

Improving Accuracy and Usage by Correctly Selecting: The Effects of Model Selection in Cognitive Diagnosis Computerized Adaptive Testing

Applied Psychological Measurement
2021, Vol. 45(2) 112–129
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621620977682
journals.sagepub.com/home/apm



Miguel A. Sorrel¹ , Francisco José Abad¹,
and Pablo Nájera¹ 

Abstract

Decisions on how to calibrate an item bank might have major implications in the subsequent performance of the adaptive algorithms. One of these decisions is model selection, which can become problematic in the context of cognitive diagnosis computerized adaptive testing, given the wide range of models available. This article aims to determine whether model selection indices can be used to improve the performance of adaptive tests. Three factors were considered in a simulation study, that is, calibration sample size, Q-matrix complexity, and item bank length. Results based on the true item parameters, and general and single reduced model estimates were compared to those of the combination of appropriate models. The results indicate that fitting a single reduced model or a general model will not generally provide optimal results. Results based on the combination of models selected by the fit index were always closer to those obtained with the true item parameters. The implications for practical settings include an improvement in terms of classification accuracy and, consequently, testing time, and a more balanced use of the item bank. An R package was developed, named *cdcatR*, to facilitate adaptive applications in this context.

Keywords

cognitive diagnosis models, computerized adaptive testing, model comparison, G-DINA, classification accuracy, item usage

Adaptive testing methodologies, originally developed in the context of traditional item response theory (IRT), are being generalized to more complex scenarios, including cognitive diagnostic computerized adaptive testing (CD-CAT; for a review, see Akbay & Kaplan, 2017; Huebner, 2010). CD-CAT is based on cognitive diagnosis models (CDMs), which are specifically

¹Autonomous University of Madrid, Spain

Corresponding Author:

Miguel A. Sorrel, Department of Social Psychology and Methodology, Autonomous University of Madrid, Ciudad Universitaria de Cantoblanco, Madrid 28049, Spain.

Email: miguel.sorrel@uam.es

developed to detect mastery and nonmastery of set of latent fine-grained skills. Some of the decisions that will affect the performance of the adaptive algorithms in this context involve the internal structure of the test specified in the Q-matrix and model selection. The identification of the Q-matrix is a laborious process where many professionals are typically involved. For example, in the development of a Q-matrix for a proportional reasoning test, a diverse group composed of mathematics researchers, mathematics educators, middle school teachers, and graduate students in psychometrics and mathematics education was involved in Tjoe and de la Torre (2014)'s study. Examinees are also generally considered using think-aloud protocols to validate the theoretical framework (e.g., Li & Suen, 2013). The usual next step is to evaluate the initial Q-matrix using empirical Q-matrix validation methods and evaluating the fit of the different specifications (e.g., Sorrel et al., 2016).

Arguably, the element that has received less attention is model selection. How to choose among the wide range of CDMs available is not an easy decision. Each CDM assumes a different cognitive process involved in responding to an item (e.g., conjunctive, disjunctive, or additive condensation rules). Besides, many CDMs have been created ranging in complexity. In this sense, recent developments have produced general CDMs such as the *generalized deterministic inputs, noisy, "and" gate* (G-DINA; de la Torre, 2011) model, the general diagnostic model (GDM; von Davier, 2005), and the log-linear CDM (LCDM; Henson et al., 2009). Reduced models are nested within these general models. Examples of reduced CDMs are the *deterministic inputs, noisy "and" gate* (DINA; Haertel, 1989) model, the *deterministic inputs, noisy "or" gate* (DINO; Templin & Henson, 2006), and the *additive* CDM (A-CDM; de la Torre, 2011). Due to its relative novelty, CD-CAT empirical applications are still scarce. A trend may be noted, however, toward the use of the same CDM for all the items in the item bank. For example, H. Y. Liu et al. (2013) applied a noncompensatory CDM, the DINA model, to a 352-item English language proficiency item bank. Reduced models have been widely used by researchers because of their simplicity of estimation and interpretation. However, these models make strong assumptions on the data and that's why their fit to the actual data should be evaluated.

Compared to reduced models, general models offer the advantage of wide applicability; they allow for all types of condensation rules within the same test. This is the case, for example, of the CD-CAT application by Sorrel, Yigit, and Kaplan (2017) where the authors applied the G-DINA model to a 76-item proportional reasoning item bank. This alternative might be more appropriate since previous empirical studies have shown that no reduced model can be deemed appropriate for all test items (e.g., de la Torre et al., 2018; de la Torre & Lee, 2013; Ravand, 2016). Nevertheless, accurate calibration of parameters in a general model is more dependent on data conditions (e.g., sample size), and the risk of capitalization of chance is higher. Owing to these shortcomings, researchers introduced item-level model comparison indices like the likelihood ratio (LR) for the purpose of relative-fit evaluation (Sorrel, Abad, et al., 2017). This allows for an intermediate situation between the two extremes (i.e., single reduced CDM vs general model). The idea is to select the most appropriate CDM for each item. A further development on the LR test, the two-step LR test (2LR) demonstrated promise as a tool for assessing item relative fit in CDMs (Sorrel, de la Torre, et al., 2017). Importantly, the 2LR test is expected to perform very well under the usual item bank calibration conditions, typically involving a large number of items (Sorrel, Abad, et al., 2017; Sorrel, de la Torre, et al., 2017).

According to previous research, model selection might have an impact on classification accuracy (Ma et al., 2016; Rojas et al., 2012) and the generalization of the item parameter estimates (Olea et al., 2012). In that respect, Rojas et al. (2012) found that single reduced models, when appropriate, led to a better classification accuracy compared to general models. This was more notable in poor-quality conditions, where it was more difficult to estimate the general model (e.g., small sample size and low item discrimination). In the context of IRT, Olea et al.

(2012) explored the consequences of fitting a complex model under poor-quality item bank calibration conditions. They found that a parameters of the three-parameter logistic model were overestimated, causing an overestimation of the precision of the trait-level estimates. Despite its potential benefits, relative model fit is not systematically evaluated in empirical applications. According to the review by Sessoms and Henson (2018), most of the studies estimated reduced CDMs like the DINA model, and 28% of them did not report model fit.

Therefore, all together, previous research indicates that test calibration conditions are highly related to the accuracy of the model estimates. This would be of major importance in the context of adapting testing where items are selected based on their parameter estimates. Considering all above, this study investigates whether item-level model comparison indices can be useful to improve CD-CATs performance in terms of classification accuracy and item usage. The rest of the article is structured as follows. First, a detailed overview on CDM, item-level model comparison, and CD-CAT is provided. Second, the design of the simulation study is described, and the results under the different conditions are presented. Finally, several implications and limitations of this study are discussed.

