Model Similarity, Model Selection, and Attribute Classification

Applied Psychological Measurement
2016, Vol. 40(3) 200–217
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621615621717
apm.sagepub.com

Wenchao Ma¹, Charles Iaconangelo¹, and Jimmy de la Torre¹

Abstract

Selecting the most appropriate cognitive diagnosis model (CDM) for an item is a challenging process. Although general CDMs provide better model-data fit, specific CDMs have more straightforward interpretations, are more stable, and can provide more accurate classifications when used correctly. Recently, the Wald test has been proposed to determine at the item level whether a general CDM can be replaced by specific CDMs without a significant loss in model-data fit. The current study examines the practical consequence of the test by evaluating whether the attribute-vector classification based on CDMs selected by the Wald test is better than that based on general CDMs. Although the Wald test can detect the true underlying model for certain CDMs, it is yet unclear how effective it is at distinguishing among the wider range of CDMs found in the literature. This study investigates the relative similarity of the various CDMs through the use of the newly developed dissimilarity index, and explores the implications for the Wald test. Simulations show that the Wald test cannot distinguish among additive models due to their inherent similarity, but this does not impede the ability of the test to provide higher correct classification rates than general CDMs, particularly when the sample size is small and items are of low quality. An empirical example is included to demonstrate the viability of the procedure.

Keywords

cognitive diagnosis model, CDM, Wald test, dissimilarity index, model selection, classification rate, G-DINA

Introduction

Compared with traditional psychometric frameworks, such as classical test theory or item response theory, where a true score or latent trait is assumed so that students can be located on a continuum based on their performance in assessments, profiles scores can be culled from cognitive diagnosis models (CDMs) to provide finer-grained information about students' strengths and weaknesses. Such information can be used to inform classroom instruction or tailor remediations.

A wide array of CDMs (see, for example, Rupp, Templin, & Henson, 2010) have been developed based on different assumptions or theories about how cognitive processes, skills, or

¹Rutgers, The State University of New Jersey, New Brunswick, USA

Corresponding Author:

Wenchao Ma, Rutgers, The State University of New Jersey, 10 Seminary Place, New Brunswick, NJ 08901, USA.
Email: wenchao.ma@rutgers.edu

attributes influence students' responses in assessments. The deterministic inputs, noisy, "and" gate (DINA; Haertel, 1989) model, an example of a conjunctive model, assigns the highest probability of answering correctly to examinees that possess all of the required attributes, as specified by the Q-matrix. Disjunctive models, however, assume that lacking a particular attribute can be off-set by possessing another. For example, the deterministic inputs, noisy, "or" gate (DINO; Templin & Henson, 2006) model assigns the highest probability of answering correctly to examinees with at least one of the required attributes (de la Torre & Douglas, 2004). Examples of other specific, interpretable CDMs are the reduced reparametrized unified model (R-RUM; Hartz, 2002), the *additive* CDM (*A*-CDM; de la Torre, 2011), and the linear logistic model (LLM; Maris, 1999).

Apart from these specific CDMs, general or saturated CDMs subsuming many widely used specific CDMs have also been developed, including the generalized DINA (G-DINA; de la Torre, 2011) model, the general diagnostic model (GDM; von Davier, 2008), and the log-linear CDM (LCDM; Henson, Templin, & Willse, 2009).

Although a multitude of CDMs are available, it is not clear how the most appropriate model for a specific test can be identified because the cognitive processes in answering items may be complicated. For example, two items may require the same attributes, yet the attributes may contribute differently to the probability of answering correctly. In addition, the saturated models can provide better model-data fit than specific CDMs, but the latter may be more appropriate for several reasons. First, reduced CDMs usually have more straightforward interpretations and require smaller sample sizes for accurate parameter estimation. Second, consistent with the parsimony principle (e.g., Beck, 1943), the simpler model is to be preferred if it is not significantly worse than the more complex model. Finally, appropriate reduced models can provide better classification rates than saturated models, especially when the sample size is small (Rojas, de la Torre, & Olea, 2012).

Considering the importance of model selection, many approaches have been investigated. For example, at the test level, model-data fit can be evaluated and compared using Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978). For examples of comparisons, see Chen, de la Torre, and Zhang (2013) and Henson et al. (2009). At the item level, Henson et al. provided a way to determine the best reduced model by visual inspection of the estimates of LCDM. In de la Torre and Lee (2013), the Wald test was used to compare the G-DINA model with DINA, DINO, and *A*-CDM, and demonstrated high power while controlling Type I error. Compared with model selection at the test level, item-level model selection does not require blanket acceptance of a model for all the items, which may avoid suboptimal choices, and enable the researchers to investigate the best CDM for each of the items. The current study aims to examine the practical consequence of the test by evaluating whether the attribute classification based on CDMs selected by the Wald test is better than classification based on a general CDM. More specifically, the impact of model similarity on model selection and attribute classification is investigated, which, in turn, necessitates the introduction of a statistic to measure the dissimilarity of a pair of models. The performance of the Wald test when the reduced model is LLM or R-RUM is then examined.

The remaining sections of the article are laid out as follows. In the "Background" section, the G-DINA model and its relation with other reduced models are briefly reviewed, along with the Wald test. The "Dissimilarity Among Reduced CDMs" section discusses the similarities among the reduced models, and the "Evaluating the Performance of the Wald Test for R-RUM and LLM" section explores the implications by examining the Type I error and power of the Wald test for LLM and R-RUM. The "Attribute Classification Accuracy" section compares the attribute classification of the saturated model (G-DINA) and the reduced models selected by the Wald test. An empirical example is included to demonstrate the procedure with real data.

The article concludes with a discussion of the limitations of the Wald test and direction for further research.

