Inferential Item Fit Evaluation in Cognitive Diagnosis Modeling

Miguel A. Sorrel, Francisco J. Abad, & Julio Olea

Universidad Autónoma de Madrid

Jimmy de la Torre

The University of Hong Kong

Juan Ramón Barrada

Universidad de Zaragoza

Author Note

Miguel A. Sorrel, Francisco J. Abad, Julio Olea, Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Madrid, Spain; Jimmy de la Torre, Faculty of Education, The University of Hong Kong, Hong Kong, Hong Kong; Juan Ramón Barrada, Department of Psychology and Sociology, Universidad de Zaragonza, Teruel, España.

This research was supported by Grant PSI2013-44300-P (Ministerio de Economia y Competitividad and European Social Fund).

Correspondence concerning this article should be addressed to Miguel A. Sorrel, Department of Social Psychology and Methodology, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, Madrid 28049, Spain, e-mail: miguel.sorrel@uam.es

Abstract

Research related to the fit evaluation at the item-level involving cognitive diagnosis models (CDMs) has been scarce. According to the parsimony principle, balancing goodness-of-fit against model complexity is necessary. General CDMs require a larger sample size to be estimated reliably, and can lead to worse attribute classification accuracy than the appropriate reduced models when the sample size is small and the item quality is poor, which is typically the case in many empirical applications. The main purpose of this study is to systematically examine the statistical properties of four inferential item fit statistics: $S - X^2$, the likelihood ratio test, the Wald test, and the Lagrange multiplier test. To evaluate the performance of the statistics, a comprehensive set of factors, namely, sample size, correlational structure, test length, item quality, and generating model, is systematically manipulated using Monte Carlo methods. Results show that the $S - X^2$ statistic has unacceptable power. Type I error and power comparisons favours LR and W tests over the LM test. However, all the statistics are highly affected by the item quality. With a few exceptions, their performance is only acceptable when the item quality is high. In some cases, this effect can be ameliorated by an increase in sample size and test length. This implies that using the above statistics to assess item fit in practical settings when the item quality is low remains a challenge.

Keywords: cognitive diagnosis models, item fit statistics, absolute fit, relative fit, Type I error, power

Inferential Item Fit Evaluation in Cognitive Diagnosis Modeling

Cognitive diagnosis models (CDMs) have been actively researched in the recent measurement literature. CDMs are multidimensional and confirmatory models specifically developed to identify the presence or absence of multiple attributes involved in the assessment items (for an overview of these models see, e.g., DiBello, Roussos, & Stout, 2007; Rupp & Templin, 2008). Although originally developed in the field of education, these models have been employed in measuring other types of constructs, such as psychological disorders (e.g., de la Torre, van der Ark, & Rossi, 2015; Templin & Henson, 2006) and situation-based competencies (Sorrel et al., 2016).

There are currently no studies comparing item characteristics (e.g., discrimination, difficulty) as a function of the kind of the constructs being assessed. However, some data suggest that important differences can be found. Specifically, notable differences are found for item discrimination, which is one of the most common index used to assess item quality. Item discrimination relates to how well an item can accurately distinguish between respondents who differ on the constructs being measured. Although it does not account for the attribute complexity of the items, a simple measure of discrimination is defined as the difference between the probabilities of correct response for those respondents mastering all and none of the required attributes. This index it is bounded by 0 and 1. In empirical applications, such as the fraction subtraction data described and used by Tatsuoka (1990) and by de la Torre (2011), one of the most widely employed datasets in CDM in the educational context, the mean discrimination power of the items was .80. In contrast, when CDMs have been applied in applications outside educational measurement the resulting discrimination estimates were found to be in the .40 range (de la Torre et al., 2015; Liu, You, Wang, Ding, & Chang, 2013; Sorrel et al., 2016; Templin & Henson, 2006). In these empirical applications, researchers typically used a sample size that varies approximately from 500 (e.g., de la Torre, 2011; Templin and Henson, 2006) to 1,000 (de la Torre et al., 2015), and an average number of items equal to 30, 12 being the minimum (de la Torre, 2011). Different CDMs were considered, including the *deterministic inputs, noisy "and" gate* (DINA; Haertel, 1989) model), the *deterministic inputs, noisy "or" gate* (DINO) model (Templin & Henson, 2006), the *additive* CDM (*A*-CDM; de la Torre, 2011), and the *generalized deterministic inputs, noisy "and" gate* (G-DINA; de la Torre, 2011) model.

Given the large number of different models, one of the critical concerns in CDM is selecting the most appropriate model from the available CDMs. Each CDM assumes a specified form of item response function (IRF). In the CDM context, the IRF denotes the probability that an item j is answered correctly as a function of the latent class. This study focused on methods assessing this assumption. Model fit evaluated at the test level simultaneously takes all the items into consideration. However, when there is model-data misfit at the test level, the misfit may be due to a (possibly small) subset of the items. Item-level model fit assessment allows to identify these misfitting items. The research focused on item fit is important because such analysis can provide guidelines to practitioners on how to refine a measurement instrument. This is a very important topic because current empirical applications reveal that no one single model can be used for all the test items (see, e.g., de la Torre et al., 2015; de la Torre & Lee, 2013; Ravand, 2015). Consequently, in this scenario, item fit statistics are a useful tool for selecting the most appropriate model for each item. The main purpose of this study was to systematically examine the Type I error and power of four item fit statistics, and provide information about the usefulness of these indexes across different plausible scenarios. Only goodness-of-fit measures with a significance test associated with them (i.e., inferential statistical evaluation) were considered in this article. The rest of the article is structured as follows. First is a brief introduction of the generalized DINA model framework. This is followed by a review of item fit evaluation in CDM, and for a presentation of the simulation study designed to evaluate the performance of the different item fit statistics. Finally, the results of the simulation study and the implications and future studies are discussed.

The Generalized DINA Model Framework

In many situations the primary objective of CDM is to classify examinees into 2^K latent classes for an assessment diagnosing K attributes. Each latent class is represented by an attribute vector denoted by $\boldsymbol{\alpha}_l = (\alpha_{l1}, \alpha_{l2}, \ldots, \alpha_{lK})$, where $l = 1,..., 2^K$. All CDMs can be expressed as $P(X_j = 1 | \boldsymbol{\alpha}_l) = P_j(\boldsymbol{\alpha}_l)$, the probability of success on item j conditional on the attribute vector l. For diagnostic purposes, the main CDM output of interest is the estimate of examinee i's $\boldsymbol{\alpha}_i = \{\alpha_{ik}\}$.

Several general models that encompasses reduced (i.e., specific) CDMs have been proposed, which include the above-mentioned G-DINA model, the general diagnostic model (GDM; von Davier, 2005), and the log-linear CDM (LCDM; Henson, Templin, & Willse, 2009). In this article, the G-DINA model, which is a generalization of the DINA model, is employed. The G-DINA model describes the probability of success on item j in terms of the sum of the effects of the attributes involved and their corresponding interactions. This model partitions the latent classes into $2^{K_j^*}$ latent groups, where K_j^* is the number of required attributes for item j. Each latent group represents one reduced attribute vector, $\boldsymbol{\alpha}_{lj}^*$, that has its own associated probability of success, written as

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

where δ_{j0} is the intercept for item *j*, δ_{jk} is the main effect due to α_k, $\delta_{jkk'}$ is the interaction effect due to α_k, and $\alpha_{k'}$ and $\delta_{j12...K_j^*}$ is the interaction effect due to $\alpha_1,...,\alpha_{K_j^*}$. Thus, without constraints on the parameter values, there are $2^{K_j^*}$ parameters to be estimated for item *j*.

