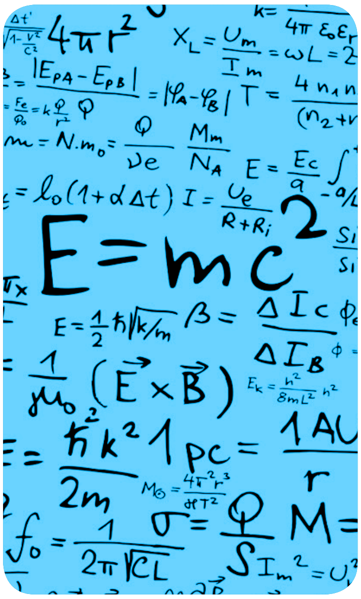


MÁSTERES de la UAM

Facultad de Psicología
/ 14-15

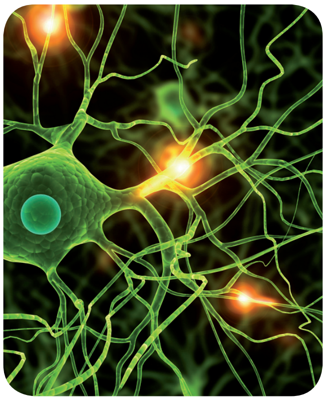
UAM
UNIVERSIDAD AUTONOMA
DE MADRID

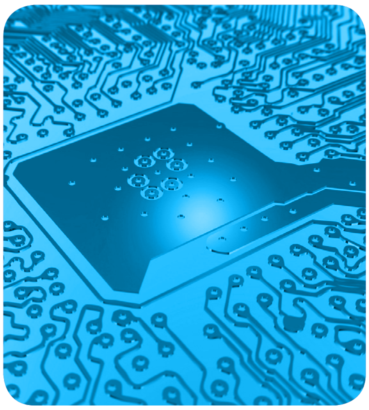
Metodologías
de las Ciencias
del Comportamiento
y de la Salud

Campus Internacional
excelencia UAM CSIC+

Item Fit Evaluation in Cognitive Diagnosis Modeling

Miguel A. Sorrel Luján





UAM
EDICIONES

Item Fit Evaluation in Cognitive Diagnosis Modeling

By

Miguel A. Sorrel

A MASTER'S THESIS

Submitted to

Universidad Autónoma de Madrid

Advisors: Julio Olea and Francisco J. Abad

In partial fulfillment of the requirements for the degree of

MASTER OF METHODOLOGY OF BEHAVIOURAL AND HEALTH SCIENCES

Interuniversitary Master's Degree (UNED, UCM, and UAM)

Madrid, Spain

June, 2015

Acknowledgement

Thank you to my advisors, Dr. Francisco José Abad and Dr. Julio Olea. They provided me with a wealth of guidance. I would also like to acknowledge Dr. Juan Ramón Barrada who crossed the pond with me and the Dr. Jimmy de la Torre for his very useful comments.

Table of contents

Abstract

In the field of cognitive diagnosis modelling there has been scarce research related to the item fit evaluation. Although general cognitive diagnosis models might provide better model–data fit than reduced models, there are several reasons that make reduced models preferable to the general models. Mainly, reduced models require smaller sample sizes, have parameters with a more straightforward interpretation, and lead to better classification rates when the sample size is small. Thus, it is relevant to assess if item fit indexes allow to select the most parsimonious model for each item. This study investigates the performance of various item fit statistics and provides information about the usefulness of these indexes on different scenarios. Five statistics were considered: RMSEA, S-X2, the LR test, the Wald test, and the LM test. Results show that the empirical significance of the LR test and the Wald test conforms more closely to the nominal significance level and these statistics have a higher statistical power when items are highly discriminative than the others statistics. However, both tests are highly affected by the item discrimination. It implies that we cannot differentiate between DINA and *A*-CDM models in practical settings when the item discrimination is low.