Background

G-DINA and Reduced Models

Used in conjunction with a binary item-attribute association matrix called a Q-matrix (Tatsuoka, 1983), the G-DINA model creates $2^{K_j^*}$ latent clusters, where K_j^* is the number of required attributes for item j , as in $K_j^* = \sum_{k=1}^{K_j^*} q_{jk}$. Here, q_{jk} represents the k th element of the j th row of the Q-matrix (Tatsuoka, 1983). To simplify the notation, the first K_j^* attributes are assumed to be the required attributes for item j , and α_{lj}^* is the reduced attribute vector consisting of the columns of the required attributes, where $l = 1, \dots, 2^{K_j^*}$. The probability of success on item j by students with attribute pattern α_{lj}^* will be denoted by $P(X_j = 1 | \alpha_{lj}^*) = P(\alpha_{lj}^*)$. Readers interested in more details about the G-DINA model are referred to de la Torre (2011). The item response function (IRF) of the G-DINA model is given by

$$g\left[P\left(\alpha_{lj}^*\right)\right] = \phi_{j0} + \sum_{k=1}^{K_j^*} \phi_{jk} \alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \phi_{jkk'} \alpha_{lk} \alpha_{lk'} + \dots + \phi_{j12\dots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}, \quad (1)$$

where $g[P(\alpha_{lj}^*)]$ is $P(\alpha_{lj}^*)$, $\log[P(\alpha_{lj}^*)]$, and $\text{logit}[P(\alpha_{lj}^*)]$ in the identity, log and logit links, respectively. In addition, ϕ_{j0} is the intercept for item j , ϕ_{jk} is the main effect due to α_k , $\phi_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$, $\phi_{j12\dots K_j^*}$ is the interaction effect due to $\alpha_1, \dots, \alpha_{K_j^*}$. The G-DINA model is a saturated model and subsumes several widely used reduced CDMs, including the DINA model, the DINO model, the A-CDM, the LLM, and the R-RUM. In this article, ϕ is used as item parameters across all these models. To obtain the DINA model, all terms in the G-DINA model in identity link, except ϕ_0 and $\phi_{12\dots K_j^*}$, are constrained to zero, that is,

$$P\left(\alpha_{lj}^*\right) = \phi_{j0} + \phi_{j12\dots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}. \quad (2)$$

For the DINO model, the IRF is given by

$$P\left(\alpha_{lj}^*\right) = \phi_{j0} + \phi_{jk} \alpha_{lk}, \quad (3)$$

where $\phi_{jk} = -\phi_{jk'k''} = \dots = (-1)^{K_j^*+1} \phi_{j12\dots K_j^*}$, for $k = 1, \dots, K_j^*$, $k' = 1, \dots, K_j^* - 1$, and $k'' > k', \dots, K_j^*$. This results in two parameters for the DINO model as well as the DINA model. By assuming all interactions are zero in the G-DINA model, the A-CDM, LLM, and R-RUM can be obtained. Specifically, the A-CDM is the constrained identity-link G-DINA model without the interaction terms. It can be formulated as

$$P\left(\alpha_{lj}^*\right) = \phi_{j0} + \sum_{k=1}^{K_j^*} \phi_{jk} \alpha_{lk}. \quad (4)$$

The LLM is the logit link G-DINA model without the interaction terms, and its IRF is given by

$$\text{logit}\left[P\left(\alpha_{lj}^*\right)\right] = \phi_{j0} + \sum_{k=1}^{K_j^*} \phi_{jk} \alpha_{lk}. \quad (5)$$

And finally, the R-RUM, the log-link G-DINA model without interaction terms, can be formulated as

$$\log \left[P \left(\alpha_{ij}^* \right) \right] = \phi_{j0} + \sum_{k=1}^{K_j^*} \phi_{jk} \alpha_{ik}. \quad (6)$$

These three models assume that mastering attribute α_{ik} raises the probability of success on item j , but the increases may be different due to the different scales they employ. As additive models, mastery of each attribute does not affect other attributes. In addition, they all have the same number of parameters for item j (i.e., $K_j^* + 1$). Furthermore, it should be noted that the LLM is equivalent to the C-RUM (Hartz, 2002), a special case of the GDM (von Davier, 2008), and that the R-RUM is also known as a special case of the generalized NIDA model (de la Torre, 2011).

The Wald Test

The Wald test was originally proposed by de la Torre (2011) as an item-level statistical test to examine whether the G-DINA model can be replaced by a reduced CDM without losing model-data fit significantly for items with more than one attribute. It was evaluated across DINA, DINO, and A -CDM by de la Torre and Lee (2013). To evaluate the appropriateness of reduced models using the Wald test, a restriction matrix $\mathbf{R}$ that can be used to test the null hypothesis that $\mathbf{R} \times \boldsymbol{\Phi}_j = \mathbf{0}$ is needed. The matrix is of dimension $(2^{K_j^*} - p) \times 2^{K_j^*}$, where p is the number of parameters in the reduced model. The rows in the restriction matrix specify the contrasts, and the columns represent the latent classes. For instance, for additive models such as R-RUM and LLM, all interactions are set to be equal to zero. If $K_j^* = 3$, the restriction matrix is

$$\begin{bmatrix} 1 & -1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 0 & 1 & 0 \\ -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \end{bmatrix},$$

which implies that $\phi_{j12} = \phi_{j13} = \phi_{j23} = \phi_{j123} = 0$. Similarly, if $K_j^* = 2$, the restriction matrix $\mathbf{R}$ is $\begin{bmatrix} 1 & -1 & -1 & 1 \end{bmatrix}$. The Wald statistic W is computed as

$$W = [\mathbf{R} \times g[P(\boldsymbol{\alpha}_{ij}^*)]]' [\mathbf{R} \times \text{Var}(g[P(\boldsymbol{\alpha}_{ij}^*)])] \times \mathbf{R}'^{-1} [\mathbf{R} \times g[P(\boldsymbol{\alpha}_{ij}^*)]],$$

where $g[P(\boldsymbol{\alpha}_{ij}^*)]$ is the IRF of the G-DINA model from Equation 1. $\text{Var}(g[P(\boldsymbol{\alpha}_{ij}^*)])$ is the variance-covariance matrix, which can be calculated as in de la Torre (2011). The Wald statistic W is assumed to be asymptotically $\chi^2_{2^{K_j^*} - p}$.