The G-DINA model is a saturated model that subsumes several widely used reduced CDMs, including the DINA model, the DINO model, the *A*-CDM, the linear logistic model (LLM; Maris, 1999), and the reduced reparametrized unified model (R-RUM; Hartz, 2002). Although based on different link functions, *A*-CDM, LLM, and R-RUM are all additive models, where the incremental probability of success associated with one attribute is not affected by those of other attributes. Ma, Iancoangelo, and de la Torre (2016) found that, in some cases, one additive model can closely recreate the IRF of other additive models. Thus, in this work we only consider three of these reduced models corresponding to the three types of condensation rules: DINA model (i.e., conjunctive), DINO model (i.e., disjunctive), and the *A*-CDM (i.e., additive). If several attributes are required to correctly answer the items, the DINA model is deduced from the G-DINA model by setting to zero all terms except for δ_{j0} and $\delta_{j12...K_j^*}$ to zero. As such, the DINA model has two parameters per item. Likewise, the DINO model also has two parameters per item, and can be obtained from the G-DINA model by setting $\delta_{jk} = -\delta_{jkk'} = \cdots = (-1)^{K_j^*+1}\delta_{j12...K_j^*}$. When all the interaction terms are dropped, the G-DINA model under the identity link reduces to the *A*-CDM, which has K_j^*+1 parameters per item. Each of these models assumes a different cognitive process in solving a problem (for a detailed description, see de la Torre, 2011).

Item Fit Evaluation

The process of model selection involves checking the model-data fit, which can be examined at test, item, or person level. Extensive studies have been conducted to evaluate the performance of various fit statistics at the test level (e.g., Chen, de la Torre, & Zhang, 2013; Liu, Tian, & Xin, 2016), and at the person level (e.g., Liu, Douglas, & Henson, 2009; Cui & Leighton, 2009). At the item level, some item fit statistics have also been recently proposed to evaluate absolute fit (i.e. the discrepancy between a statistical model and the data) and relative fit (i.e. the discrepancy between two statistical models). The parsimony principle dictates that from a group of models that fit equally well, the simplest model should be chosen. The lack of parsimony, or overfitting, may result in a poor generalization performance of the results to new data because some residual variation of the calibration data is captured by the model. With this in mind, general CDMs should not be always the preferred model. In addition, as pointed out by de la Torre and Lee (2013), there are several reasons that make reduced models preferable to general models. First, general CDMs are more complex, thus requiring a larger sample size to be estimated reliably. Second, reduced models have parameters with a more straightforward interpretation. Third, appropriate reduced models lead to better attribute classification accuracy than the saturated model, particularly when the sample size is small and the item quality is poor (Rojas, de la Torre, & Olea, 2012). In this line, Ma et al. (2016) found that a combination of different appropriate reduced models determined by the W test always produced a more accurate classification accuracy than the unrestricted model (i.e., the G-DINA model). In the following, we will describe some of the statistics that may be computed in this context.

Absolute Fit

Absolute item fit is typically assessed by comparing the item performance on various groups to the performance levels predicted by the fitted model. A χ^2-like statistic is used to make this comparison. Different statistics have emanated from traditional item response theory (IRT), and the main difference among them is how the groups are formed. There are two main approaches. In the first one, respondents are grouped based on their latent trait estimates and observed frequencies of correct/incorrect responses for these groups are obtained. Yen's (1981) Q_1 statistic is computed using this approach and has been adapted to CDM (Sinharay & Almond, 2007; Wang, Shu, Shang, & Xu, 2015). Its performance has been compared to that of the posterior predictive model checking method (Levy, Mislevy, & Sinharay, 2009). Q_1 type I error was generally well kept below .05 and was preferred to the posterior predictive model checking method. The main problem with this approach is that observed frequencies are not truly observed because they cannot be obtained without first fitting a certain model. This will lead to a model-dependent statistic that makes it difficult to determine the degrees of freedom (Orlando & Thissen, 2000; Stone & Zhang, 2003). In the second approach, the statistic is formulated based on the observed and expected frequencies of correct/incorrect responses for each summed score (Orlando & Thissen, 2000). The main advantage of this approach is that the observed frequencies are solely a function of observed data. Thus, the expected frequencies can be compared directly to observed frequencies in the data. A χ^2-like statistic, referred to as $S-X^2$ (Orlando and Thissen, 2000), is then computed as

$$S-X^2{}_j = \sum_{s=1}^{J-1} N_s \frac{(O_{js}-E_{js})^2}{E_{js}(1-E_{js})} \sim \chi^2(J\text{-}1\text{-}m), \tag{2}$$

where s is the score group, J is the number of items, N_s is the number of examinees in group s, and O_{js} and E_{js} are, the observed and predicted proportions of correct responses for item j for group s, respectively. The model-predicted probability of correctly responding item j for examinees with sum score s is defined as

$$P(x_{ij}=1|S_i=s) = \frac{\sum_{l=1}^{2^K} P\left(x_{ij}=1|\alpha_l\right) P\left(S_i^j=s-1|\alpha_l\right) P\left(\boldsymbol{\alpha}\right)}{\sum_{l=1}^{2^K} P\left(S_i=s|\alpha_l\right) P\left(\boldsymbol{\alpha}\right)}, \tag{3}$$

where $P\left(S_i^j = s-1 \middle| \alpha_l\right)$ is the probability of obtaining the sum score $s-1$ in the test composed of all the items except item *j*, and $P\left(\boldsymbol{\alpha}\right)$ defines the probability for each of the latent. Model-predicted joint likelihood distributions for each sum score are computed the recursive algorithm developed by Lord and Wingersky (1984) and detailed in Orlando and Thissen (2000). The statistic is assumed to be asymptotically χ^2 distributed with $J - 1 - m$ degrees of freedom, where *m* is the number of item parameters.

Relative Fit

When comparing different nested models there are three common tests than can be used (Buse, 1982): likelihood ratio (LR) test, Wald (W), and Lagrange multiplier (LM) tests. In the CDM context, the null hypothesis (H_0) for these tests assumes that the reduced model (e.g., *A*-CDM) is the "true" model, whereas the alternative hypothesis (H_1) states that the general model (i.e., G-DINA) is the "true" model. As such, H_0 defines a restricted parameter space. For example, for an item *j* measuring two attributes in the *A*-CDM model, we restrict the interaction term to be equal to 0, whereas this parameter is freely estimated in the G-DINA model. It should be noted that the three procedures are asymptotically equivalent (Engle, 1983). In all the three cases, the statistic is assumed to be asymptotically χ^2 distributed with $2^{K_j^*} - p$ degrees of freedom, where *p* is the number of parameters of the reduced model.

Let $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ denote the maximum likelihood estimates of the item parameters under H_0 and H_1, respectively (i.e., restricted and unrestricted estimates of the population parameter). Although all three tests answer the same basic question, their approaches to answering the question differ slightly. For instance, the LR test requires estimating the models under H_0 and H_1; in contrast, the W test requires estimating only the model under H_1, whereas the LM test requires estimating only the model under H_0.