Keywords**:** *cognitive diagnosis modeling, model fit, simulation study, validity*

Theoretical introduction

Cognitive diagnosis modeling

Over the last years there has been an increase of interest in psychometric models referred to as cognitive diagnosis models (CDMs). Based on the review of existing labels for these models that have been used in the literature (e.g., *cognitively diagnostic models*, Henson & Douglas, 2005, *cognitive psychometric models*, Rupp, 2007; *multiple classification models*, Haertel, 1989; *structured IRT models*, Rupp & Mislevy, 2007), Rupp and Templin (2008) offered the following definition:

> Diagnostic classification models (DCM) are probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables. The predictor variables are combined in compensatory and non-compensatory ways to generate latent classes. DCM enable multiple criterion-referenced interpretations and associated feedback for diagnostic purposes, which is typically provided at a relatively fine-grain size. This feedback can be, but does not have to be, based on a theory of response processing grounded in applied cognitive psychology. Some DCM are further able to handle complex sampling designs for items and respondents, as well as heterogeneity due to strategy use. (p.226).

Following this definition, we find that cognitive diagnosis modeling (CDM) is an interdisciplinary approach to diagnostic assessment. So that the psychological processes underlying performance on items can be modeled, cognitive models need to be developed for those domains that need to be assessed. In this regard, CDM establishes a link between cognitive psychology and statistical modeling.

In the field of education, researchers have proposed different theories about how students represent knowledge and develop competence in a subject domain (e.g. Mathematics). In addition, CDM can facilitate inferences more relevant to learning. That is the reason why are widely employed in *cognitively diagnostic educational assessment* (Leighton & Gierl, 2007; Nichols, Chipman, & Brennan, 1995). Despite the fact that few empirical studies have been published out of the educational context, CDMs can be applied to other contexts. Two good examples are the study of Templin and Henson (2006) who demonstrate how the hypothesized underlying factors contributing to pathological gambling can be measured with the *deterministic input, noisy "or" gate* (DINO) model, and the study of García, Olea, and de la Torre (2014) who found that CDMs could achieve an accurate fit to the responses of a situational judgement test (SJT) measuring 6 professional competencies based on the great eight model (Bartram, 2005). Moreover, SJTs conforms another promising context of use for CDMs. Some authors consider that one of the most critical issues for the future of SJT research is to provide enough evidence about the constructs included therein. In essence, experts call for a new approach to the nature of the construct in SJTs. In a recent review of SJTs (Weekley, Hawkes, Guenole, & Ployhart, 2015), it is recognized that, among the current and principal lines of research in SJTs, the application of CDMs to SJTs is included.

Unlike traditional Item Response Theory (IRT) models, which generally model continuous latent variables, the latent variables in CDMs are discrete, consisting either of dichotomous (e.g., mastery vs non-mastery), or polytomous levels (e.g., "good performance", "fair performance", and "poor performance"). Over the last two decades, several CDMs that can be successfully applied across a wide variety of settings have been developed. At present, most of research has been focused on CDMs for dichotomous attributes. In contrast, only a few CDMs accommodating polytomous attributes can be found. Thus, taking into account that basic topics are still needed of further investigation, in this work we will focus on CDMs for dichotomous attributes.

Through this section we will illustrate the main characteristics of CDMs with empirical data: their multidimensional nature, their confirmatory nature, the complexity of their loading structure, and the type of latent predictor variables they contain (Rupp and Templin, 2008). Data were taken from the administration of a SJT composed of 23 items that presented situations about various student-related issues (teamwork, studying for exams, organizing, accomplishing assignments, interpersonal skills, social responsibility, perseverance, integrity). This database was used by Sorrel, Olea, Abad, Aguado, and Lievens (2015) to propose that CDMs could represent a new psychometric approach to obtain evidence of validity for SJTs, to assess their reliability, and to score the different skills or abilities that are theoretically measured by the test. Items 1 and 2 are shown in Figure 1.

ITEM 1: When studying for an exam, do you find that you reach best results when:

a. you start planning and setting aside time in advance
b. work in a clean environment, even if it means taking time away from studying
c. wait for inspirations before becoming involved in most important study tasks
d. wait until the last day or so to study, knowing that you have to get it done now

The most effective response to this situation would be:

ITEM 2: Your professor announces in class that undergraduate students are needed to help run subjects for his upcoming study. While you would not receive any formal sort of extra credit, the professor would appreciate any volunteers. Given the following choices, which option would you choose?

a. Examine your schedule and offer to volunteer a couple hours a week when it is personally convenient.
b. Examine your schedule and offer to volunteer as many hours as you can.
c. Realize that you would have to give up some of your free time and choose not to volunteer.
d. Offer to run subjects only if you are paid.