Dissimilarity Among Reduced CDMs

The power of a test to distinguish one model from another depends, in part, on the similarities between models. Thus, in certain circumstances, where one model is highly similar to another model, data-based approaches will have difficulty distinguishing the two models. At this point, it is not clear how similar the different CDMs are, and thus, an examination of how well each model can approximate the behavior of the others is warranted. For the purposes of this article, dissimilarity was defined as a function of the differences in probabilities of success across all

possible latent groups between two models. In particular, the dissimilarities between two CDMs (e.g., $M1$ and $M2$) on a particular item j can be written as

$$DS(M1, M2) = \min_{\Phi} \sum_l^{2^{K_j^*}} \left| P_{M1}(\alpha_{lj}^*) - \tilde{P}_{M2}(\alpha_{lj}^*) \right|, \quad (7)$$

where $P_{M1}(\alpha_{lj}^*)$ and $\tilde{P}_{M2}(\alpha_{lj}^*)$ are the success probabilities of α_{lj}^* based on Models $M1$ and $M2$, respectively, and Φ is the parameter vector of $M2$. It should be noted that $P_{M1}(\alpha_{lj}^*)$ is fixed in each condition, whereas $\tilde{P}_{M2}(\alpha_{lj}^*)$ is not. Specifically, the $\Phi = \{\phi_{jk}\}$ is estimated to minimize the sum of the absolute differences in each element of two vectors, $P_{M1}(\alpha_{lj}^*)$ and $\tilde{P}_{M2}(\alpha_{lj}^*)$. In this way, the probabilities of each latent class in $M2$ are adjusted in an attempt to approximate the probabilities in $M1$. When $DS(M1, M2) = 0$, $M2$ can approximate $M1$ perfectly, implying that the same model-data fit would be obtained regardless of which model was used to fit an item.

For example, assume that $M1$ is DINA model. Let $P_{M1}(00)$, $P_{M1}(10)$, and $P_{M1}(01)$ be equal to 0.25, and $P_{M1}(11) = 0.81$; also, assume that $M2$ is A -CDM, then

$$\tilde{P}_{M2}(00) = \phi_0,$$

$$\tilde{P}_{M2}(10) = \phi_0 + \phi_1,$$

$$\tilde{P}_{M2}(01) = \phi_0 + \phi_2, \text{ and}$$

$$\tilde{P}_{M2}(11) = \phi_0 + \phi_1 + \phi_2,$$

where ϕ_0 , ϕ_1 , and ϕ_2 are parameters to be estimated such that $DS(M1, M2)$ is minimized. The optimization function gives the following solution: $\phi_0 = 0.2$, $\phi_1 = 0.3$, and $\phi_2 = 0.3$, which leads to $\tilde{P}_{M2}(00) = 0.2$, $\tilde{P}_{M2}(10) = 0.5$, $\tilde{P}_{M2}(01) = 0.5$, and $\tilde{P}_{M2}(11) = 0.8$. The resulting $DS(M1, M2) = |0.25 - 0.2| + |0.25 - 0.5| + |0.25 - 0.5| + |0.81 - 0.8| = 0.56$.

It should be noted that the dissimilarity index is not symmetric. That is, $DS(M1, M2)$ is not necessarily equal to $DS(M2, M1)$. In addition, when one item is known to follow a specific CDM (i.e., $M1$), $DS(M1, M2)$ can be used to evaluate the extent to which another CDM (i.e., $M2$) can approximate $M1$. A small value of $DS(M1, M2)$ indicates that $M2$ can approximate $M1$ well, whereas a large value suggests that it cannot.

An Example of Equivalency

As mentioned earlier, when $DS(M1, M2) = 0$, $M2$ can approximate $M1$ perfectly. In the following, a simple example is used to illustrate that LLM can approximate A -CDM perfectly under some specific conditions. Assume that item j follows A -CDM, requires two attributes, and each attribute contributes equally to the probability of success—that is, $\phi_1 = \phi_2$. If the guessing parameter is equal to the slip parameter, then the LLM can approximate the item parameters equally well. It can be shown as follows: denote the guessing probability as g , which means $P_j(00) = g$ and $P_j(11) = 1 - g$. Because $\phi_1 = \phi_2$, $P_j(10) = P_j(01) = 0.5$. When adopting LLM, let $P_j(00) = g$ and $P_j(11) = 1 - g$ as well. Then, $P_j(10)$ and $P_j(01)$ can be calculated as

$$\left[1 + \left[\exp \frac{\text{logit}(1 - g) + \text{logit}(g)}{2} \right]^{-1} \right]^{-1},$$

which is, after calculation, equal to 0.5, regardless of g . Thus, in this specific case, LLM and A -CDM are equivalent and therefore indistinguishable under a specific situation. However, in more realistic situations, different models may not be exactly equivalent, yet may be very similar.

Dissimilarities Based on a Simulation Study

To explore more general conditions, a simulation study was designed to illustrate the similarities among the DINA model, DINO model, A -CDM, LLM, and R-RUM for an item with two attributes ($K_j^* = 2$). All examinees can be grouped into four latent classes (i.e., 00, 10, 01, and 11). For each of the five models (i.e., DINA, DINO, A -CDM, LLM, and R-RUM), the true probabilities of success of the latent classes were calculated. The same five models were used to approximate the probabilities of success of each generating model, and the dissimilarities between the models were calculated using Equation 7.