Before describing in greater detail these three statistical tests, it is necessary to mention a few points about the estimation procedure in CDM. The parameters of the G-DINA model can be estimated using the marginalized maximum likelihood estimation (MMLE) algorithm as described in de la Torre (2011). By taking the derivative of the log-marginalized likelihood of the response data, $l(\mathbf{X})$, with respect to the item parameters, $P_j(\boldsymbol{\alpha}_{lj}^*)$, we obtain the *estimating* function:

$$\frac{\partial l(\mathbf{X})}{\partial P_j(\boldsymbol{\alpha}_{lj}^*)} = \left[\frac{1}{P_j(\boldsymbol{\alpha}_{lj}^*)\left(1-P_j(\boldsymbol{\alpha}_{lj}^*)\right)}\right]\left[R_{\alpha_{lj}^*} - P_j(\boldsymbol{\alpha}_{lj}^*)I_{\boldsymbol{\alpha}_{lj}^*}\right], \quad (4)$$

where $I_{\boldsymbol{\alpha}_{lj}^*}$ is the number of respondents expected to be in the latent group $\boldsymbol{\alpha}_{lj}^*$, and $R_{\boldsymbol{\alpha}_{lj}^*}$ is the number of respondents in the latent group $\boldsymbol{\alpha}_{lj}^*$ expected to answer item *j* correctly. Thus, the MMLE estimate of $P_j(\boldsymbol{\alpha}_{lj}^*)$ is given by $\hat{P}_j(\boldsymbol{\alpha}_{lj}^*) = R_{\boldsymbol{\alpha}_{lj}^*} / I_{\boldsymbol{\alpha}_{lj}^*}$. Estimating functions are also known as *score* functions in the LM context. The second derivative of the log-marginalized likelihood with respect to $P_j(\boldsymbol{\alpha}_{lj}^*)$ and $P_j(\boldsymbol{\alpha}_{l'j}^*)$ can be shown to be (de la Torre, 2011)

$$-\sum_{i=1}^{I}\left\{p(\boldsymbol{\alpha}_{lj}^* \mid \mathbf{X}_i)\frac{X_{ij} - P_j(\boldsymbol{\alpha}_{lj}^*)}{P_j(\boldsymbol{\alpha}_{lj}^*)\left[1 - P_j(\boldsymbol{\alpha}_{lj}^*)\right]}\right\}\left\{p(\boldsymbol{\alpha}_{l'j}^* \mid \mathbf{X}_i)\frac{X_{ij} - P_j(\boldsymbol{\alpha}_{l'j}^*)}{P_j(\boldsymbol{\alpha}_{l'j}^*)\left[1 - P_j(\boldsymbol{\alpha}_{l'j}^*)\right]}\right\}, \quad (5)$$

where $p(\boldsymbol{\alpha}_{lj}^* \mid \mathbf{X}_i)$ represents the posterior probability that examinee *i* is in latent group $\boldsymbol{\alpha}_{lj}^*$. Using $\hat{P}_j(\boldsymbol{\alpha}_{lj}^*)$ and the observed **X** to evaluate Equation 4, we obtain the information matrix for the parameters of item *j*, $\mathbf{I}(\hat{\mathbf{P}}_j^*)$, and its inverse corresponds to the variance-covariance matrix, $Var(\hat{\mathbf{P}}_j^*)$, where $\hat{\mathbf{P}}_j^* = \{\hat{P}_j(\boldsymbol{\alpha}_{lj}^*)\}$ denotes the probability estimates.

Likelihood ratio test. As previously noted, the LR test requires the estimation of both unrestricted and restricted models. The likelihood function is defined as the probability of observing **X** given the hypothesis. It is defined as $L(\tilde{\boldsymbol{\theta}})$ for the null hypothesis and $L(\hat{\boldsymbol{\theta}})$ for the alternative hypothesis. The LR statistic is computed as twice the difference between the logs of the two likelihoods:

$$LR = 2\left[\log L(\hat{\boldsymbol{\theta}}) - \log L(\tilde{\boldsymbol{\theta}})\right] \sim \chi^2\left(2^{K_j^*} - p\right), \tag{6}$$

where $\log L(\boldsymbol{\theta}) = \log \prod_{i=1}^{I} \sum_{l=1}^{L} L(\mathbf{X}_i|\boldsymbol{\alpha}_l)\, p(\boldsymbol{\alpha}_l)$ and $L(\mathbf{X}_i|\boldsymbol{\alpha}_l) = \prod_{j=1}^{J} P(\boldsymbol{\alpha}_{lj})^{X_{ij}}[1 - P(\boldsymbol{\alpha}_{lj})^{1-X_{ij}}$. Having a test composed of *J* items, the application of the LR test at the item level implies that $J_{K_j^*>1}$ comparisons will be made, where $J_{K_j^*>1}$ is the number of items measuring at least $K = 2$ attributes. For each of the $J_{K_j^*>1}$ comparisons, a reduced model is fitted to a target item, whereas the general model is fitted to the rest of the items. This model is said to be a restricted model because it has less parameters than an unrestricted model where the G-DINA is fitted to all the items. The LR test can be conducted to determine if the unrestricted model fits the data significantly better than the restricted model comparing the likelihoods of both the unrestricted and restricted models (i.e., $L(\hat{\boldsymbol{\theta}})$ and $L(\tilde{\boldsymbol{\theta}})$, respectively). Note that the likelihoods here are computed at the test level.

Wald test. The W test takes into account the curvature of the log-likelihood function, which is denoted by $C(\hat{\boldsymbol{\theta}})$, and defined by the absolute value of $\partial^2 \log L / \partial \boldsymbol{\theta}^2$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. In CDM research, de la Torre (2011) originally proposed the use of the Wald test to compare general and specific models at the item level under the G-DINA framework. For item *j* and a reduced model with *p* parameters, this test requires setting up $\mathbf{R}_j$, a $(2^{K_j^*} - p) \times 2^{K_j^*}$ restriction matrix with specific constraints that make the saturated model to be equivalent to the reduced model of interest. The Wald statistic is computed as

$$W_j = \left[\mathbf{R}_j \times \hat{\mathbf{P}}_j^*\right]' \left[\mathbf{R}_j \times Var\left(\hat{\mathbf{P}}_j^*\right) \times \mathbf{R}_j'\right]^{-1} \left[\mathbf{R}_j \times \hat{\mathbf{P}}_j^*\right] \sim \chi^2\left(2^{K_j^*} - p\right), \tag{7}$$

where $\hat{\mathbf{P}}_j^*$ are the unrestricted estimates of the item parameters.

Lagrange multiplier test. The LM test is based on the slope of the log-marginalized likelihood $S(\boldsymbol{\theta}) = \partial \log L / \partial \boldsymbol{\theta}$, which is called the *score* function. By definition, $S(\boldsymbol{\theta})$ is equal to zero when evaluated at the unrestricted estimates of $\boldsymbol{\theta}$ (i.e., $\hat{\boldsymbol{\theta}}$), but not necessarily when evaluated at the restricted estimates (i.e., $\tilde{\boldsymbol{\theta}}$). The score function is weighted by the information matrix to derive the LM statistics. Following the parameter estimation under the G-DINA framework, the score function can be assumed to be as indicated by Equation 3. The LM statistic for item *j* is defined as

$$LM_j = S_j(\tilde{\mathbf{P}}_j^*)' Var(\tilde{\mathbf{P}}_j^*) S_j(\tilde{\mathbf{P}}_j^*) \sim \chi^2\left(2^{K_j^*} - p\right), \tag{8}$$

where $\tilde{\mathbf{P}}_j^*$ are the restricted estimates of the item parameters. It should be noted that all item parameters are estimated under the restricted model.