Cognitive Diagnosis Modeling

CDMs are confirmatory latent class models that are receiving increasing attention in the literature (for an overview of these models see, e.g., Rupp & Templin, 2008). The goal of CDM is to classify respondents as masters or nonmasters on a set of prespecified list of K discrete attributes (e.g., skills, cognitive processes, and disorders). This latent attribute vector or latent class can be denoted by α_l , for $l = 1$ to 2^K . These models emerged in the field of education and have also been applied in other settings such as clinical psychology and competency modeling (e.g., Sorrel et al., 2016; Templin & Henson, 2006). Multiple models have been proposed. Most of these models are represented in general CDMs, such as the aforementioned G-DINA model. For each item, this model partitions the latent classes into $2^{K_j^*}$ latent groups, where K_j^* is the number of attributes being measured by item j . The item response function (IRF) of the G-DINA model is then given by:

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

where δ_{j0} is the intercept or baseline probability for item j , δ_{jk} is the main effect due to α_{lk} , $\delta_{jkk'}$ is the interaction effect due to α_{lk} and $\alpha_{lk'}$, and $\delta_{j12\dots K_j^*}$ is the interaction effect due to $\alpha_1, \dots, \alpha_{K_j^*}$. Thus, there are $2^{K_j^*}$ parameters to be estimated for item j .

Reduced CDMs can be formed by constraining some of the parameters of this general model. In this article, we consider three of these reduced models: DINA, DINO, and A -CDM. In the DINA model, all terms in Equation 1 except the baseline probability and the highest interaction term are set to 0. The DINA model only has two parameters per item: the guessing parameter represented by δ_{j0} , and the slip parameter is represented by $1 - (\delta_{j0} + \delta_{j12\dots K_j^*})$. This is a non-compensatory model where the highest probability of success is only achieved when all the attributes required by the item have been mastered. On the contrary, the DINO model is a compensatory model. There are also only two parameters per item, namely δ_0 and δ_{j1} , with the important exception that δ_{j1} is constrained so that some lower-order terms will be canceled by the corresponding higher-order terms (de la Torre, 2011). Respondents will have a high probability of success, provided they master at least one required attribute. Finally, the A -CDM is an additive model where all the interaction terms are dropped, and thus each mastered attribute contributes to the probability of success independently. Given that all these reduced models are

nested within the more general G-DINA model, item-level model comparison statistics can be computed to compare the relative fit of different models.

Item-Level Model Comparison

Among all the competing models, the Occam's razor principle dictates that the simplest should be chosen. One of the reasons for doing that is to avoid the capitalization on chance problem. Different studies pointed out that particularly when the sample size is small and the item quality is poor, an appropriate reduced CDM will lead to a higher accuracy (Ma et al., 2016; Rojas et al., 2012). The advantage of using a reduced CDM will be greater the more complex the item structure is. In the case of minimum complexity (i.e., one-attribute items), it is irrelevant which CDM is applied because all of them are equivalent. In contrast, more complex items will lead to different IRF specifications according to the different CDMs. In this sense, the DINA and DINO models will always have two parameters per item regardless of the item complexity, but the number of item parameters will linearly and exponentially grow for the A-CDM and G-DINA models, respectively. Sample size requirements for estimation of item parameters in complex structures will be stricter.

Model comparison can be conducted at the item level using relative fit statistics such as the Wald and LR tests (for a comparison of these tests see, e.g., Sorrel, Abad, et al., 2017; Sorrel, de la Torre, et al., 2017). They allow selecting the most appropriate reduced model for each item. The traditional implementation of the LR test requires both the general and the reduced CDM to be estimated. Sorrel, de la Torre, et al. (2017) proposed a more efficient approximation with the advantage of only requiring the more general model to be estimated, referred to as two-step LR test (2LR). Item-level maximum likelihoods for the competing models are estimated using the following formula:

$$L(\boldsymbol{\varphi}_j | \mathbf{R}_j, \mathbf{I}_j) = \prod_{l=1}^{2^{K_j^*}} P^{(m)}(\boldsymbol{\alpha}_{lj}^*)^{R_{\alpha_{lj}^*}} [1 - P^{(m)}(\boldsymbol{\alpha}_{lj}^*)]^{(I_{\alpha_{lj}^*} - R_{\alpha_{lj}^*})}, \quad (2)$$

where $P^{(m)}(\boldsymbol{\alpha}_{lj}^*)$ represents the probability of correctly answering the item j for respondents in the latent group l based on the item parameters of the model of interest, $\mathbf{R}_j = \{R_{\alpha_{lj}^*}\}$, and $\mathbf{I}_j = \{I_{\alpha_{lj}^*}\}$. $I_{\alpha_{lj}^*}$ and $R_{\alpha_{lj}^*}$ represent, respectively, the number of respondents in latent group l and the number of respondents in latent group l who correctly answered the item, both of them based on the attribute joint distribution estimated for the more general model. The null hypothesis states that the likelihoods of the reduced and general models are equal, meaning that the fit of the reduced model is not significantly worse than the fit of the general model. When the null hypothesis is not rejected, the reduced model would be preferred due to the reasons mentioned above. The 2LR statistic is computed as

$$2LR_j = 2 \left[\log L(\mathbf{P}_j | \mathbf{R}_j, \mathbf{I}_j) - \log L(\boldsymbol{\varphi}_j | \mathbf{R}_j, \mathbf{I}_j) \right], \quad (3)$$

where \mathbf{P}_j and $\boldsymbol{\varphi}_j$ are the vectors of G-DINA and reduced model item parameters, respectively. The statistic is asymptotically chi-square distributed with $2^{K_j^*} - np_j$ degrees of freedom, where np_j indicates the number of reduced model parameters for item j . The Wald and 2LR statistics performed well under different scenarios of sample size, test length, generating model, and item quality, with the 2LR test providing slightly better Type I error and power rates (Sorrel, de la Torre, et al., 2017).

Cognitive Diagnosis Computerized Adaptive Testing

This is a new area of application that has been aided from the developments in traditional CAT. Unfortunately, because latent variables in CDMs are discrete, item selection methods based on the Fisher information cannot be applied in CD-CAT. However, several item selection indices have been proposed for CD-CAT, including the G-DINA model discrimination index (GDI; Kaplan et al., 2015). This index yielded shorter test administration times compared to the other item selection methods (e.g., modified posterior weighted Kullback–Leibler). The next item to be selected by the adaptive algorithm is the one with the highest GDI:

$$s = \operatorname{argmax}_{j \in B_q} GDI = \sum_{l=1}^{2^k} \pi(\alpha_{lj}^*)^{(t)} [P(\alpha_{lj}^*) - \bar{P}_j]^2, \quad (4)$$

where α_{lj}^* defines a reduced latent attribute pattern, $\pi(\alpha_{lj}^*)^{(t)}$ the posterior probability of α_{lj}^* at step t of the algorithm, $P(\alpha_{lj}^*)$ the conditional probability of success on item j given by the pattern α_{lj}^* , and the average success probability is computed as $\bar{P}_j = \sum_{l=1}^{2^{K_j^*}} \pi(\alpha_{lj}^*)^{(t)} P(\alpha_{lj}^*)$. It can be noted from Equation 4 that GDI does not rely on a point estimator of α_i . Instead, the index considers all the information available at the current step (i.e., $\pi(\alpha_{lj}^*)^{(t)}$). This is one strength of the method as it allows avoiding the problem of multiple maxima that is likely to occur at the initial steps of a CAT (Sorrel, Barrada, et al., 2020).