Simulation design. For generating the probability of success, $P(00)$ increased from 0.01 to 0.49 with a step of 0.02, and $P(11)$ increased from 0.51 to 0.99 also with a step of .02. For the DINA and DINO models, there were 625 success probabilities. For the A -CDM, LLM, and R-RUM, $P(10)$ increased from $P(00)$ to $P(11) - P(00)$ with a step of 0.02, resulting in more than 16,000 success probabilities for each model. The numbers of success probabilities were not identical because monotone constraints were set as in $P(00) \leq P(01)$, $P(10) \leq P(11)$. Note that no responses were simulated in this study, nor were the models fitted to data. Instead, only probabilities of success were generated, and the ϕ parameters of the models were estimated to minimize the difference in the probability of success between the two models. This excluded the confounding factors of simulation conditions from the results, allowing for a more direct comparison of the statistical properties of the models. This design allowed us to examine the extent $M2$ can approximate the success probabilities of $M1$.

Results. In Figure 1, the boxplots of the dissimilarity indices are given. Clearly, when the approximating model was the same as the generating model, the dissimilarity was equal to zero. When the DINA model was the generating model, the DINO model and A -CDM could not approximate the success probabilities (each with average dissimilarity of 0.5); the LLM and R-RUM approximated the probabilities quite well under many conditions, as evidenced by the average dissimilarity of 0.18 and 0.13, respectively. When the generating model was the DINO model, the DINA model, A -CDM, and R-RUM performed poorly with an average dissimilarity of 0.5. When the generating model was an additive model, the DINA and DINO models were typically not able to approximate the probabilities of success. For example, when generating model was R-RUM, the average dissimilarity of DINA and DINO was 0.36 and 0.51, respectively. However, the dissimilarity indices for the three additive models (i.e., A -CDM, LLM, and R-RUM) were generally small, and the average dissimilarities for all pairs of additive models were not greater than 0.15, implying that the probabilities of success generated by one of them were easily approximated by the others. The similarity of the additive models can account for the performance of the Wald test that is discussed in following sections.

Evaluating the Performance of the Wald Test for R-RUM and LLM

Although the Wald test was shown to be a promising approach to compare the G-DINA model against the DINA model, DINO model, and A -CDM (de la Torre & Lee, 2013), its statistical properties in terms of the LLM and R-RUM have not been established. The LLM and R-RUM are log and logit link G-DINA model with constraints, and previous studies did not investigate

Figure 1. Dissimilarities among different CDMs.

Note. M1 is the generating model and M2 is the approximating model. The vertical axis is the dissimilarity index. CDM = cognitive diagnosis model; DINA = deterministic inputs, noisy “and” gate; DINO = deterministic inputs, noisy “or” gate; A-CDM = additive CDM; LLM = linear logistic model; R-RUM = reduced reparametrized unified model.

whether the link function affects the performance of the Wald test. In addition, because the additive models (i.e., A-CDM, LLM, and R-RUM) are more similar to each other than they are to the DINA and DINO models, the power of the Wald test to distinguish among additive models

Table 1. Simulation Study Q-Matrix.

Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	0	0	0	0
7	0	1	0	0	0
8	0	0	1	0	0
9	0	0	0	1	0
10	0	0	0	0	1
11	1	1	0	0	0
12	1	0	1	0	0
13	1	0	0	1	0
14	1	0	0	0	1
15	0	1	1	0	0
16	0	1	0	1	0
17	0	1	0	0	1
18	0	0	1	1	0
19	0	0	1	0	1
20	0	0	0	1	1
21	1	1	1	0	0
22	1	1	0	1	0
23	1	1	0	0	1
24	1	0	1	1	0
25	1	0	1	0	1
26	1	0	0	1	1
27	0	1	1	1	0
28	0	1	1	0	1
29	0	1	0	1	1
30	0	0	1	1	1

requires close examination. Therefore, a simulation study was conducted to examine the Type I error and power of the Wald test for LLM and R-RUM.

Simulation Design

To evaluate the performance of the Wald test in selecting a reduced model, a simulation study was conducted, and five factors were considered, namely, sample size, item quality, attribute distribution, generating model, and fitted model. The Type I error was examined for LLM and R-RUM. For the power of the Wald test, the DINA model, DINO model, *A*-CDM, and R-RUM were used to analyze data generated from LLM; the DINA model, DINO model, *A*-CDM, and LLM were used to analyze data generated from R-RUM. The number of attributes and items were fixed to $K = 5$ and $J = 30$, respectively; the Q-matrix used in the study is given in Table 1.

The three sample sizes were $N = 500, 1,000$, and $2,000$. Item quality was defined by the slip (s) and guessing (g) parameters, which were equal to $1 - P(1)$ and $P(0)$, respectively. Items with $s, g \in U(0.05, 0.15)$ were labeled high quality, items with $s, g \in U(0.15, 0.25)$ medium quality, and items with $s, g \in U(0.25, 0.35)$ low quality. Attribute patterns followed two different distributions in this study: uniform distribution, where all possible attribute patterns are

equally likely, and higher order distribution (de la Torre & Douglas, 2004), where the probability of mastering attribute k for individual i is defined as

$$P_{ik} = \frac{\exp[1.7 \times (\theta_i - \delta_k)]}{1 + \exp[1.7 \times (\theta_i - \delta_k)]}.$$

Here θ_i is the ability of i th person, which is drawn from the standard normal distribution, and δ_k is the difficulty of the k th attribute. The difficulties of the five attributes were fixed as -1 , -0.5 , 0 , 0.5 , and 1 . When generating data was based on the A -CDM, LLM, and R-RUM, all main effects (ϕ_{jk}) for each item were set to be equal so that each required attribute had equal contribution to answering correctly. For example, when $K_j^* = 2$, the ϕ_{jk} for the LLM was calculated as follows:

$$\begin{aligned}\text{logit}[g_j] &= \text{logit}[P(00)] = \phi_{j0}, \\ \text{logit}[P(10)] &= \text{logit}[P(01)] = \phi_{j0} + \phi_j, \\ \text{logit}[P(11)] &= \text{logit}[1 - s_j] = \phi_{j0} + 2\phi_j,\end{aligned}$$

where $\phi_j = \phi_{j1} = \phi_{j2}$. For R-RUM, the main effects were computed in a similar fashion, with the only difference being that the log-link was used (e.g., $\log P(11)$ rather than $\text{logit}P(11)$).