Before these statistics can be used with real data, we need to ensure that they have good statistical properties. This is even more crucial for $S - X^2$, LR, and LM tests because they have not been examined before in the CDM context. There have been, however, noteworthy studies on $S - X^2$ in the IRT framework by Orlando and Thissen (2000, 2003) and Kang and Chen (2008). Its Type I error was generally found to be close to the nominal level. The LM test has also been applied within the IRT framework. It has been shown to be a useful tool for evaluating the assumption of the form of the item characteristics curves in the two- and three- parameter logistic models (Glas, 1999; Glas & Suárez-Falcón, 2003). However, item quality was not manipulated in these previous studies and its effect is yet to be determined. This factor has been found to be very relevant in many different contexts using the relative item fit indices, as is the case of the evaluation of differential item functioning

(DIF). For example, prior research using the LR test in DIF have found that the statistical power of the LR test to detect DIF increases with increases in item discrimination (Wang & Yeh, 2003).

The W test is the only one that has been employed before in the CDM context for assessing fit at the item level. However, we found only two simulation studies examining their statistical properties. Although these works have contributed to our state of knowledge in this field, many questions related to the usefulness of these statistics with empirical data remained open. De la Torre and Lee (2013) studied the W test in terms of Type I error and power, and they found that it had a relative accurate Type I error and high power, particularly with large samples and items measuring a small number of attributes. In their case, the number of items was fixed to 30 and item quality was not manipulated. Items were set to have a mean discrimination power of approximately .60. Recently, Ma et al. (2016) extended the findings of de la Torre and Lee (2013) by including two additional reduced models (i.e., LLM and R-RUM). In their simulation design, they also considered two additional factors, item quality, and attribute distribution. They found that, although item quality strongly influenced the Type I error and power, the effect of the attribute distribution (i.e., uniform or high-order) was negligible. As a whole, although these studies have shed some light on the performance of the W test, the impact of other important factors or levels not explicitly considered in these studies remains unclear. This study aims to fill this gap, as well as examine the potential use of $S-X^2$, LR, and LM tests for item fit evaluation in the CDM context.

Method

A simulation study was conducted to investigate the performance of several item fit statistics. Five factors were varied and their levels were chosen to represent realistic scenarios detailed in the introduction. These factors are: (1) generating model (*MOD*; DINA model, *A*-

CDM, and DINO model); (2) test length (*J*; 12, 24, and 36 items); (3) sample size (*N*; 500 and 1,000 examinees); (4) item quality or discrimination, defined as the difference between the maximum and the minimum probabilities of correct response according to the attribute latent profile (*IQ*; .40, .60, and .80); and (5) correlational structure (*DIM*; uni- and bidimensional scenarios).

The following are details of the simulation study. The probabilities of success for individuals who mastered none (all) of the required attributes were fixed to .30 (.70), .20 (.80), and .10 (.90) for the low, medium, and high item quality conditions, respectively. For the *A*-CDM, an increment of $.40/K_j^*$, $.60/K_j^*$, and $.80/K_j^*$ was associated with each attribute mastery for the low, medium, and high item quality conditions, respectively. The number of attributes was fixed to $K = 4$. The correlational matrix of the attributes has an off-diagonal element of .5 in the unidimensional scenario, and 2×2 block diagonal submatrices with a correlation of .5 in the bidimensional scenario. The Q-matrices used in simulating the response data and fitting the models are given in the online annex 1. There were the same number of one-, two-, and three-attribute items.

The 3×3×2×3×2 ($MOD \times J \times N \times IQ \times DIM$) between-subjects design produces a total of 108 factor combinations. For each condition, 200 data sets were generated and DINA, *A*-CDM, DINO, and G-DINA models were fitted. Type I error was computed as the proportion of times that we reject H_0 when the fitted model is true. Power was computed as the proportion of times that a wrong reduced model is rejected. For example, in the case of the DINA model, power was computed as the proportion of times that we reject H_0 when the generating model is the *A*-CDM or the DINO model. Type I error and power were investigated using .05 as the significance level. With 200 replicates, the 95% confidence interval for the Type I error is given by $.05 \pm 1.96\sqrt{.05(1-.05)/200} = [.02,.08]$. For the purposes of this work, a power of at least .80 was considered adequate. The power analysis may not be interpretable when the Type I error for the statistics compared is very disparate. To make meaningful comparisons, it was necessary to approximate the distribution of the item-fit statistic under the null hypothesis. In doing so, we used the results from the simulation study. A nominal alpha (α_n) for which the actual alpha (α_a) was equal to .05 was found for all cases (i.e., simulation conditions of the design) where the Type I error was either deflated or inflated (i.e., $\alpha_a \notin [.02, .08]$). In these cases, this adjusted value was used as α_n producing a value for power which could then be compared with the other statistical tests.

As a mean to summarize and better understand the results of the simulation study, separate ANOVAs were performed for each of the item fit statistics. Dependent variables were the Type I error and power associated with each statistical test for all items with the five factors as between-subjects factors. Due to the large sample size, most effects were significant. For this reason, omega squared ($\hat{\omega}^2$), measure of effect size, was chosen to establish the impact of the independent variables. We considered the following guidelines for interpreting $\hat{\omega}^2$ (Kirk, 1996): Effect sizes in the intervals [.010, .059), [.059, .138), and [.138, ∞) were considered small, medium, and large effects, respectively. In addition, a cutoff of $\hat{\omega}^2 \geq 0.138$ was used to establish the most salient interactions. We checked that the estimates of observed power (i.e., post-hoc power) were greater than .80. The code used in this article was written in R. Some functions included in the CDM (Robitzsch et al., 2015) and GDINA (Ma & de la Torre, 2016) packages were employed. The R code can be requested by contacting the corresponding author.

Results

Due to space constraints, we only discuss effects sizes and report marginal means for the most relevant effects. Type I error and power of the item fit statistics for the three reduced models in their entirety are shown in the online annexes 2 and 3.

Type I Error

The effect size $\hat{\omega}^2$ values and marginal means associated with each main effect on the Type I error are provided in Table 1. $S-X^2$ is the only statistic with a Type I error that was usually close to the nominal level. The marginal means are always within the [.02, .08] interval, with the grand mean being .06. We only find a small effect of item quality ($\hat{\omega}^2$ = .01) and the generating model ($\hat{\omega}^2$ = .03): Type I error was slightly larger in the low and medium item quality conditions and for the *A*-CDM. None of the interactions had a salient effect.

The Type I error of the LR, W, and LM tests were very similar. Type I error was only acceptable for the high item quality conditions, which was the factor with the greatest effect ($\hat{\omega}^2$ = .33, .71, and .30 for LR, W and LM tests, respectively). When the item discrimination is low or medium, the Type I error was inflated. This makes it difficult to interpret the marginal means for all other factors, because conditions with low, medium, and high item discrimination are mixed. That was why the marginal means were generally much larger than the upper-limit of the confidence interval (i.e., .08). All things considered, the grand means of the three tests were inflated: .19, .29, and .14 for LR, W, and LM tests, respectively. Only one of the two-way interactions had a salient effect: Generating model × Item quality. As can be observed from the Figure 1, there were large differences between the marginal means for the different levels of generating model across the levels of item quality. The Type I error was closer to the nominal level when item quality got higher, with the exception of the DINO model, where Type I error was more inflated with medium quality items. Marginal means for the high quality conditions were within the confidence interval for all models in the case of LR and W test. When the generating model is *A*-CDM the LM test tended to be conservative (i.e., Type I error dropped close to 0).