The most effective response to this situation would be:

Figure 1. Items 1 and 2 of the SJT. Most appropriate answers are shown in bold

As we said before, CDMs are inherently confirmatory, as shown by their loading structure. The loading structure of a CDM, which is commonly known as Q-matrix (Tatsuoka, 1983), is a mapping structure that indicates the skills required for successfully answering each individual item. In the CDMs literature there is a consistent notation that will be employed in this work. Respondents (e.g., learners, patients, applicants) are indexed by $i = 1, . . . , I$, assessment items are indexed by $j = 1, . . . , J$, and attributes (e.g., borrowing numbers, a diagnostic criteria for Pathological Gambling, a professional competency) are indexed by $k = 1, . . . , K$. Observed responses of respondent i to item j are denoted Xij, while the skill profile vector of a respondent is denoted α_i, such that α_{ik} indexes whether respondent i has mastered skill k ($\alpha_{ik} = 1$) or not ($\alpha_{ik} = 0$).

In the database that we are employing to illustrate, Sorrel et al. (2015) identified four attributes based on the test specifications and the expert's ratings obtained by a Delphi method. The four attributes underlying performance on the SJT are *Study habits*, *Study attitudes*, *Helping others*, and *Generalized compliance*. More details about these attributes are provided in Annex 1.

A Q-matrix can be viewed as a cognitive design matrix that makes explicit the internal structure of a test. A portion of the Q-matrix used for the illustration purpose is displayed in Table 1. The Q-matrix is a *J* x *K* matrix of zeros and ones, where the element on the *j*th row and *k*th column of the matrix, q_{jk} indicates whether skill *k* is required to correctly answer item *j* (q_{jk} = 1) or not (q_{jk} = 0).

Table 1

Q-matrix

Item	*Study habits*	*Study attitudes*	*Helping others*	*Generalized compliance*
1	1	0	0	0
2	0	1	1	0
3	1	0	0	0
4	0	1	1	1
5	1	1	0	0

Note. 1 = the attribute is required to choose the most effective response option; 0 = the attribute is not required to choose the most effective response option.

As can be seen from the Table 1, two items involved only one attribute, two items involved two attributes, and one item involved three attributes. Item 1, shown in Figure 1, measures Study habits. Students who engage in regular acts of studying probably will answer this item correctly. Item 2, which is also shown in Figure 1, measures Study habits and Helping others. Probably, students who approve the broader goals of education (e.g. education should be within everyone's reach) and tend to help others will correctly answer this item.

Confirmatory Factor Analysis (CFA) models and IRT models usually have a *simple structure*, that is, each item loads only in one factor (for a detailed discussion, see McDonald, 1999). Factors as defined in these models are generally broader dimensions (e.g. number ability). On the contrary, in the case of CDMs factors, commonly referred to as attributes, are narrowly defined (e.g. fraction subtraction). Each item typically requires more than one attribute. This leads to a *complex loading structure* where each item is specified in relation to multiple attributes. This complex loading structure, in terms of multidimensional IRT, is known as *within-item multidimensionality* (Adams,Wilson & Wang, 1997) and is reflected in the "1s" of the Q-matrix as it happen, for example, in the componential IRT models (Embretson, 1999; Fischer, 1995; Van der Linden & Hambleton, 1997).

We will now graphically compare a few prototypical models that we have discussed. Figure 2 depicts three different psychometric models so that we could better understand the difference between simple structure and complex structure. Note that the bars for categorical variables reflect *thresholds* (i.e. the probability of a respondent possessing or mastering dichotomous attributes and probabilities of correct response for dichotomous observed responses). In these figures, these bars are located at arbitrary points to simplify the illustrations. The number of underlying attributes has been reduced to two for didactic purposes. Figure 2A and 2B shows a bi-dimensional CFA and IRT models with simple structures and contrast it with Figure 2C which shows a bi-dimensional CDM with a complex loading structure (i.e. items 1, 3, 6, and 9 load on both dimensions). In this way CDMs could be understood as an extension of traditional multidimensional IRT and CFA models that are particularly suitable to a complex loading structure.