For each condition, 500 replications were used. The G-DINA model was used to analyze all data sets, and the Wald statistics were calculated to evaluate whether each multi-attribute item could be fitted by different reduced models (i.e., DINA, DINO, A -CDM, LLM, and R-RUM) without significant loss of model-data fit. The code was written in Ox (Doornik, 2014).

Results

Type I error. Table 2 provides the Type I error rate at a significance level of .05 for the LLM and R-RUM across sample sizes, item qualities, attribute distributions, and number of required attributes. The results are summarized across items with the same K_j^* . When item quality was high, the Type I error rate was close to the nominal significance level (i.e., within $\pm .025$) across all conditions. In contrast, when item quality was low, all the Type I error rates were inflated. For medium item quality and uniform distribution, with one exception, the Type I error rate was close to the nominal α when $N \geq 1,000$; for the same item quality but higher order distribution, the Type I error was close to the nominal α when $K_j^* = 3$ and the model was LLM, or when $K_j^* = 2$, $N = 2,000$, and the model was R-RUM. The remaining conditions had inflated Type I error.

Power. The power of the Wald test for R-RUM when $K_j^* = 2$ is provided in Table 3. The results for the $K_j^* = 3$ condition as well as the results for LLM were similar, and hence, the corresponding tables have been omitted. These findings will be briefly discussed and are available, as are all other results, from the first author upon request. As with de la Torre and Lee (2013), a power of at least .8 is considered as adequate, and of at least .9 as excellent. Similarly, the Wald statistic was compared with the empirical χ^2 distributions to address any inflation of the Type I error. Item quality strongly influenced the observed power; the sample size did as well, though to a lesser extent. Referring to the $K_j^* = 2$ condition (Table 3), when item quality was high, the power of the Wald test was adequate or excellent (ranging from 0.81 to 1.00), with only two exceptions (i.e., 0.68 and 0.70), both of which occurred under $N = 500$, the higher order distribution, and the data fitted with the LLM. When item quality dropped to medium, the power was adequate or

Table 2. Type I Error of the Wald Test for the LLM and R-RUM ($\alpha = 0.05$).

K_j^*	Attribute distribution	Item quality	LLM			R-RUM		
			500	1,000	2,000	500	1,000	2,000
2	Higher order	High	0.045	0.047	0.042	0.042	0.045	0.041
		Medium	0.107	0.097	0.082	0.112	0.097	0.059
		Low	0.301	0.285	0.222	0.336	0.259	0.189
	Uniform	High	0.041	0.037	0.046	0.038	0.051	0.042
		Medium	0.077	0.064	0.056	0.072	0.047	0.069
		Low	0.331	0.253	0.150	0.329	0.212	0.163
3	Higher order	High	0.026	0.028	0.038	0.042	0.038	0.042
		Medium	0.047	0.044	0.042	0.125	0.124	0.121
		Low	0.221	0.220	0.154	0.400	0.399	0.379
	Uniform	High	0.053	0.057	0.034	0.039	0.051	0.039
		Medium	0.087	0.065	0.052	0.092	0.073	0.053
		Low	0.352	0.373	0.250	0.483	0.458	0.249

Note. LLM = linear logistic model; R-RUM = reduced reparametrized unified model.

better when the fitted model was DINA or DINO (with one exception), but was much lower for fitted models *A*-CDM and LLM. For example, under $N = 500$ and the higher order distribution condition, the power for fitted models *A*-CDM and LLM was smallest—0.28 and 0.17, respectively. The power of these two models was improved by increases in the sample size. For low-quality items, the power was very low, particularly when using the empirical distribution, with a minimum value of 0.06.

Similar patterns can be seen under the $K_j^* = 3$ condition. High-item quality and large sample size led to adequate or excellent power across all fitted models. When conditions were less favorable (i.e., smaller sample size, lower item quality), power dropped to below adequate level.

Although item quality and sample size strongly affected the observed power, the fitted model was of equal or greater importance. When the generating model was R-RUM or LLM, the power of the DINO model was the highest across all fitted models. Likewise, the power for the DINA model was typically higher than that of the additive models. The Wald test did not distinguish among the additive models well when the item quality was medium or low—this occurred when the generating model was R-RUM or LLM.

Attribute Classification Accuracy

In this section, the Wald test was applied in a simulation study to evaluate its performance as a model selection method. In particular, attribute classification based on the selected models were compared with those of the saturated model to examine the benefits the procedure can confer.

Simulation Design

In this simulation study, the factors controlled were sample size, item quality, attribute distribution, and generating model. The settings of the first three factors were the same as in the simulation study above. The generating models were the DINA model, DINO model, *A*-CDM, LLM, and R-RUM. When generating data was based on the *A*-CDM, LLM, and R-RUM, all main

Table 3. Power of the Wald test for the R-RUM ($\alpha = .05$ and $K_j^* = 2$).