Goal of the Present Study

This study aims to explore the impact of item bank calibration on the CD-CAT performance. Specifically, it is assessed to what extent a better performance can be obtained when an appropriate reduced CDMs is chosen for each item using the 2LR test. Hypothetically, the 2LR test will show a very good performance under the usual item bank calibration conditions. Thus, it is expected that this index will be useful in improving CD-CAT performance. Compared to a situation where a general model is estimated for all the items, a combination of models derived by the 2LR test will require estimating fewer parameters. Thus, these parameters will be estimated more accurately, having an impact on the classification accuracy. In addition, item usage under the different item bank calibration conditions will be explored. A simulation study was conducted to address these research questions. Only low item discrimination conditions are considered because a larger improvement can be expected in these situations.

Method

A simulation study was conducted to evaluate the classification accuracy and item usage obtained with each of the model calibrations described: G-DINA, 2LR-derived combination of models, DINA, DINO, and A -CDM. For comparison purposes, true item parameters were also considered, which allows obtaining an estimation of the upper limit for the classification accuracy. Factors and levels were selected based on a literature review of current CDM and CD-CAT empirical applications. Three data factors were varied, namely the calibration sample size ($N = 250, 500$, and $2,000$ respondents), the item bank length ($J = 165$ and 330 items), and the Q-matrix complexity ($Q\text{-str} = \text{simple and complex Q-matrix structure}$). More specifically, Q-matrix complexity was understood as the number of attributes being measured by each of the items. Two levels were considered for this factor. In the simple Q-matrix condition, 45 one-, 60 two-, and 60 three-attribute items were generated. On the contrary, in the complex Q-matrix

condition, 60 two-, 60 three-, and 30 four-attribute items were generated, and 15 additional one-attribute items were also included to ensure identifiability of the Q-matrix (Xu, 2017). In the $J = 330$ item conditions, these numbers were doubled.

The 2LR test is an inferential test and thus a significance level needs to be selected. We report the results for $\alpha = .05$ (2LR-None) and also correcting for multiple comparisons using the *Holm* correction procedure (2LR-Holm). The following procedure was used to determine the most appropriate CDM for each item. All reduced models (i.e., tests) whose p -values were significant were rejected. All reduced models with a nonsignificant p -value defined the set of candidate reduced models A_j . The set A_j can include all combinations by taking 0 to 3 elements from the list of tested reduced models (i.e., DINA, DINO, and A-CDM). The G-DINA model was retained if all reduced models were rejected at this step (i.e., $A_j = \{\}$). Whenever the set A_j included more than one element, the model with the largest p -value was selected as the best model for that item. We expected a better performance of the procedure based on the adjusted p -values (i.e., 2LR-Holm) considering the fact that a large number of tests were being considered. For example, in the 165 items and simple Q-matrix condition, there were 120 items measuring more than one attribute. Given that we considered three possible reduced models, 360 tests were conducted. Multiple comparison corrections might be conservative. This can be translated into an increase in the number of “false” reduced model candidates in the set A_j . Choosing the model with the largest p -value provided a higher classification accuracy compared to using the G-DINA model alone in a previous study (Ma et al., 2016).

Ten item banks were constructed for each simulated condition. The following text describes these item banks in terms of item discrimination, number of attributes, and data generating model. Item parameters were generated randomly from the following distributions: $P(0) \sim U(0.20, 0.40)$ and $P(1) \sim U(0.60, 0.80)$. For the A-CDM model, the main effect of each attribute was set to be $P(0) + (P(1) - P(0))/K_j^*$. This condition has been referred to as low item quality in previous research (e.g., Ma et al., 2016; Sorrel, Abad, et al., 2017; Sorrel, de la Torre, et al., 2017). The number of attributes was fixed to 5, which is a reasonable value considering current CDM empirical applications (Sessoms & Henson, 2018) and simulation studies (e.g., Cheng, 2009, 2010; Kaplan et al., 2015). The generating, true model used in the data generation process was always a combination of the same number of DINA, DINO, and A-CDM items.

For each condition and item bank, we generated a validation sample consisting of 5,000 response patterns generated uniformly from the space of possible $2^5 = 32$ latent classes. A CD-CAT based on each of the model calibrations was applied to each of these validation samples. The first item was randomly chosen from the medium discriminating items in the item bank. The item selection rule was the G-DINA discrimination index (GDI; Kaplan et al., 2015). Different studies have discussed the relationships among several of the item selection rules available and have highlighted *the strong relationship among them* (Wang et al., 2020; Yigit et al., 2019). GDI was chosen on the basis of computational time. Results can be expected to be generalizable to other item selection rules. Conditional results on different CD-CAT length conditions were explored: starting from five up to 30 items. However, most of the results were described assuming that the CD-CAT length was fixed to 30 items (i.e., fixed-length stopping rule), a reasonable test length that provides sufficient classification accuracy considering prior research (Kaplan et al., 2015). The scoring method was the maximum likelihood estimation method.

To assess the quality of the item bank calibration, we compared the item parameters estimates with the true item parameters. The root mean squared error (RMSE) was computed for each method, and averaged across the item banks. The formula for RMSE is

Table 1. Model Selection Rates for the 2LR Test ($\times 100$).

N	J	Q-str	MC correction					
			None			Holm		
			✓	G	×	✓	G	✓
250	165	Complex	66	17	17	79	0	21
		Simple	79	7	14	85	0	15
	330	Complex	80	6	14	85	0	15
		Simple	85	4	12	88	0	12
500	165	Complex	76	18	6	92	1	7
		Simple	89	6	6	94	0	6
	330	Complex	90	5	5	95	0	6
		Simple	92	4	4	95	0	5
2,000	165	Complex	90	10	0	100	0	0
		Simple	94	6	0	100	0	0
	330	Complex	94	6	0	100	0	0
		Simple	95	5	0	100	0	0
	Grand mean		86	8	7	93	0	7

Note. Cells with values higher or equal than 80 are shown in bold. 2LR = two-step likelihood ratio; ✓/× = True (generating)/False (nongenerating) reduced CDM is retained; G = G-DINA is retained; MC = multiple comparison; N = calibration sample size; J = item bank length; Q-str = Q-matrix structure; CDM = cognitive diagnosis model; G-DINA = generalized deterministic inputs, noisy, "and" gate.

$$RMSE = \sqrt{\sum_{j=1}^J \sum_{l=1}^{2^K} \frac{(P_j(\alpha_l) - \hat{P}_j(\alpha_l))^2}{J \cdot 2^K}}, \quad (5)$$

where $P_j(\alpha_l)$ and $\hat{P}_j(\alpha_l)$ represent the true and estimated probabilities of success for item j associated to latent class l , respectively. Two dependent variables were used for the comparison between the different CD-CAT applications: pattern recovery (i.e., proportion of correctly classified vectors) and item usage for the different item types. Item usage was computed from the item exposure rates. We computed the relative proportion of items administered within each item type category (i.e., q -vector complexity or generating model). Model estimation was conducted using the GDINA *R* package (Ma & de la Torre, 2017). Finally, we developed the *R* package *cdcatR* (Sorrel, Nájera, & Abad, 2020) that allows for application of the 2LR and CD-CAT analyses in *R*.