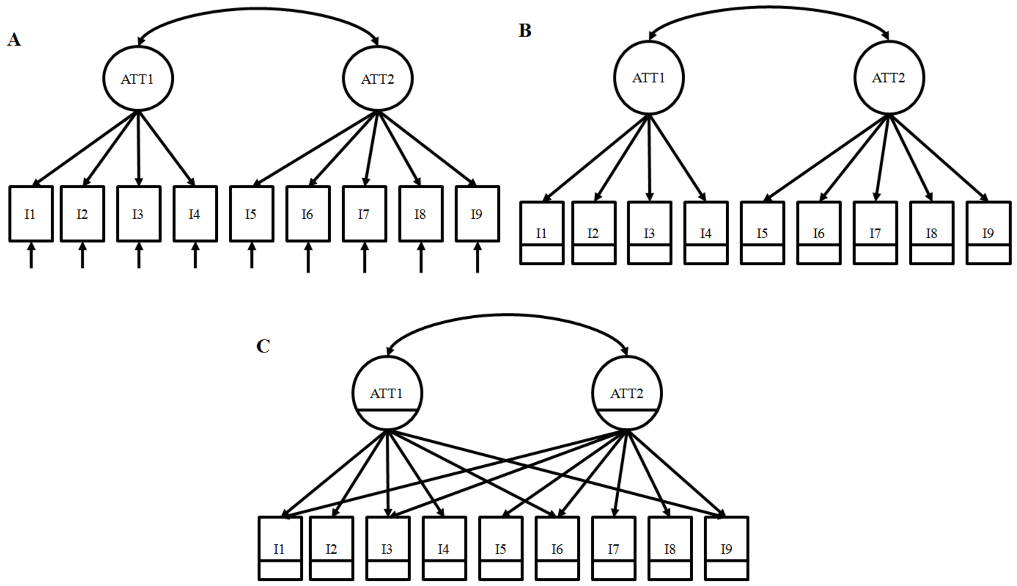


Figure 2. Representation of the different prototypical models. Model A = Two-dimensional CFA model with simple loading structure; Model B = Two-dimensional IRT model with simple loading structure; Model C = Two-dimensional CDM with complex loading structure.

In short, CDMs are latent class models (Haagenars & McCutcheon, 2002) that classify respondents into some latent classes according to similarity of their responses to test items. They are called *restricted* latent class models because the number of latent classes is restricted by the number of attributes involved in answering items of a test.

With K attributes underlying performance on a given test, the respondents will be classified into 2^K latent classes (the number 2 indicates that there are two possible outcomes for each attribute: mastery or non-mastery). Latent classes are indexed by $l = 1, \ldots, 2^K$. For example, with four attributes required to perform successfully on the items of a given test, test takers will be classified into $2^4 = 16$ latent classes. Table 2 shows the attribute class probabilities of a sample composed of 138 respondents which were classified these 16 latent classes.

Table 2.

Latent class probabilities

Latent Class	Attribute profile	Class probability	Class expected frecuency
1	0000	.09	12
2	1000	.00	0
3	0100	.06	8
4	1100	.00	0
5	0010	.01	1
6	1010	.04	6
7	0110	.01	1
8	1110	.04	6
9	0001	.10	14
10	1001	.11	15
11	0101	.08	11
12	1101	.08	11
13	0011	.00	0
14	1011	.16	22
15	0111	.00	0
16	1111	.21	29

The main output of CDM for each respondent is a vector of estimates denoting in terms of probability the state of mastery of the *i*th respondent on each of the attributes. These probabilities are converted in dichotomous scores (i.e., mastery or non-mastery) by comparing them to a cut-off score (usually .5; de la Torre, Hong, & Deng, 2010; Templin & Henson, 2006) to define these attribute profiles.