N	Attribute distribution	Item quality	Theoretical distribution				Empirical distribution			
			DINA	DINO	A-CDM	LLM	DINA	DINO	A-CDM	LLM
500	Higher order	High	0.96	1.00	0.88	0.68	0.99	1.00	0.89	0.70
		Medium	0.79	1.00	0.40	0.27	0.82	1.00	0.28	0.17
	Uniform	Low	0.57	0.82	0.48	0.36	0.36	0.66	0.20	0.08
		High	0.98	1.00	1.00	0.81	0.99	1.00	1.00	0.81
1,000	Higher order	Medium	0.80	1.00	0.46	0.31	0.84	1.00	0.41	0.26
		Low	0.55	0.77	0.43	0.35	0.27	0.52	0.15	0.07
	Uniform	High	1.00	1.00	0.97	0.91	1.00	1.00	0.97	0.90
		Medium	0.95	1.00	0.54	0.41	0.96	1.00	0.40	0.28
2,000	Higher order	Low	0.66	0.91	0.42	0.30	0.43	0.78	0.16	0.06
		High	1.00	1.00	1.00	0.96	1.00	1.00	1.00	0.97
	Uniform	Medium	0.96	1.00	0.65	0.45	0.99	1.00	0.67	0.47
		Low	0.59	0.85	0.32	0.29	0.39	0.71	0.10	0.07
	Higher order	High	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.98
		Medium	1.00	1.00	0.69	0.55	1.00	1.00	0.63	0.50
	Uniform	Low	0.82	0.98	0.34	0.25	0.66	0.94	0.13	0.06
		High	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Uniform	Medium	1.00	1.00	0.91	0.74	1.00	1.00	0.89	0.72
		Low	0.71	0.96	0.24	0.22	0.65	0.94	0.11	0.09

Note. R-RUM = reduced reparametrized unified model; DINA = deterministic inputs, noisy “and” gate; DINO = deterministic inputs, noisy “or” gate; A-CDM = additive cognitive diagnosis model; LLM = linear logistic model.

effects (ϕ_{jk}) for each item were set to be equal so that each required attribute had equal contribution. Based on these settings, 100 data sets were generated for each condition.

To determine the most appropriate CDM for each item, the G-DINA model was fitted to each data set, and the Wald statistics for five reduced models were calculated for each item and the reduced models whose p values were less than 0.05 were rejected. If all reduced models were rejected for an item, the G-DINA model was used as the best model; if the procedure failed to reject more than one reduced CDM with the same simplicity (i.e., DINA and DINO, or A -CDM, LLM, and R-RUM), the model with the largest p value was selected as the most appropriate. In other words, if at least one reduced model was retained, and if (a) the DINA or DINO model was one of the retained models, then the model (i.e., DINA or DINO model) with the larger p value was selected as the best model; but if (b) both DINA and DINO were rejected, the reduced model with the largest p value was selected as the best model for this item. Note that when the p values of several reduced models were greater than 0.05, the DINA and DINO models were preferred over the A -CDM, LLM, and R-RUM because of their simplicity. After the best model for each item was determined, the model or the combination of different models for different items was estimated using the estimation methods presented below. It should be noted that the Wald test is only necessary for the items with two or more attributes.

The item parameters of the G-DINA model can be estimated using marginalized maximum likelihood estimation (MMLE) with the Expectation–Maximization (EM) algorithm (de la Torre, 2011). In this study, the item parameters of DINA, DINO, A -CDM, R-RUM, LLM, and G-DINA were estimated using a similar EM algorithm. Given that the estimation is conducted item by item, for a specific item j , in the E-step, $N(\alpha_{lj}^*)$, the number of examinees expected to be grouped into the latent class l and, $R(\alpha_{lj}^*)$, the expected number of examinees answering item j correctly in the latent class l were calculated. In the M-step, the following log-likelihood function was maximized:

$$\sum_{l=1}^{2^{K_j^*}} R(\alpha_{lj}^*) \log P(\alpha_{lj}^*) + [N(\alpha_{lj}^*) - R(\alpha_{lj}^*)] \log [1 - P(\alpha_{lj}^*)].$$

The stopping criterion (for the log-likelihood) was .001; person attributes were estimated using the expected a posteriori (EAP) method.

The person parameter recovery was evaluated using the proportion of correctly classified attribute vectors (PCV), as well as the proportion of correctly classified attributes (PCA). They are here defined as,

$$\text{PCV} = \frac{\sum_{r=1}^{\text{Rep}} \sum_{i=1}^N I[\alpha_i = \hat{\alpha}_i]}{N \times \text{Rep}}, \text{ and } \text{PCA} = \frac{\sum_{r=1}^{\text{Rep}} \sum_{i=1}^N \sum_{k=1}^K I[\alpha_{ik} = \hat{\alpha}_{ik}]}{N \times K \times \text{Rep}},$$

where Rep is the number of replications, and $I[\alpha_i = \hat{\alpha}_i]$ and $I[\alpha_{ik} = \hat{\alpha}_{ik}]$ evaluate whether the estimated attribute vector and the estimated individual attributes, respectively, match the generated values.

Results

The classification rates when attributes followed the higher order distribution are given in Table 4. The true model was also fitted to the data to serve as the baseline. The results show that across all conditions, the true model always provided the most accurate attribute classification,

Table 4. PCV of α From Higher Order Distribution (Percentage).

N	Generating model	Low quality			Moderate quality			High quality		
		True	Fitted	G-DINA	True	Fitted	G-DINA	True	Fitted	G-DINA
500	DINA	46.11	38.30	36.30	73.52	72.74	71.00	92.08	92.01	91.59
	DINO	46.60	31.88	26.08	73.51	72.01	69.40	91.98	91.88	91.29
	A-CDM	33.87	29.58	26.13	64.60	63.73	61.69	88.71	88.47	88.12
	LLM	34.02	30.70	25.81	65.15	64.43	61.87	89.11	89.04	88.56
1,000	R-RUM	33.37	30.52	27.45	65.98	64.96	63.29	89.18	88.97	88.75
	DINA	48.47	44.54	42.81	73.71	73.31	72.43	92.19	92.11	91.92
	DINO	48.37	38.71	34.44	74.12	73.84	73.05	92.11	92.06	91.91
	A-CDM	39.68	36.08	33.66	66.35	65.91	64.90	89.26	89.16	88.97
2,000	LLM	39.25	35.77	33.33	66.25	65.90	64.85	89.67	89.60	89.41
	R-RUM	39.78	37.01	34.77	66.89	66.48	65.52	89.76	89.73	89.52
	DINA	49.52	48.18	46.96	74.49	74.40	73.94	92.15	92.11	92.01
	DINO	49.43	45.68	43.34	74.10	73.98	73.69	92.10	92.06	91.97
	A-CDM	42.19	40.86	39.24	67.15	66.93	66.51	89.56	89.52	89.44
	LLM	42.24	41.07	39.51	66.79	66.61	66.21	89.58	89.56	89.47
	R-RUM	42.09	41.00	39.65	67.54	67.22	66.94	89.90	89.87	89.78