Insert Table 1 here

None of the other interactions for the LR, W, and LM tests were relevant so the main effects could be interpreted. However, as noted above, Type I error was generally acceptable only in the high item quality condition. Sample size and test length affected the performance of the three statistics: Sample size had a small effect for the LR, W, and LM tests ($\hat{\omega}^2$ = .01, .03, and .01, respectively); whereas test length had a small effect on the Type I error of the LR and LM tests ($\hat{\omega}^2$ = .02 and .03, respectively), and a large effect in the case of W test ($\hat{\omega}^2$ = .17). The Type I error was closer to the nominal level as the sample size and the test length increased. As can be observed in the online annex 2, there were cases where Type I error was within the confidence interval when the test length and the sample size were large (i.e., J = 24 or 36 and N = 1,000). Finally, correlational structure had a small effect in the case of the LM test ($\hat{\omega}^2$ = .02). The Type I error for the LM test was inflated in the bidimensional conditions compared to the unidimensional conditions, although differences were small.

Insert Figure 1 here

Power

The $\hat{\omega}^2$ values and marginal means associated with each main effect on the power are provided in Table 2. For most of the conditions involving high quality items, it was not necessary to correct α_a. For example, we corrected α_a for the LR tests only in some of the conditions (i.e., J = 12 and N = 500). The pattern of effects of the manipulated factors on the power was very similar for all the tests. However, power of the LR and W tests was almost always better than those of the $S-X^2$ and LM tests - the grand means across models were .75, .78, .25, and .46 for LR, W, $S-X^2$, and LM tests, respectively. Again, item quality had the greatest effect with an average $\hat{\omega}^2$ = .74. Power was usually lower than .80 in the low item quality conditions for all the statistics. This factor was involved in all the salient high-order interactions: Sample size × Item quality (Figure 2), Test length × Item quality (Figure 3), Test length × Item quality × Correlational structure (Figure 4), and Test length × Item quality × Generating model (Figure 5). Here follows a description of each of these interactions.

Insert Table 2 here

As noted before, power increased as the item quality got better. This effect interacted with the sample size and test length (see Figures 2 and 3). In the case of the $S-X^2$ and LM tests, the improvement on the power associated with moving from low to medium quality items was similar for the different levels of sample size and test length, but this gain is generally much bigger when we move from medium to high quality items in the case of the N = 1,000, J = 24, and J = 36 conditions. The pattern of results for the LR test was similar to the one observed for the W test. Thus only the W test was depicted in Figure 3. Power in the medium quality items conditions was already close to 1.00 when N = 1,000 and J = 24 or 36. This is why there is a small room for improvement when we move to high quality item conditions because of this ceiling effect.

Insert Figures 2 and 3 here

In the case of the LM test, we found that the three-way Correlational structure × Test length × Item quality had a salient effect on the power for rejecting *A*-CDM when was false. As can be seen from Figure 4, only test length and item quality had a noteworthy effect on the LM power in the bidimensional scenario.

Insert Figure 4 here

There is a salient interaction effect of the item quality and the generating model factors affecting all the statistics. As can be observed from Table 2, in general, the main effect of the generating model indicates that, for $S-X^2$, LR, and W tests, the DINA model was easier to reject when the data were generated with the DINO model, and vice versa. Power for rejecting *A*-CDM was generally higher when data were generated with the DINA model. The effect on the power of LM was different: the power for rejecting DINA and DINO models was higher for data generated using the *A*-CDM, and the power for rejecting *A*-CDM was close to 0, regardless the generating model - .09 and .13 for data generated with DINA and DINO models, respectively. In short, LM tended to reject models different from *A*-CDM. In the case of the $S - X^2$ power, power increased as the item quality got better, but the increment was larger for models which were easier to distinguish (i.e., DINA vs. DINO, *A*-CDM vs. DINA). This relationship between item quality and generating models was affected by the test length in the case of LR, W, and LM tests. This three-way interaction was very similar for the LR and W tests, so we only depicted it for the W test (see Figure 5). Power was always equal to 1.00 in the high item quality conditions, regardless the test length. In the medium item quality conditions, power was also very high when comparing the more distinguishable models (i.e., DINA vs. DINO, *A*-CDM vs. DINA), even when test was composed by a small number of items (J = 12). In the low item quality conditions, the LR and W test only can differentiate between the DINA and DINO models, but only if the number of items was at least 24. In the case of the LM test, this three-way interaction had only a salient effect on the power for rejecting DINA and *A*-CDM models. However, power were generally only acceptable for rejecting DINA and DINO models when the generating model is *A*-CDM, regardless the test length and the quality of the items.

Insert Figure 5 here

Discussion

Even though the interest in CDMs began in response to the growing demand for a better understanding of what students can and cannot do, CDMs have being recently applied to data from different contexts such as psychological disorders (de la Torre et al., 2015; Templin & Henson, 2006) and competency modeling (Sorrel et al., 2016). Item quality has been found to be typically low outside of the educational context. In addition, according to the literature this is an expected result of applications where the attributes are specified *post hoc* (i.e., CDM are retrofitted; Rupp & Templin, 2008).The proper application of a statistical model requires the assessment of model-data fit. One important question that is raised by these new applications is how item quality may affect the available procedures for assessing model fit. While extensive studies have been conducted to evaluate the performance of various fit statistics at the test (e.g., Chen et al., 2013, Liu et al., 2016) and person levels (e.g., Liu et al., 2009; Cui & Leighton, 2009), the item-level is probably the one who has received less attention in previous literature. The statistical properties of the of the item fit statistics remains unknown (e.g., $S-X^2$, LR, and LM tests) or need further investigation (e.g., W test). Taking the above into account, this study provides information about the usefulness of these indexes on different plausible scenarios.

In order to employ item fit statistics in practical use, it is necessary that Type I error is close to the nominal value and that they have a great power to reject false models. In the case of the statistic evaluating absolute fit, $S-X^2$, although it has been found to have a satisfactory Type I error, its power is far from reaching acceptable values. These results are in line with previous studies assessing the performance of χ^2-like statistics in the context of the DINA model (Wang et al., 2015). Here we extent these results to compensatory and additive models (i.e., DINO and *A*-CDM). In conclusion, given its poor performance in terms of power, decisions cannot be made based only on this indicator. There are, however, a number of possible solutions for dealing with this problem that need to be considered in future studies. For example, Wang et al. (2015) have shown how the Stone's (2000) method can be applied to avoid low power in the case of the DINA model. As far as we know, this method has not yet being included in the software available.