Generally, CDMs can be grouped into three families as shown in Table 3. These are some of the widely employed CDMs. A more detailed classification could be found in Rupp, Templin, and Henson (2010). Considering the manner in which the latent predictor variables are combined, CDM can be divided into compensatory and non-compensatory models. In *non-compensatory latent-variable models*, a low value on one latent variable cannot be compensated by a high value on another latent variable whereas in *compensatory latent-variable models* a low value on one latent variable can be compensated by a high value on another latent variable. Non-compensatory models are better aligned with cognitive theory in some cases in which it is strongly believed that the respondent must have mastered all the attributes within the item in order to get the item correct. For example, if a fraction substraction item measures separate a whole number from a fraction, subtracts numerators, find a common denominator, and reduce answers to the simplest form, all of these operations/attributes are required to answer the item correctly, and a lack of certain attributes cannot be compensated by possessing other attributes. General CDMs allow for both types of relationships within the same test.

Table 3.

CDM Types.

CDM type	Examples	Author(s)
Non-Compensatory	1) deterministic-input, noisy-and-gate (DINA) model	Junker & Sijsma (2001)
	2) non-compensatory reparamatrized unified model (NC-RUM)	DiBello et al. (1995); Hartz (2002).
Compensatory	1) deterministic-input, noisy-or-gate (DINO) model	Templin & Henson (2006)
	2) compensatory reparamatrized unified model (C-RUM)	Hartz (2002)
General	1) general diagnositc model (GDM)	Von Davier (2005)
	2) log-linear CDM (LCDM)	Henson, Templin, & Willse (2009)
	3) generalized DINA model (G-DINA)	de la Torre (2011)

A critical concern is selecting the most appropriate model from the available CDMs. To a great extent, the process of determining the most appropriate model is a validation process given that the results of statistical models are meaningless when the model fit is poor. 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) 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.

This study investigates the performance of various item fit statistics and provides information about the usefulness of these indexes on different scenarios within the generalized DINA (G-DINA) model framework developed by de la Torre (2011). In the next sections we will describe the G-DINA framework, the item fit evaluation in CDM, and the objectives of the current study.

The generalized DINA model framework

As shown by de la Torre (2011), many of the widely known CDM can be represented via the *generalized deterministic inputs, noisy "and" gate* (G-DINA; de la Torre, 2011) model, which is a generalization of the *deterministic inputs, noisy "and" gate* (DINA; de la Torre, 2009; Junker & Sijtsma, 2001) model. Thus G-DINA allows to estimate a different model of each item on the same test. The G-DINA model describes the probability of success on item *j* in terms of the sum of the effects of involved attributes and their 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 $\boldsymbol{\alpha}_{lj}^*$ and has its own associated probability of success, written as

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

where δ_{j0} is the intercept for item *j*, δ_{jk} is the main effect due to α_k, $\delta_{jk\,k'}$ is the interaction effect due to α_k and $\alpha_{k'}$, and $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1,\ldots,\alpha_{K_j^*}$. Thus, there are $2^{K_j^*}$ parameters to be estimated for item *j*.

In this work we will focus on two reduced models which are nested in the G-DINA model: DINA model and *additive* CDM (*A*-CDM; de la Torre, 2011). DINA model is one of the most widely used CDMs. In the case of *A*-CDM, as we will see it has more parameters per item than the DINA model. It is therefore interesting to compare the performance of the item fit statistics for these two different models. These two models are described below.

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_0 and $\delta_{j12\ldots K_j^*}$. Therefore, the probability of success could be written as

$$P(\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \delta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} . \quad (2)$$

That is, in the DINA model, except for the attribute vector $\boldsymbol{\alpha}_{lj}^* = 1_{K_j^*}$, the $2^{K_j^*-1}$ latent groups have identical probability of correctly answer the item j. As such, the DINA model has two parameters per item, commonly known as guessing and slipping parameters.

When all the interaction terms are dropped, the G-DINA model reduces to *A*-CDM. The probability of a correct response for the *A*-CDM is given by

$$P(\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\, \alpha_{lk} . \quad (3)$$

This model indicates that mastering attribute α_k increases the probability of success on item j by δ_{jk} independently of the contributions of the others attributes. The *A*-CDM has $K_j^* + 1$ parameters per item.