Results

Calibration Sample Results: Model Selection

Table 1 includes the results for the average 2LR performance across all the test bank replications.¹ The overall performance was generally acceptable regardless of the condition. The large number of comparisons caused that the Type I error rate increased. Accordingly, the correct reduced CDM was selected more frequently when the Holm correction was used (grand mean of 86 vs 93). Even in the small calibration sample size conditions, the true reduced model was selected at least in 79% of the items when the p -value was adjusted for multiple comparisons. To facilitate the interpretation of the effects of calibration sample size, item bank length, and Q-matrix complexity a graphical representation is included in Figure 1. The calibration sample size factor had the largest effect. Small sample size conditions affected the power of the

Table 2. Average Root Mean Squared Error Results.

Calibration sample size	Item bank length	Q-matrix structure	Model		
			G-DINA	2LR-None	2LR-Holm
250	165	Complex	0.101	0.081	0.064
		Simple	0.070	0.058	0.054
	330	Complex	0.088	0.062	0.056
		Simple	0.066	0.053	0.051
500	165	Complex	0.067	0.051	0.040
		Simple	0.048	0.038	0.036
	330	Complex	0.058	0.040	0.036
		Simple	0.046	0.036	0.034
2,000	165	Complex	0.031	0.021	0.018
		Simple	0.024	0.018	0.016
	330	Complex	0.029	0.019	0.017
		Simple	0.023	0.017	0.016

Note. The minimum value in each row is shown in bold. G-DINA = generalized deterministic inputs, noisy, “and” gate; 2LR = two-step likelihood ratio.

statistic, thus increasing the number of times that an incorrect reduced CDM was retained. As the calibration sample size increased, the selection rates improved. In the $N = 2,000$ conditions, the true reduced model was always selected under 2LR-Holm. The other two factors affected more the percentage of times the G-DINA model was retained, but these effects depended on the application of the correction for multiple comparisons. When the correction was made (i.e., 2LR-Holm), the G-DINA model was hardly ever retained, and these factors had only a small effect on the performance of the statistic. In contrast, for the 2LR-None, the most unfavorable conditions (i.e., a small item bank length and a complex Q-matrix structure) resulted in a higher proportion of G-DINA selections to the detriment of correct CDM selections. The good performance of the 2LR test allowed dramatically reducing the number of parameters to be estimated. For example, in the $J = 330$ and simple Q-matrix condition, the GDINA model estimated 1,620 parameters, whereas the combination of models selected by the 2LR test estimated an average of 770.6 to 779.7 parameters for the different levels of calibration sample size. Table 2 includes the RMSE results that represent the impact on the accuracy of the item parameter estimates (Equation 5). RMSE values were always lower for the combination of models selected by the 2LR test results compared to the G-DINA model. As shown in Table 1, including the Holm correction improved the performance of the 2LR test. Accordingly, RMSE values were smaller when models were selected including the Holm correction. Differences among the models became smaller as the sample size and the item bank length increased. In addition, the complex Q-matrix condition is a more challenging situation because it includes more parameters. This resulted in larger RMSE values. Item parameter recovery was expected to affect the performance of the adaptive algorithms based on those estimates as we explore in the next section.

Validation Sample Results: Pattern Recovery

Pattern recovery results are shown in Figures 2 and 3 for the 165 and 330 item bank length conditions, respectively. For comparison purposes, the upper limit of the pattern recovery is represented in black. In the following, we describe the most notable findings.

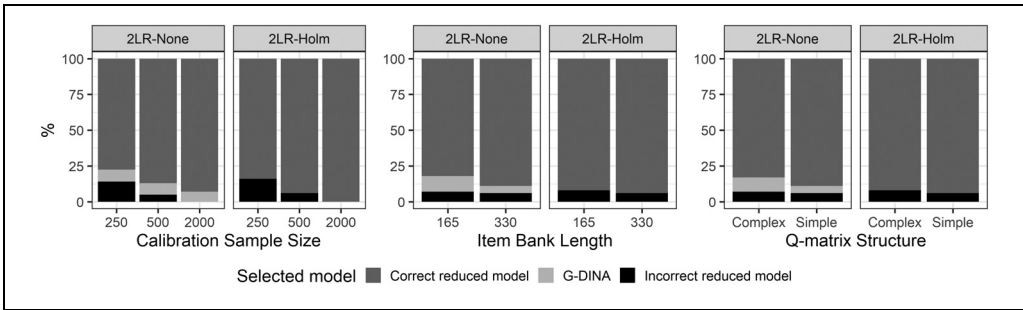


Figure 1. Representation of the 2LR test results by factor.

Note. 2LR = two-step likelihood ratio; G-DINA = generalized deterministic inputs, noisy, "and" gate.

General versus reduced CDMs. The true underlying model for the item bank was a combination of DINA, DINO, and *A*-CDM items. Thus, as expected, estimating the same reduced model (i.e., DINA, DINO, or *A*-CDM) for all the items in the item banks resulted in a poorer performance of the CD-CAT compared to that of the CD-CAT based on the G-DINA model that subsumes all of them. Among the reduced models, CD-CATs based on the DINA and DINO models performed similarly, and CD-CATs based on the *A*-CDM performed considerably worse in all conditions. As indicated in the previous section, the 2LR test generally flagged the most appropriate model for each item. Consequently, CD-CAT based on that combination of models usually had a very good overall performance. Indeed, the performance of this combination of models was always equal or better compared to that of the G-DINA model. For a 30-item CD-CAT, the average improvement in pattern recovery that was obtained by the 2LR test along with the Holm correction was 0.044, and ranged from 0.001 to 0.234.

Multiple comparison correction. Including the Holm correction always led to a better performance of the CD-CAT. This was related to the results described in the model selection section. Differences were more notable when the calibration sample and the item bank length were small, and the Q-matrix was complex.

Calibration sample size. The sampling estimating error was smaller when the sample size was large (i.e., $N = 2,000$), and then the results for pattern recovery for the more general model (i.e., G-DINA) were close to the upper limit. The same can be said for the combination of models selected by the 2LR test results, given that this statistic performed very well under this condition. In contrast, the G-DINA model was not accurately estimated under small sample conditions (i.e., $N = 500, 250$), and that's why the CD-CAT based on the G-DINA model parameters performed poorly when the calibration sample size became smaller. The 2LR test results performed generally close to the upper limit, given that reduced CDMs were easier to estimate under small calibration sample size conditions.