Overall the Type I error and power comparisons favour LR and W tests over the LM test. However, and more importantly, Type I error is only acceptable (i.e., $\alpha \cong .05$) when the item quality is high: with a very few exceptions, Type I error with medium and low quality items is generally inflated. We tentatively attribute these results to the noise in the estimation of the item parameters and the standard errors in those conditions. This also applies in other contexts such as the evaluation of differential item functioning (e.g., Bai, Sun, Iaconangelo, & de la Torre, 2016). Particularly in the case of the LR test, in medium item quality conditions this can be compensated by an increase in the number of respondents and items when the true model is DINA or *A*-CDM. For the DINO model Type I error is highly inflated even in those conditions, which is not consistent with the previous results of de la Torre and Lee (2013). On the other hand, when we correct the actual alpha so that it corresponds to the nominal level, we found that the power is still generally high in the medium item quality conditions. Monte Carlo methods can be used in practical settings to approximate the distribution of the statistics under the null hypothesis as it is done in the simulation study (e.g., Rizopoulus, 2006). All things considered, this means that, most likely, we will not choose an incorrect model if we use LR or W test and the item quality is at least medium, which is consistent with de la Torre and Lee's results for the W test. However, this does not mean that CDMs cannot be applied in poor quality items conditions. In these situations the model fit of the test should be assessed as a whole and it should be ensured that the derived attribute scores are valid and reliable. Another promising alternative is to employ a strategy that makes the best of each statistic. According to our results, $S-X^2$, LR, and W statistics can be used simultaneously as a useful tools for assessing item fit in empirical applications. Among all the models fitting the data according to the $S-X^2$ statistic, we will choose the one pointed by the LR or the W test as the most appropriate model.

Even though the LR test was found to be relatively more robust than the W test, the power of W test was slightly higher. Another advantage of using the W test is that it requires only the unrestricted model to be estimated. In contrast, the LR test required $J_{K_j^*>1} NR + 1$ models to be estimated, where *NR* is the number of reduced models to be tested. For example, for one of the conditions with 36 items and 1,000 examinees the computation of the LR and W test requires 2.44 minutes and 6 seconds, respectively. In other words, the W test was 24 times faster than the LR test. Furthermore, in a real scenario, multiple CDMs can be fitted within the same test. Thus, a more exhaustive application of the LR test would require comparing the different combinations of the models, and lead to substantially longer time to implement the LR test. Future studies should explore how this limitation can be addressed.

Although we introduced the LM test as an alternative for assessing fit at the item level, we found that its performance is highly affect by the underlying model: it tended to keep *A*-CDM and reject DINA and DINO models. This test focuses on the distance between the restricted and the unrestricted item parameter estimates. A possible explanation for this poor performance is that the computation of this difference (i.e., the score function) relies on a good estimation of the attribute the joint distribution. In this regard, Rojas et al. (2012) found that fitting an incorrect reduced CDM may have a great impact on the attribute classification accuracy, affecting the estimation of the attribute joint distribution, and thus the performance of this test.

To fully appreciate the current findings, some caveats are in order. A first caveat relates to the number of attributes. In certain application fields the number of attributes can be high. For example, Templin and Henson (2006) specifies 10 attributes corresponding to the 10 *DSM–IV–TR* criteria for pathological gambling. Thus, it is recommended that future research examine the effect of the number of attributes. Second, all items were simulated to have the same discrimination power. In a more realistic scenario, discriminating and nondiscriminating items are mixed. Third, we focus on inferential statistical evaluation. Future studies should consider other approximations. For example, goodness-of-fit descriptive measures have been shown to be useful in some situations. Chen et al. (2013) found that fit measures based on the residuals can be effectively used at the test level. Kunina-Habenicht, Rupp, & Wilhelm (2012) found that the distributions of the RMSEA and MAD indexes can be insightful when evaluating models and *Q*-matrices in the context of the log-linear model framework. New studies might try to extend this results to other general frameworks.

References

Bai, Y., Sun, Y., Iaconangelo, C. & de la Torre, J. (2016, July). *Improving the Wald test DIF detection under CDM framework*. Paper presented at the International Meeting of Psychometric Society, Asheville, NC.

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange Multiplier tests: an expository note. *American Statistician*, *36*, 153-157.

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.

Cui, Y., & Leighton, J.P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*, 429-449.

de la Torre, J, van der Ark, L.A., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*. doi:10.1177/0748175615569110

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. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179-199.

DiBello, L., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, *Psychometrics*) (pp. 979–1027). Amsterdam: Elsevier.

Engle, R.F. (1983). Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics. In M. D. Intriligator & Z. Griliches (Eds.), *Handbook of Econometrics* (Vol. II) (pp. 796-801). New York, NY: Elsevier.

Glas, C. A. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika*, *64*, 273-294.

Glas, C. A., & Suárez-Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*, 87-106.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333–352.

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois, 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.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and psychological measurement*, *56*, 746-759.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59–81.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, *33*, 519-537.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement*, *33*, 579-598.

Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics, 41*, 3-26.

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.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score" equatings". *Applied Psychological Measurement*, *8*, 453-461.

Kang, T., & Chen, T. T. (2008). Performance of the Generalized S-X2 Item Fit Index for Polytomous IRT Models. *Journal of Educational Measurement*, *45*, 391-406.

Ma, W., Iaconangelo, C., & Torre, J. de la. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*, 200–217.

Ma, W., & de la Torre, J. (2016). GDINA: The generalized DINA model framework. R package version 0.9.8. http://CRAN.R-project.org/package=GDINA

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187-212.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289-298.

Ravand, H. (2015). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, doi: 0734282915623053.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of statistical software*, *17*, 1-25.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2015). CDM: Cognitive Diagnosis Modeling. R package version 4. 6-0. http://CRAN.R-project.org/package=CDM

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 meeting of the National Council on Measurement in Education, Vancouver, Canada.

Rupp. A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*, 219-E262.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitively diagnostic models - a case study. *Educational and Psychological Measurement*, *67*, 239-257.

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.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness of fit test statistic in IRT models. *Journal of Educational Measurement*, *37*, 58-75.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*, 331-352.

Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In Frederiksen, N., Glaser, R., Lesgold, A., & Safto, M. (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453–488). Hillsdale: Erlbaum.

Templin J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305.

von Davier, M. (2005). *A General diagnostic model applied to language testing data*. ETS Research Report. Princeton, New Jersey: ETS.

Wang, C., Shu, Z., Shang, Z., & Xu, G. (2015). Assessing item level fit for the DINA model. *Applied Psychological Measurement*. doi:10.1177/0146621615583050

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*, 479-498.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245-262.

Table 1

Marginal Means and Effect Sizes of the ANOVA Main Effects for the Type I Error

Item fit statistic	*N*			*DIM*			*J*				*IQ*				*MOD*				Grand mean
	$\hat{\omega}^2$	500	1,000	$\hat{\omega}^2$	UNI	BI	$\hat{\omega}^2$	12	24	36	$\hat{\omega}^2$	LD	MD	HD	$\hat{\omega}^2$	DINA	A-CDM	DINO	
$S-X^2$	.00	.07	.07	.00	.06	.07	.00	.06	.07	.07	**.01**	.07	.07	.06	**.03**	.06	.08	.06	.06
LR	**.01**	.21	.18	.00	.19	.20	**.02**	.22	.19	.17	**.33**	.30	.23	.06	**.14**	.15	.17	.27	.19
W	**.03**	.31	.27	.00	.29	.29	**.17**	.36	.28	.24	**.71**	.51	.29	.08	**.18**	.24	.27	.36	.29
LM	**.01**	.15	.13	**.02**	.13	.15	**.03**	.16	.14	.13	**.30**	.16	.20	.07	**.60**	.16	.01	.26	.14

Note. Effect size values greater than .010 are shown in bold. Shaded cells correspond to Type I error in the [.02, .08] interval. N = Sample size; DIM = Correlational structure; J = Test length; IQ = Item quality; MOD = Generating model; Uni: Unidimensional; Bi: Bidimensional; LD = Low discrimination; MD = Medium discrimination; HD = High discrimination.