So that we may better understand the differences between these two reduced models, Figure 3 depicts the parameter estimates for G-DINA for two items taken from the SJT test. The vertical axis shows the point estimate for each parameter and the associated standard-error band (i.e. parameter value ± standard error); the red horizontal line indicates the value of 0 as visual reference point. We can identify a likely candidate CDM for each item. For example, the pattern of parameter estimates for Item A shows that the interaction effect is essentially 0 ($\delta_{A12} = -.09$). This pattern is consistent with the *A*-CDM model where mastering each has a positive effect on the probability of success independently of the contributions of the others attributes ($\delta_{A1} = .30$ and $\delta_{A2} = .47$). The pattern of estimates for Item B, on the other hand, shows that both main effects estimates could potentially be 0 ($\delta_{A1} = .05$ and $\delta_{A2} = 0$), which is a pattern that is consistent with the DINA model.

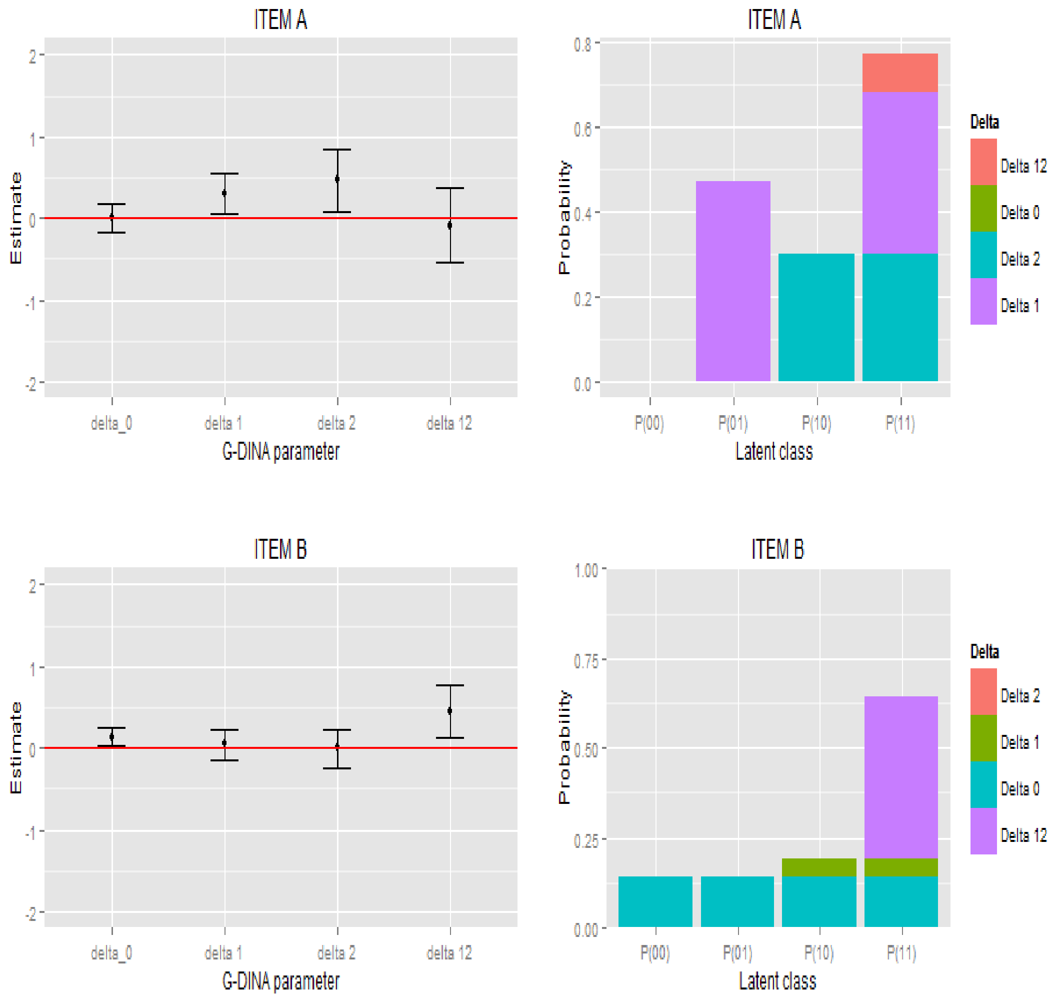


Figure 3. Parameter estimates for G-DINA model for Items A and B and the derived probabilities of success for the different latent classes.