Q-matrix complexity. It was always harder to recover the attribute vector when the Q-matrix was complex. This decrement in accuracy was more pronounced for the G-DINA model as the number of item parameters to be estimated was higher. For example, in the complex Q-matrix for an item measuring four attributes $2^4 = 16$ parameters were estimated under the G-DINA model. If the DINA model fitted that particular item according to the 2LR test, only two parameters were estimated. In this line, even when the calibration sample size was 250, the CD-CAT based on the 2LR test selection was still relatively close to the upper limit. In the $J = 330$ condition, for a 30-item CD-CAT, the average pattern recovery was 0.86, 0.81, and 0.64 for CD-CATs based on

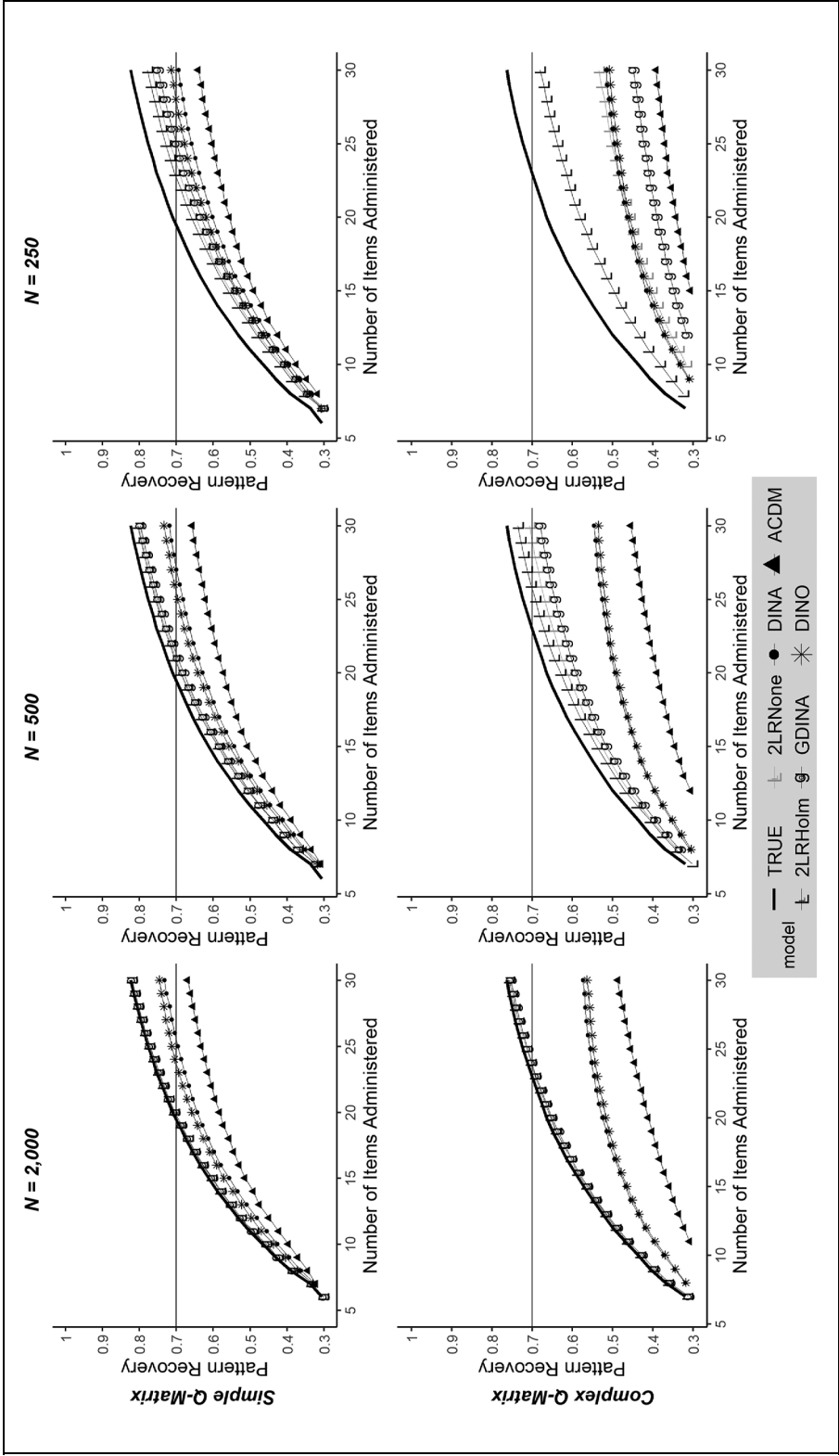


Figure 2. Pattern recovery according to fitted model and number of items administered. Note. Item bank length is 165 items. A horizontal line is included at *Pattern recovery* = 0.70 for interpretation purposes. The pattern recovery was always equal or higher for 2LR-Holm than for 2LR-None. DINA = deterministic inputs, noisy, “and” gate; DINO = deterministic inputs, noisy, “or” gate; A-CDM = additive cognitive diagnosis model; 2LR = two-step likelihood ratio; G-DINA = generalized DINA.