Table 2

Marginal Means and Effect Sizes of the ANOVA Main Effects for the Power for Rejecting a False Reduced Model

Fitted, false model	Item fit statistic	*N*			*DIM*			*J*				*IQ*				Generating, true model (MOD)			Grand mean
		$\hat{\omega}^2$	500	1,000	$\hat{\omega}^2$	Uni	Bi	$\hat{\omega}^2$	12	24	36	$\hat{\omega}^2$	LD	MD	HD	$\hat{\omega}^2$	A-CDM	DINO	
DINA	$S-X^2$	**.25**	.16	.25	.00	.21	.20	**.40**	.12	.22	.27	**.81**	.07	.13	.42	**.53**	.13	.29	.21
	LR	**.13**	.68	.79	**.03**	.76	.71	**.26**	.62	.76	.82	**.78**	.35	.85	1.00	**.18**	.67	.80	.73
	W	**.14**	.74	.84	**.07**	.82	.75	**.22**	.70	.81	.86	**.76**	.48	.89	1.00	**.22**	.72	.86	.79
	LM	**.02**	.66	.70	.00	.67	.69	**.13**	.63	.67	.74	**.62**	.53	.61	.90	**.82**	.95	.42	.68
																	DINA	DINO	
A-CDM	$S-X^2$	**.29**	.23	.35	**.03**	.28	.31	**.15**	.24	.30	.34	**.86**	.09	.16	.63	**.20**	.34	.24	.29
	LR	**.18**	.64	.75	.00	.69	.69	**.39**	.56	.72	.80	**.89**	.22	.85	1.00	**.09**	.73	.65	.69
	W	**.22**	.65	.77	.00	.71	.71	**.37**	.60	.72	.81	**.89**	.27	.87	1.00	**.04**	.73	.69	.71
	LM	**.04**	.09	.13	**.15**	.07	.15	**.16**	.05	.13	.15	**.59**	.05	.02	.28	**.05**	.09	.13	.11
																	DINA	A-CDM	
DINO	$S-X^2$	**.21**	.22	.31	**.03**	.25	.28	**.43**	.17	.27	.35	**.86**	.07	.17	.55	**.58**	.37	.16	.26
	LR	**.10**	.78	.86	.01	.81	.83	**.32**	.71	.85	.91	**.76**	.51	.96	1.00	**.34**	.91	.73	.82

	W	**.13**	.80	.88	**.01**	.83	.85	**.38**	.73	.87	.92	**.79**	.56	.97	1.00	**.39**	.92	.76	.84
	LM	**.06**	.57	.63	.00	.60	.60	**.01**	.58	.61	.62	**.28**	.51	.61	.69	**.90**	.25	.96	.60

Note. Effect size values greater than .010 are shown in bold. Shaded cells correspond to power in the [.80, 1.00] interval. N = Sample size; DIM = Correlational structure; J = Test length; IQ = Item quality; MOD = Generating model; Uni: Unidimensional; Bi: Bidimensional; LD = Low discrimination; MD = Medium discrimination; HD = High discrimination.

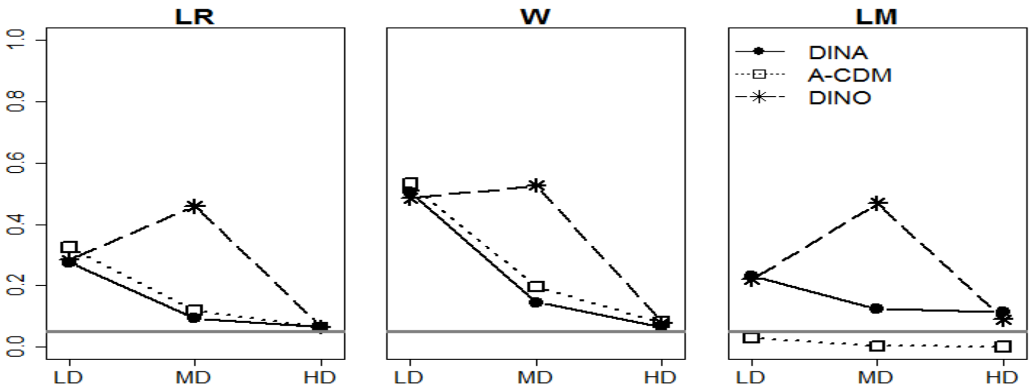

Figure 1. Two-way interaction of Generating model × Item quality with LR, W, and LM Type I error as dependent variables. The horizontal gray line denotes the nominal Type I error (α = .05).

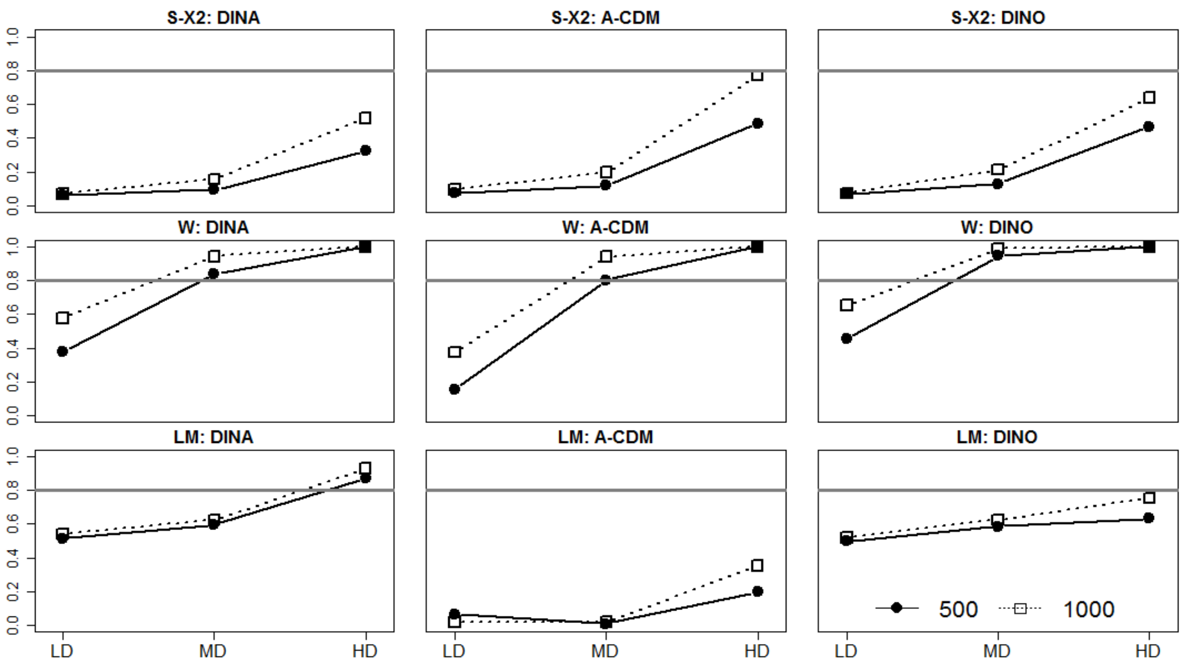

Figure 2. Two-way interaction of Sample size × Item quality with $S-X^2$, W, and LM power as dependent variables. The horizontal gray line represents a statistical power of .80.