Note. PCV = proportion of correctly classified attribute vectors; G-DINA = generalized DINA; DINA = deterministic inputs, noisy “and” gate; DINO = deterministic inputs, noisy “or” gate; A-CDM = additive cognitive diagnosis model; LLM = linear logistic model; R-RUM = reduced reparametrized unified model.

followed by the CDMs selected by the Wald test; the G-DINA model performed the worst. For example, when $N = 500$, item quality was low and generating model was LLM, the attribute classification rate for models selected by the Wald test was 30.70%, which is lower than the classification rate of 34.02% from the true model, but higher than the classification rate of 25.81% from the G-DINA model. In addition, when item quality was low and sample size was small (i.e., $N = 500$), the differences between the classification rates of the selected model and the saturated model were more apparent. For example, from Table 4, when the generating models were the DINO and DINA model, the selected models outperformed the G-DINA model by 5.8% and 2.0%, respectively. The corresponding differences for the A -CDM, LLM, and R-RUM models ranged from 3.1% to 4.9%. However, when the sample size was large and item quality was high, the selected models performed similarly to the G-DINA model, and both provided classification rates comparable with the true model. For instance, when $N = 2,000$ and item quality was high, the differences in classification rate for the true model, selected model, and the G-DINA model were not greater than 0.14%. Furthermore, attribute distribution affected the classification. In particular, when the attributes had a higher order distribution—a more realistic assumption than the uniform distribution—the selected models led to larger improvement. Similar patterns occurred in the case of the classification of individual attributes, or the PCA, though the magnitude of improvement was attenuated.

An Empirical Example

In this example, an empirical analysis was conducted to illustrate the application of the Wald test to evaluate the most appropriate CDM at the item level with real data. The data used were obtained from a Dutch-language version of the Millon Clinical Multiaxial Inventory-III (MCMI-III; Millon, Millon, Davis, & Grossman, 2009; Rossi, van der Ark, & Sloore, 2007), one of the most widely used self-report instruments for clinical assessment (Camara, Nathan, & Puente, 2000). It is designed to identify patients with particular mental disorders. Because of the comorbidity of mental disorders, many items in this instrument simultaneously measure several categories or disorders. The CDM framework can account for the multidimensional nature of the items.

Thirty items were used in the analysis, with the clinical scales somatoform (Scale H), thought disorder (Scale SS), and major depression (Scale CC) functioning as the attributes (i.e., $K = 3$). Each of these items measures one to three studied disorders. The item numbers as they were used in the current study, and appear in the MCMI-III manual (Millon et al., 2009), are given in Table 5.

To illustrate the impact of item quality on classification rates, two 10-item tests (i.e., A and B) were constructed, where Test A consisted of the more highly discriminating items and Test B of less discriminating items. The discrimination of each item was determined by calculating the G-DINA model discrimination divergence, which is an empirical tool, developed by de la Torre and Chiu (2015), for identifying and correcting misspecified entries in a Q-matrix. Tests A and B had identical Q-matrices, and each test included five one-attribute items, four two-attribute items, and one shared item with three attributes. Two sets of responses of 300 examinees to the items on Tests A and B were drawn randomly from a total of 739 examinees and analyzed.

To examine the improvements of the Wald test in attribute classification, a baseline examinee attribute specification had to be established for comparison. Establishing attribute specification based on a longer test with larger sample size allowed us to then compare the increase in classification agreement that can be achieved by using the proposed procedure. The baseline examinee attribute specification was found by fitting the G-DINA model to the full data set,

Table 5. Q-Matrix for the Empirical Analysis.

Item number		H	SS	C
Current	Manual			
1 ^a	1	1	0	1
2 ^a	4	1	0	1
3 ^b	11	1	0	0
4	22	0	1	0
5 ^a	34	0	1	1
6 ^a	37	1	0	0
7 ^a	44	0	0	1
8 ^a	55	1	0	1
9 ^a	56	0	1	0
10 ^a	68	0	1	0
11	72	0	1	0
12	74	1	0	1
13 ^b	78	0	1	0
14 ^a	83	0	1	0
15 ^b	102	0	1	0
16	104	0	0	1
17 ^b	107	1	0	1
18 ^b	111	1	0	1
19 ^b	117	0	1	0
20	128	0	0	1
21 ^b	130	1	0	1
22	134	0	1	0
23	142	0	1	1
24 ^{ab}	148	1	1	1
25 ^b	150	0	0	1
26 ^b	151	0	1	1
27	154	0	0	1
28	162	0	1	0
29	168	0	1	0
30	171	0	0	1

Note. The column labeled “Manual” provides the corresponding item numbers in the MCMI-III manual (Millon, Millon, Davis, & Grossman, 2009). Please refer to pages 175-179 of the manual for item descriptions. MCMI-III = Millon Clinical Multiaxial Inventory-III.

^aThe items in Test A.

^bThe items in Test B.

including all 30 items and 739 respondents. Using these results to create the Q-matrix, the examinee attributes were then estimated based on the selected CDMs for Test A and Test B.

This involved two steps: first, the G-DINA model was fitted to the two data sets, each consisting of 300 examinees and 10 items. In the second step, the Wald test was used, and the most appropriate CDM for each multi-attribute item was determined based on the same model selection rule as described in the simulation study above. The *p* values of the Wald statistics and the selected model for multi-attribute items are given in Table 6. Based on the selected models, the two tests were calibrated and respondents’ attributes patterns were estimated. To compare the classification rate of selected models with the G-DINA model, the attribute profiles of these 300 respondents based on the G-DINA model were estimated.