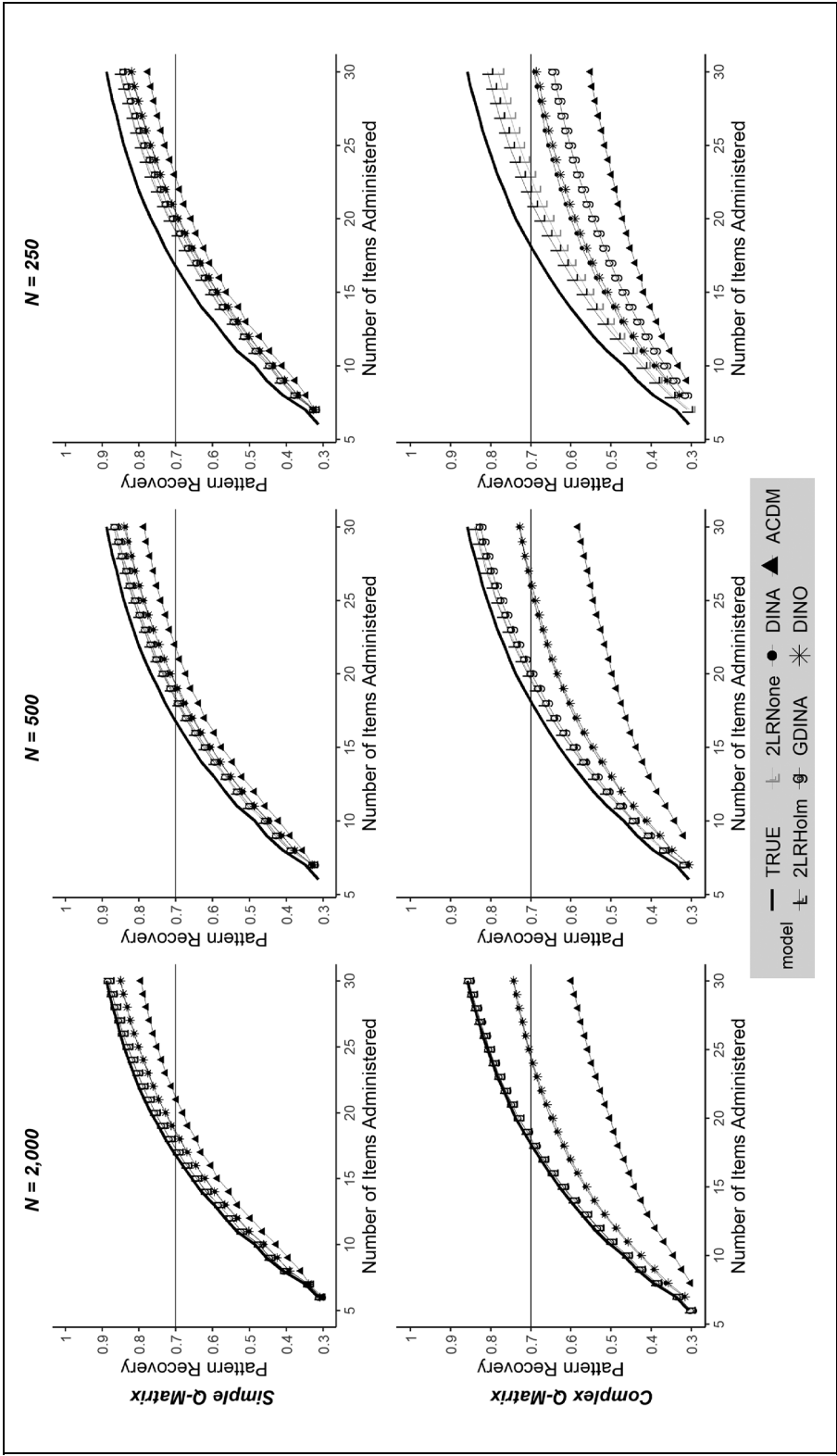


Figure 3. Pattern recovery according to fitted model and number of items administered.

Note. Item bank length is 330 items. A horizontal line is included at *Pattern recovery* = 0.70 for interpretation purposes. The pattern recovery was always equal or higher for 2LR-Holm than for 2LR-None. DINA = deterministic inputs, noisy, “and” gate; DINO = deterministic inputs, noisy, “or” gate; A-CDM = additive cognitive diagnosis model; 2LR = two-step likelihood ratio; G-DINA = generalized DINA.

Table 3. Average Item Usage Results for the 330-Item Banks in the Most and Least Ideal Conditions.

Most ideal condition: $Q\text{-str}$ = simple structure and $N = 2,000$

Model	One- attribute (#90)	Two- attribute (#120)	Three- attribute (#120)	DINA- items (#80)	DINO- items (#80)	A-CDM- items (#80)
TRUE	0.40	0.36	0.24	0.30	0.30	0.00
2LR-Holm	0.39	0.37	0.25	0.31	0.30	0.00
2LR-None	0.39	0.37	0.25	0.31	0.30	0.00
G-DINA	0.38	0.36	0.25	0.31	0.30	0.00
DINA	0.62	0.24	0.14	0.37	0.00	0.01
DINO	0.62	0.23	0.14	0.00	0.37	0.01
A-CDM	0.98	0.02	0.00	0.00	0.00	0.02

Least ideal condition: $Q\text{-str}$ = complex structure and $N = 250$

Model	One- attribute (#30)	Two- attribute (#120)	Three- attribute (#120)	Four- attribute (#60)	DINA- items (#100)	DINO- items (#100)	A-CDM- items (#100)
TRUE	0.20	0.46	0.28	0.07	0.41	0.39	0.00
2LR-Holm	0.20	0.46	0.27	0.07	0.40	0.39	0.02
2LR-None	0.19	0.45	0.27	0.09	0.39	0.39	0.03
G-DINA	0.17	0.41	0.27	0.15	0.39	0.37	0.06
DINA	0.40	0.40	0.16	0.04	0.50	0.01	0.09
DINO	0.40	0.38	0.18	0.05	0.01	0.51	0.09
A-CDM	0.54	0.40	0.05	0.01	0.11	0.09	0.26

Note. Maximum values within each item característica (i.e., complejidad y modelo) are shown in bold (± 0.03 differences are not considered). The number of items for each item type category (J_c) is indicated in the table by # J_c . DINA = deterministic inputs, noisy, “and” gate; DINO = deterministic inputs, noisy “or” gate; A-CDM = additive cognitive diagnosis model; 2LR = two-step likelihood ratio; G-DINA = generalized DINA.

the true item parameters, 2LR test along with the Holm correction selection and G-DINA estimates, respectively.

Item bank length. Increasing the item bank length always led to a better performance of the CD-CAT. Item parameters were generated using a uniform distribution, then augmenting the bank length increases the number of high-quality items, for which 2LR test model selection rates are higher. Results based on the 2LR-Holm test selection of models always achieved values greater than 0.70, even in the more problematic condition.

Validation Sample Results: Item Usage

Average item usage results across the 10 item banks are shown in Table 3. Only the large item bank length condition ($J = 330$) is considered to prevent the different item types in the item bank from being exhausted by the selection algorithm. In addition, due to space limits, only the most and least ideal data conditions are presented (i.e., simple vs complex Q -matrix, large vs small calibration sample size). The most notable results are listed in the following text: (a) simpler items were typically preferred when using GDI. It should be noted that in the complex Q -matrix condition there were only 30 one-attribute items. Probably, highly discriminating one-attribute items were exhausted. The use of one-attribute items could have been greater if a wider range of one-attribute items were available; (b) the patterns of item usage for both

Table 4. True and Estimated Item Parameters for Two Different Item Types.

One-attribute item: Item 2 with q -vector = {01000}					
Models	GDI	$P(0)$	$P(1)$		
TRUE	0.065	0.246	0.754		
2LR-Holm	0.072	0.210	0.747		
G-DINA	0.072	0.210	0.747		
DINA	0.071	0.211	0.744		
DINO	0.072	0.210	0.748		
A-CDM	0.072	0.209	0.747		
DINA two-attribute item: Item 50 with q -vector = {01100}					
Models	GDI	$P(00)$	$P(10)$	$P(01)$	$P(11)$
TRUE	0.068	0.202	0.202	0.202	0.798
2LR-Holm	0.068	0.203	0.203	0.203	0.801
G-DINA	0.068	0.196	0.199	0.214	0.801
DINA	0.068	0.204	0.204	0.204	0.801
DINO	0.009	0.195	0.411	0.411	0.411
A-CDM	0.034	0.119	0.374	0.386	0.640

Note. GDI/probability differences with respect to the true values greater than 0.03/0.10 are shown in bold. 2LR-Holm and 2LR-None provided identical results. GDI = G-DINA model discrimination index; 2LR = two-step likelihood ratio; DINA = deterministic inputs, noisy, "and" gate; G-DINA = generalized DINA; DINO = deterministic inputs, noisy "or" gate; A-CDM = additive cognitive diagnosis model.

