 0$, an informal look can be had at the issue of latent variable granularity. Summary statistics for θ estimates corresponding to α classifications are given in Table 8. Here, it can be seen that for many of the dimensions, $\hat{\theta}$ corresponding to $\hat{\alpha} = 1$ were tightly accumulated around 1, having 1 as a median in every

Table 8. Summary Statistics for $\hat{\theta}$ Conditional on $\hat{\alpha}$.

$\hat{\alpha}$	Mean $\hat{\theta}$		SD $\hat{\theta}$		Median $\hat{\theta}$	
	1	0	1	0	1	0
$\hat{\alpha}_1$	0.768	0.054	0.335	0.155	1.000	0.000
$\hat{\alpha}_2$	0.930	0.251	0.122	0.215	1.000	0.222
$\hat{\alpha}_3$	0.782	0.464	0.304	0.317	1.000	0.458
$\hat{\alpha}_4$	0.866	0.181	0.251	0.274	1.000	0.000
$\hat{\alpha}_5$	0.859	0.244	0.270	0.321	1.000	0.073
$\hat{\alpha}_6$	0.800	0.250	0.349	0.311	1.000	0.000
$\hat{\alpha}_7$	0.944	0.278	0.118	0.222	1.000	0.250
$\hat{\alpha}_8$	0.905	0.396	0.231	0.378	1.000	0.444

case. For Attributes 3 and 8, $\hat{\theta}$ corresponding to $\hat{\alpha} = 0$ had median values far from 0, but were near 0 for other attributes. Overall, these ability estimates polarized quite nicely depending on attribute classifications, providing support for treating attributes as binary valued when modeled as operating conjunctively.

Discussion

CDMs can be thought of as alternatives to unidimensional and multidimensional item response models when diagnostic information is desired. However, merely aiming for diagnostic information or assuming a conjunctive response process does not change the interpretation of the latent trait. If the interpretation is broad, it may be unrealistic to model the latent variable as binary, or even discrete. However, some traits may amount to simple steps in solving a problem and may be validly viewed as binary, possibly in the presence of broader abilities also required for the problem. Whether classification or scoring is desired, it is worth considering the interpretation of the latent variables involved and modeling them appropriately. In some instances, it may be reasonable to assume that the latent variables are a mixture of discrete and continuous variables, and may be modeled accordingly. The study’s aim was to introduce such a model, by combining a continuous model and a conjunctive CDM that are relatively simple, and have a clear interpretation when combined. In addition, recognizing that modeling CCMs can become extremely difficult when the number of latent traits becomes large, the CCM was introduced, a model with continuous traits that can be interpreted as a continuous generalization of the NIDA model. It only has person parameters with direct interpretations as the probability of applying a skill directly and can be fitted in very large dimensions.

The DINA-NIRT is a simple model with few parameters that may prove useful when latent variables of mixed type seem plausible. Examples of fitting this model with both 1 and 2 continuous traits have been given, but it can readily be extended, especially if the latent traits follow a simple structure as proposed by applications to testlets when doing cognitive diagnosis. One feature of the MCMC algorithm for estimation that made it reliable and effective was stripping out parameters related to the population distribution of α and treating them as fixed parameters. This is much easier with binary latent traits than with joint likelihood approaches involving continuous latent traits, because the simple support of the binary traits fixes the scale and it does not suffer from identifiability issues. When estimated through MCMC, another viewpoint of this is that the authors begin with a misspecified prior distribution for α that treats it as uniformly distributed on the possible values and independent of θ . With enough items, the likelihood

dominates the prior, and it was observed that the true distribution of α can be accurately estimated by the empirical distribution of $\hat{\alpha}$.

Multidimensional item response models, whether they are compensatory or noncompensatory, become very difficult to fit once the dimension of the latent trait exceeds 3 or 4, due to complicated numerical integration in such high dimensions. The CCM provides an alternative that makes some rather strict assumptions to keep the parameterization as uncomplicated as possible, but can be used with several latent variables and has a clear interpretation. It can be seen as more general than the NIDA model, and will fit the data well in at least as many situations. In addition, even when it does display misfit, and may be inappropriate for scoring in high stakes situations, it can offer useful diagnostic information on a continuum.

To summarize, CDMs and NIRT models have very similar interpretations, and primarily differ in the granularity of their latent variables. Hybrid models may afford a chance to fit more realistic models, rather than misspecifying the types of the latent variables, merely for the purpose at hand. The DINA-NIRT is one simple example of a model that could be used for mixed types, though any combination of a conjunctive CDM and a noncompensatory IRT model may work through taking a product of their IRFs. In the difficult case where many continuous latent traits are required along with a noncompensatory assumption, the CCM is a practical model that may prove useful, even in situations when fitting other NIRT models may not be possible.

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) received no financial support for the research, authorship, and/or publication of this article.

References

- Bradshaw, L., & Templin, J. (2013). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*. Advance online publication. doi:10.1007/s11336-013-9350-4
- Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, 37, 598-618.
- De Carlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, 35, 8-26.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179-199.
- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.
- DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-389). Mahwah, NJ: Lawrence Erlbaum.
- Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika*, 49, 175-186.
- Embretson, S. E. (1997). *Multicomponent latent trait models*. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305-322). New York, NY: Springer-Verlag.
- Hartz, S., Roussos, L., Henson, R., & Templin, J. (2005). *The fusion model for skills diagnosis: Blending theory with practicality*. Unpublished manuscript.
- Henson, R., Templin, J., & Irwin, P. (2009, April). *Ancillary random effects: A way to obtain diagnostic information from existing large scale tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

- Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74, 191-210.
- Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement*, 25, 258-272.
- Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, 33, 519-537.
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 54, 184-212.
- Sinharay, S., & Almond, G. (2007). Assessing fit of cognitive diagnostic models, a case study. *Educational and Psychological Measurement*, 67, 239-257.
- Sinharay, S., Johnson, M., & Stern, H. S. (2006). Posterior predictive assessment of item response models. *Applied Psychological Measurement*, 30, 298-321.
- Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51, 337-350.
- Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*, 12, 55-73.
- Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Legold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Lawrence Erlbaum.
- von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report RR-05-16). Princeton, NJ: Educational Testing Service.
- Whitley, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45, 479-494.
- Yamamoto, K. (1989). *HYBRID model of IRT and latent class models* (ETS Research Report RR-89-41). Princeton, NJ: Educational Testing Service.