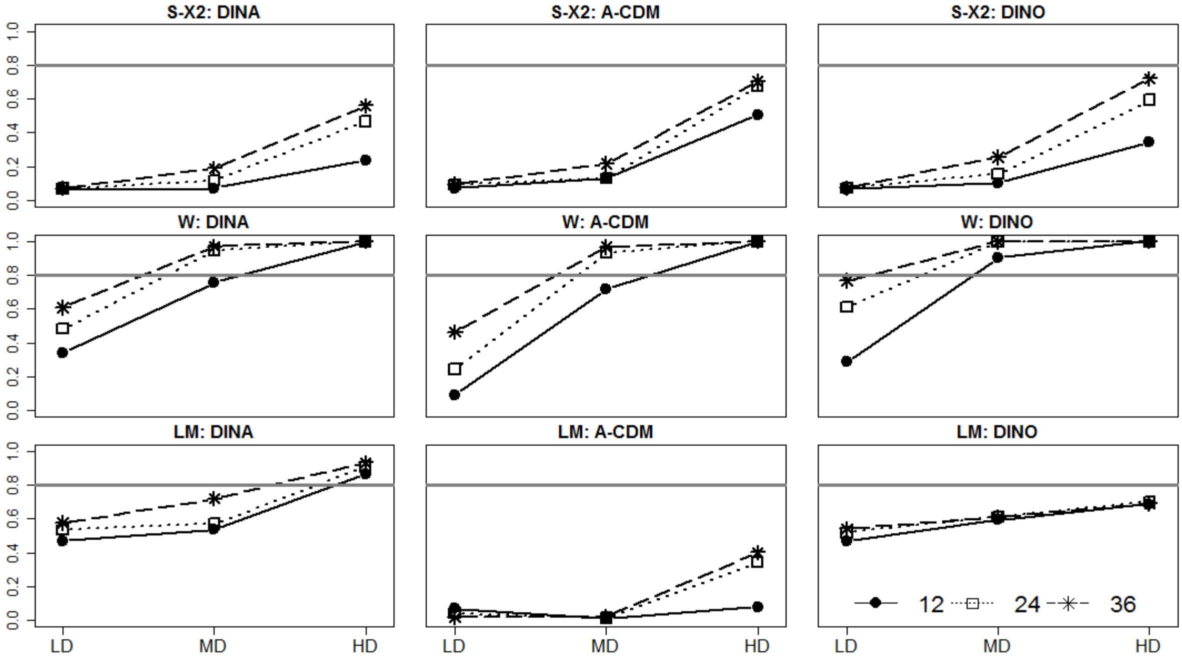

Figure 3. Two-way interaction of Test length × Item quality with $S-X^2$, LR, and LM power as dependent variables. The horizontal gray line represents a statistical power of .80.

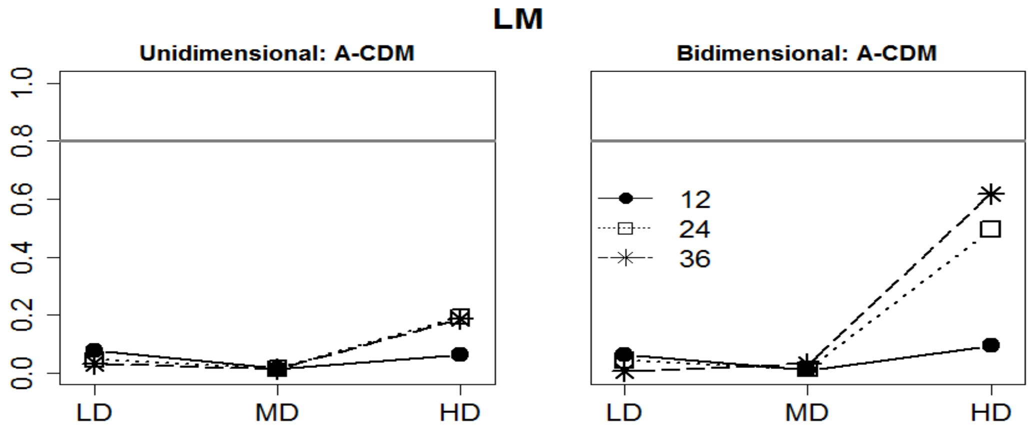

Figure 4. Thee-way interaction of Correlational structure × Item quality × Test length with LM power for rejecting A-CDM when is false as dependent variable. The horizontal gray line represents a statistical power of .80.

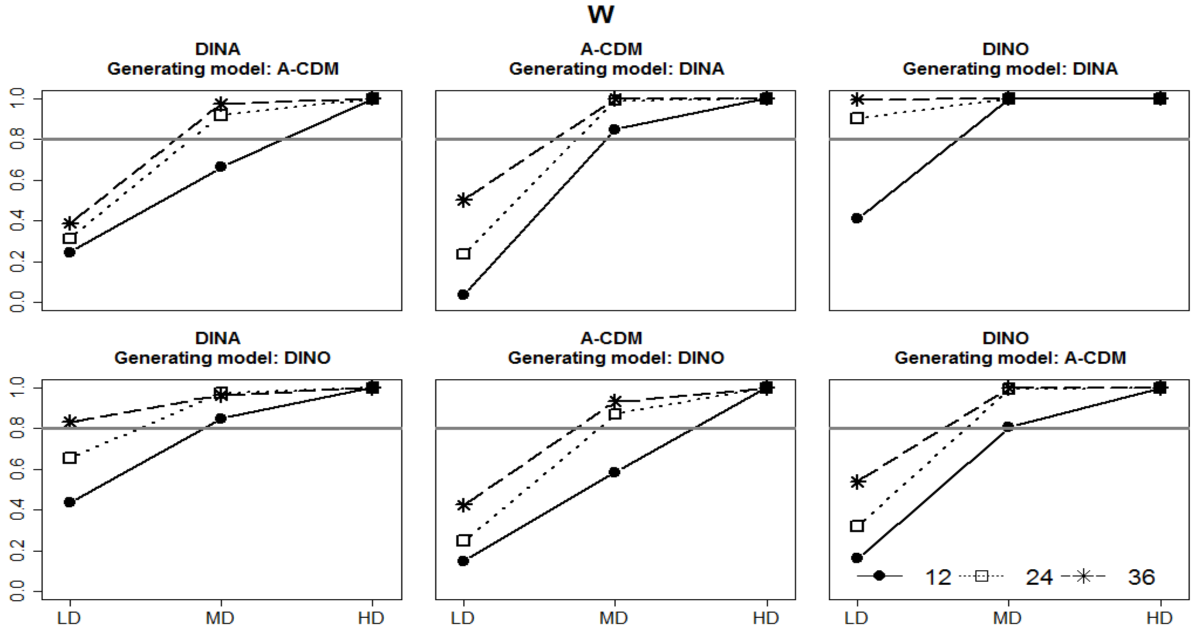

Figure 5. Thee-way interaction of Generating model × Item quality × Test length for W test power for rejecting DINA, *A*-CDM, and DINO when they are false as dependent variables. The horizontal gray line represents a statistical power of .80.