$\alpha = 0.05$ and Holm implementations of the 2LR test were quite similar, and were the ones closest to the pattern corresponding to the true estimates. These patterns were also similar to those of the CAT based on the G-DINA model, except for a smaller use of four-attribute items in the complex Q-matrix condition; (c) when the data were calibrated using a single reduced CDM, one-attribute items were generally preferred. It should be noted that all the CDMs are equivalent when the number of attributes being measured by the item is one. In addition, items following a different model were seldom used, and items following that specific reduced models were mostly used. The former was more pronounced when the Q-matrix was simple, whereas the latter was more pronounced when the Q-matrix was complex; (d) items following the A-CDM model were seldom administered. This was most noticeable in the simple Q-matrix conditions, where, even when the item bank was calibrated using the A-CDM model, A-CDM items were rarely administered. In this situation, the algorithm generally administered one-attribute items.

All above has to do with the fact that item parameters were not properly estimated when the calibrated reduced model differed from the true generating model. This is illustrated in Table 4 where the estimated parameters for two items in the $N = 2,000$ and simple Q-matrix condition are presented. Item 2 was a one-attribute item, and thus all the CDMs provided essentially the same item parameters. This might explain why one-attribute items were usually used under any condition. Item 50 was a two-attribute item following the DINA model. As can be seen from the table, the estimated GDI for DINO and A-CDM was quite low, whereas the G-DINA model, the model derived from the 2LR test, and the G-DINA model provided similar results close to the GDI that was specified in the data generation.

Discussion

In current empirical studies, a single reduced CDM is applied to all items in the item bank (e.g., H. Y. Liu et al., 2013). Generally, this might not be a suitable approach, given that

reduced models make strong assumptions about the data, so they might not be appropriate for all items. Accounting for the heterogeneity of CDMs even within the same test found in some empirical studies (de la Torre et al., 2018; Ravand, 2016), the use of general CDMs emerged as a good alternative (Sorrel, Yigit, & Kaplan, 2017). A recent literature review of CDM fixed-form applications found that only 31% of the studies used general CDMs (Sessoms & Henson, 2018), so this alternative is not the most common. It has a limitation that the estimation of general CDMs is much more challenging, typically requiring a larger sample size to be estimated accurately. Considering this, the present study explored whether the classification accuracy can be improved using comparison indices to select the most appropriate model for each item. The results indicated that implementing item-level model comparison indices such as 2LR test (Sorrel, de la Torre, et al., 2017) improved the accuracy of the CD-CAT under all the simulated conditions. Accordingly, the same accuracy can be obtained with fewer items administered. Time saving has been traditionally considered one of the advantages of CATs. This study shows that a key element to maximize this advantage is model selection. This time saving might be of major importance, for example, in classroom settings because it would allow teachers designing classroom specific activities to optimize student learning. A test of this type can be applied in a weekly basis to develop a learning profile for the students in the class and adapt instruction accordingly (Wu, 2018). The lack of appropriate software programs for conducting adaptive applications based on CDM might have hampered the development of more empirical applications. To facilitate these applications and encourage new ones, the code was turned into an R package named *cdcatR* (Sorrel, Nájera, & Abad, 2020).

Regarding the manipulated factors, we found that the accuracy improvement by the use of the 2LR test can be expected to be larger when the Q-matrix structure is complex (i.e., large proportion of items measuring more than one attribute) and the calibration sample size is small. Otherwise, if the same reduced CDM (e.g., DINA) is applied to all items in a situation in which items follow several different CDMs, the resulting accuracy will be generally much lower. This might be ameliorated in a certain way if the sample size is large and the Q-matrix has a simple structure, as in H. Y. Liu et al. (2013)'s study, but still a CD-CAT based on a general model or a combination of appropriate models would provide better accuracy results. Furthermore, even if a similar accuracy is obtained with the application of a single reduced model, there will be a poor use of the item bank. Specifically, items following a different reduced CDM will not be selected by the adaptive procedure. This is due to a severe underestimation of the model discrimination when an incorrect reduced CDM is specified for an item. Results of this study indicate that this inefficient use of the item bank can be tackled through the use of model selection indices. On the other hand, procedures based on a general model (e.g., G-DINA) will lead to optimal results provided the general model is accurately estimated. This will generally be the case when the sample size is large and the number of parameters to be estimated is small (e.g., Sorrel, Yigit, & Kaplan, 2017). Otherwise the classification accuracy can be compromised. In any case, the main finding of this study is that we can improve classification accuracy and make a better use of the item bank using item-level model fit indices to select the most appropriate CDMs for each item. Importantly, it will not have any negative impact. This study considers DINA, DINO, and *A*-CDM models, but other different constrained versions of the G-DINA model can be easily included in the set of compared models. Given the large number of comparisons, we encourage researches and practitioners to use a procedure to control the Type I error rate such as the Holm correction. It is worth noting that these methodologies are indeed very easy to implement. It only takes a few seconds to conduct the model selection analysis.

Findings from this study can serve future research in several ways. First, both calibration sample size and Q-matrix complexity are factors that greatly influence the adaptive algorithm. Results regarding the calibration sample size are consistent with those in the recent study by

Huang (2018). Simulation studies in this context should consider these two factors in order to ensure findings that are broadly generalizable. Second, we found that simpler items were typically preferred by GDI. One of the possible reasons is that all models are equivalent when the item measures only one attribute, whereas the models are more and more different as the item complexity increases. If the appropriate reduced model is not correctly specified, the item discrimination would be severely underestimated. On another note, items following *A*-CDM were not generally administered. These results are in line with previous research using GDI (Kaplan et al., 2015; Yigit et al., 2019). Third, most of the item selection methods rely on parametric models. There is a recent nonparametric item selection (NPS; Chang et al., 2019) method that has demonstrated promising results. It uses the nonparametric classification (NPC; Chiu & Douglas, 2013) method to estimate the attribute profiles. Nonparametric models have been shown to outperform the parametric ones when the sample size is especially low (i.e., 30–100). The main drawback of the NPC is that it needs to specify a condensation rule that determines the ideal response pattern for each item and latent class. According to our results, it is expected that assuming an incorrect condensed rule will result in a loss in classification accuracy. Finally, it was not uncommon to see that multiple reduced CDMs obtained a fit similar to that of the general model. Like other previous studies (e.g., Ma et al., 2016), we retained the reduced CDM with the highest *p*-value. Despite the good results achieved, future studies might address whether descriptive measures of fit such as AIC or BIC can improve the model selection rates.