PCV and PCA were calculated from the attribute patterns estimated based on selected models by the Wald test and the G-DINA and are given in Table 7. Compared with the G-DINA model,

Table 6. The *p* values of the Wald Test.

Test	Item	<i>p</i> values					Selected Model
		DINA	DINO	A-CDM	LLM	R-RUM	
A	1	.00	.00	.96	.50	.01	A-CDM
	2	.00	.00	.54	.84	.33	LLM
	5	.01	.00	.32	.46	.52	R-RUM
	8	.00	.00	.78	.14	.00	A-CDM
	24	.00	.00	.00	.18	.01	LLM
B	17	.09	.02	.24	.73	.94	DINA
	18	.00	.02	.49	.12	.02	A-CDM
	21	.00	.00	.69	.82	.26	LLM
	24	.00	.00	.00	.45	.00	LLM
	26	.13	.00	.17	.83	.74	DINA

Note. Entry 0 represents the *p* value is less than .01. DINA = deterministic inputs, noisy “and” gate; DINO = deterministic inputs, noisy “or” gate; A-CDM = additive cognitive diagnosis model; LLM = linear logistic model; R-RUM = reduced reparametrized unified model.

Table 7. Classification Rates of G-DINA and Selected Models (in Percentage).

Test	PCV		PCA	
	G-DINA	Selected	G-DINA	Selected
A	75.3	75.7	89.2	89.7
B	61.7	66.7	81.6	84.1

Note. G-DINA = generalized DINA; Selected = models selected based on Wald test; PCV = proportion of correctly classified attribute vectors; PCA = proportion of correctly classified attributes.

when item discrimination was high, the selected CDMs provided very limited improvement in terms of the PCA and PCV. When item discrimination was low, the selected CDMs offered a 5% improvement in terms of the PCV (61.7%-66.7%) and a 2.5% improvement in terms of the PCA (81.6%-84.1%).

Discussion

This study extended the findings of de la Torre and Lee (2013) in several ways. First, it evaluated the power and Type I error rate of the Wald test across several popular additive models—the LLM and R-RUM. Second, it expanded the conclusions and implications of the 2013 study by demonstrating that not only can the Wald test identify correct reduced models, but it can also be used to improve attribute classification, especially under unfavorable conditions. Not only are simpler models preferred for their parsimony, using the correct reduced model leads to better attribute classification (Rojas et al., 2012). The Wald test provides a way to empirically select the reduced models, rather than assume them a priori.

Analysis of real test data here and elsewhere (e.g., de la Torre & Lee, 2013) shows that no one model can be used for all the test items. These results indicated that in many, if not all, test applications, no single reduced model can be expected to satisfactorily fit all the items. The Wald test clears this impasse by empirically selecting a reduced model for each item.

This study also shows that additive models can present some additional hurdles in terms of model selection due to the functional similarity of the *A*-CDM, LLM, and R-RUM. Although the Wald test did not differentiate well among the three, extending the procedure to these models revealed that the procedure can still improve attribute classification even when the generating model was *A*-CDM, LLM, or R-RUM. In fact, the selection of the true additive model was not essential to improving the classification rate over the G-DINA model. This is not surprising, given that the similarity study suggested that each additive model could closely re-create the latent group success probabilities of the other additive models. Beyond that, however, it is not clear how functional similarity affects the performance of the Wald test. For example, it is not clear to what extent the additive models must be dissimilar for the Wald test to exhibit adequate or good power. Further research into the relationship between model similarity and the power of the Wald test is needed. This study only investigated dissimilarities among different reduced CDMs when $K_j = 2$. To better understand the index, its behavior when items measure more than two attributes needs to be investigated.

This article reveals that the Type I error of the Wald test can be inflated under certain conditions. One way of addressing the inflation of Type I error is by using a different method of computing the standard errors. Another option would be to adjust the significance level by accounting for the multiple comparisons. Alternatively, another test entirely, such as the likelihood ratio test or Lagrange multiplier test, could be used to select a model at the item level. Although these different tests would most likely perform similarly with large samples, it is not clear how they would perform, comparatively, under less favorable test conditions, such as modest sample sizes and lower item qualities. Implementing one or more of these procedures may lead to better control of the Type I error and possibly improved attribute classification.

Finally, it is worth noting that the generalizability of the findings of this study are dependent upon the limitations of the design of the simulation study, such as using fixed test length, assuming the *Q*-matrix is correct, and constraining all the attributes of the additive model to be equal. To further generalize these findings, future research should consider a wider range of conditions with fewer constraints.

Acknowledgments

The authors thank Gina Rossi for access to the data used in the Empirical Example section.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by National Science Foundation Grant DRL-0744486.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.
- Beck, L. W. (1943). The principle of parsimony in empirical science. *The Journal of Philosophy*, 40, 617-633.
- Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, 31, 141-154.

- Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, 50, 123-140.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179-199.
- de la Torre, J., & Chiu, C.-Y. (2015). A general method of empirical Q-matrix validation. *Psychometrika*. Advance online publication. doi:10.1007/s11336-015-9467-8
- de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.
- de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, 50, 355-373.
- Doornik, J. A. (2014). Ox: An object-oriented matrix language (Version 7.00) [Computer software]. Available from <http://www.doornik.com/>
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 301-321.
- Hartz, S. M. (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.
- Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74, 191-210.
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, 64, 187-212.
- Millon, T., Millon, C., Davis, R., & Grossman, S. (2009). *MCMI-III manual* (4th ed.). Minneapolis, MN: Pearson Assessments.
- Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Paper presented at the annual meeting of the National Council of Measurement in Education, Vancouver, British Columbia, Canada.
- Rossi, G., van der Ark, L., & Sloore, H. (2007). Factor analysis of the Dutch-language version of the MCMI-III. *Journal of Personality Assessment*, 88, 144-157.
- Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345-354.
- Templin, J. L., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287-305.
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61, 287-307.