A few limitations of this study are worth mentioning. First, the Q-matrices involved in this study were assumed to be known. This represents a plausible scenario in the context of CD-CAT where a lot of resources are invested and the Q-matrix construction is probably guided by a strong theory. This is the case, for example, of H. Y. Liu et al. (2013) and Sorrel, Yigit, & Kaplan (2017) studies. Over the recent years, a number of methods for Q-matrix empirical validation have been proposed. These include the GDI method (de la Torre & Chiu, 2016; Nájera et al., 2019). This method requires an initial Q-matrix that will contain, presumably, some misspecifications. The method is aimed to detect and correct those misspecifications. Second, to keep the scope of this study manageable, a few simplifications about factors affecting the CD-CAT performance were made. These included fixing the number of attributes, using a single method in estimating the attribute vectors, and focusing on the unconstrained CD-CAT where neither exposure control nor content balancing was considered. The reason for that is that usually CDM applications are relatively low-stakes and, accordingly, test security is not a big concern. But if the test is high-stakes (e.g., personnel selection), exposure control becomes necessary. This can be done, for example, using the Symptom–Hetter algorithm or dynamic binary searching procedures (Zheng & Wang, 2017). These two methods are viable solutions to reduce item exposure without seriously affecting accuracy. In the data generation process, there was no reason to consider any particular attribute joint distribution. Therefore, latent classes were sampled from a uniform distribution. This favors the item bank calibrations. Different studies might explore the effect of the attribute joint distribution assuming a particular prior. Responses to all items were available for the item bank calibration. Matrix sampling designs are more common in practical settings to minimize testing time and fatigue. In those situations, when the data set is incomplete, the item parameter and pattern recovery will be worse. Nonetheless, it could be hypothesized that the calibration based on the general model would be more negatively affected because less information is available. The results of a preliminary simulation study conducted in response to the comments of one of the reviewers showed that the benefit of using the 2LR rather than the G-DINA model was even higher under incomplete designs. Matrix sampling designs are unexplored in the context of CDM. Future research is required to address this issue. Finally, this study focuses on what has come to be called low item quality in previous simulations studies (e.g., Ma et al., 2016; Sorrel, Abad, et al., 2017)

because a higher accuracy improvement was expected. This is nonetheless a realistic scenario considering the range of item discrimination values reported in empirical applications outside educational measurement (de la Torre et al., 2018; H. Y. Liu et al, 2013; Sorrel et al., 2016; Templin & Henson, 2006).

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 grants PSI2013-44300-P and PSI2017-85022-P (Ministerio de Economía y Competitividad and European Social Fund) and the UAM-IIC Chair «Psychometric Models and Applications».

ORCID iDs

Miguel A. Sorrel  <https://orcid.org/0000-0002-5234-5217>

Pablo Nájera  <https://orcid.org/0000-0001-7435-2744>

Notes

1. Results in terms of selection rate, pattern recovery, and item usage based on the Wald test can be expected to be very similar (Sorrel, Abad, et al., 2017; Sorrel, de la Torre, et al., 2017). However, the two-step LR test was found to be faster than the Wald test (GDINA v. 2.7.8). This is most likely due to the computation of standard errors in the Wald test. It is to be expected that with short tests the differences in computation time will be attenuated. Y. Liu et al. (2019) can be consulted for a discussion on how to estimate the standard errors that are required in the Wald test.

References

- Akbay, L., & Kaplan, M. (2017). Transition to multidimensional and cognitive diagnosis adaptive testing: An overview of CAT. *The Online Journal of New Horizons in Education-January*, 7(1), 206–214.
- Chang, Y. P., Chiu, C. Y., & Tsai, R. C. (2019). Nonparametric CAT for CD in educational settings with small samples. *Applied Psychological Measurement*, 43, 543–561.
- Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, 74, 619–632.
- Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement*, 70, 902–913.
- Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, 30, 225–250.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179–199.
- de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, 81, 253–273.
- 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.
- de la Torre, J., van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, 51, 281–296.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, 26, 301–321.

- 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.
- Huang, H. Y. (2018). Effects of item calibration errors on computerized adaptive testing under cognitive diagnosis models. *Journal of Classification*, 35, 437–465.
- Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research & Evaluation*, 15, Article 3.
- Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, 39, 167–188.
- Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment*, 18, 1–25.
- Liu, H. Y., You, X. F., Wang, W. Y., Ding, S. L., & Chang, H.-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification*, 30, 152–172.
- Liu, Y., Andersson, B., Xin, T., Zhang, H., & Wang, L. (2019). Improved Wald statistics for item-level model comparison in diagnostic classification models. *Applied Psychological Measurement*, 43, 402–414.
- Ma, W., & de la Torre, J. (2020). *GDINA: The generalized DINA model framework* (R package version 2.7.8). <https://CRAN.R-project.org/package=GDINA>
- Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, 40, 200–217.
- Nájera, P., Sorrel, M. A., & Abad, F. J. (2019). Reconsidering cutoff points in the general method of empirical Q-matrix validation. *Educational and Psychological Measurement*, 79, 727–753.
- Olea, J., Barrada, J. R., Abad, F. J., Ponsoda, V., & Cuevas, L. (2012). Computerized adaptive testing: The capitalization on chance problem. *The Spanish Journal of Psychology*, 15, 424–441.
- Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, 34, 782–799.
- 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 presentation]. Meeting of the National Council on Measurement in Education, Vancouver, Canada.
- Rupp, A. A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, 6, 219–262.
- Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, 16, 1–17.
- Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*, 41, 614–631.
- Sorrel, M. A., Barrada, J. R., de la Torre, J., & Abad, F. J. (2020). Adapting cognitive diagnosis computerized adaptive testing item selection rules to traditional item response theory. *PLOS ONE*, 15(1), Article e0227196. <https://doi.org/10.1371/journal.pone.0227196>
- Sorrel, M. A., de la Torre, J., Abad, F. J., & Olea, J. (2017). Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models. *Methodology*, 13, 39–47.
- Sorrel, M. A., Nájera, P., & Abad, F. J. (2020). *cdcatR: Cognitive diagnostic computerized adaptive testing in R* (R package version 1.0.1). <https://CRAN.R-project.org/package=cdcatR>
- Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, 19, 506–532.
- Sorrel, M. A., Yigit, H. D., & Kaplan, K. (2017, April). *CD-CAT implementation of the proportional reasoning assessment* [Paper presentation]. Annual meeting of the National Council of Education, San Antonio, TX.
- Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287–305.
- Tjoe, H., & de la Torre, J. (2014). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal*, 26, 237–255.

- von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, 2005(2), i–35.
- Wang, W., Song, L., Wang, T., Gao, P., & Xiong, J. (2020). A note on the relationship of the Shannon entropy procedure and the Jensen–Shannon divergence in cognitive diagnostic computerized adaptive testing. *SAGE Open*, 10(1). <https://doi.org/10.1177/2158244019899046>
- Wu, H.-M. (2018). Online individualized tutor for improving mathematics learning: A cognitive diagnosis model approach. *Educational Psychology*, 39, 1218–1232.
- Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, 45, 675–707.
- Yigit, H. D., Sorrel, M. A., & de la Torre, J. (2019). Computerized adaptive testing for cognitively based multiple-choice data. *Applied Psychological Measurement*, 43, 388–401.
- Zheng, C., & Wang, C. (2017). Application of binary searching for item exposure control in cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, 41, 561–576.