Item fit evaluation in CDM

As Sinharay and Almond (2007) noted, model checking is a crucial part of any model-based statistical analysis: not only provides a vital sanity check that the theory underlying the model can actually predict the phenomena observed in the data, but also suggests improvements to the model. To a great extent, the first concern refers to absolute fit (i.e. the discrepancy between a statistical model and the data), whereas the second one is related to the relative fit (i.e. the discrepancy between two statistical models).

Regarding absolute fit evaluation at item level, the S-X2 item fit statistic (Orlando & Thissen, 2000, 2003) for dichotomous data, which emanates from IRT, has been adapted to the field of CDM based on test scores (i.e. number of correct scores; Sinharay & Almond, 2007). The S-X2 statistic is defined as

$$S - X_j^2 = \sum_{k=1}^{J-1} N_k \frac{(O_{jk} - E_{jk})^2}{E_{jk}(1 - E_{jk})}, \quad (4)$$

where k is the test score, N_k is the number of examinees in group k, and O_{jk} and E_{jk} are, respectively, the observed and predicted proportions of correct responses for group k. The degrees of freedom equals $J - 1 - m$, where m is the number of item parameters estimated and J the number of items.

Within the CDM field, a statistic called item-fit RMSEA has been proposed to absolute fit at the item level (Kunina-Habenicht, Rupp & Wilhelm, 2009). The RMSEA statistic is defined as

$$RMSEA_j = \sqrt{\sum_{l=1}^{2^K} p(\boldsymbol{\alpha}_l)\left[P_{expected}\,(X_j = 1|\boldsymbol{\alpha}_l) - P_{observed}\,(X_j = 1|\boldsymbol{\alpha}_l)\right]^2}, \quad (5)$$

where l denotes the latent class and $p(\boldsymbol{\alpha}_l)$ is the estimated class probability of $\boldsymbol{\alpha}_l$.

Considering the relative fit evaluation, when comparing different nested models there are three common tests than can be used (Buse, 1982): the likelihood ratio (LR) test, the Wald (W) test, and the Lagrange multiplier (LM) test (sometimes called score test). The null hypothesis (H_0) for all three tests is that the reduced model is the "true" model whereas the alternative hypothesis (H_1) sets that the general model is the "true" model". That is, H_0 defines a restricted parameter space. For example, for an item j measuring two attributes in the A-CDM model we assume that the interaction term is equal to 0. We can represent this belief in the form of a null hypothesis and its alternate: H_0: $\delta_{j12} = 0$ and H_1: $\delta_{j12} \neq 0$. A large test statistics indicate that the null hypothesis is false so that the reduced model could not be assumed. These three procedures are asymptotically equivalent (Engle, 1983). In all cases, the statistic is assumed to be asymptotically χ^2distributed with $g = 2^{K_j^*} - p$ 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 under H_0 and H_1 respectively (i.e. restricted and unrestricted estimates of the population parameter). While all three tests address the same basic question, they are slightly different (see Figure 4). In the case of the LR test, we estimate the model under H_0 and under H_1 and look at the loss in the likelihood. If we plot the log-likelihood function, the value of LR can be obtain directly from the values of $\log L(\theta)$ at $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$. When computing the W test, we estimate the model only under H_1 and look at the distance $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}$. Finally, in the case of the LM test, we estimate $\tilde{\boldsymbol{\theta}}$ under H_0 and see if the restricted maximum likelihood estimates are near the unrestricted estimates.

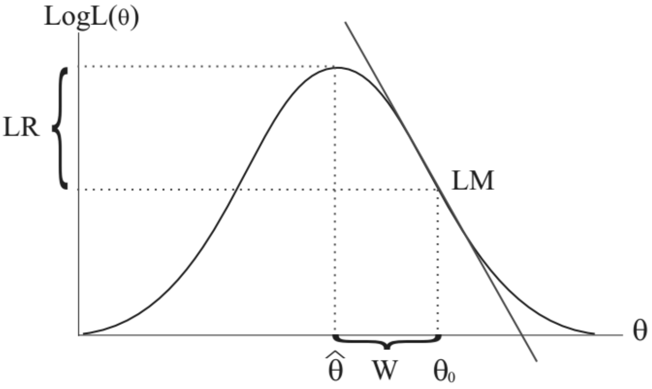


Figure 4. The relation among likelihood ratio test, Wald test, and Lagrange multiplier test.

In this section these three statistical test will be described in greater detail and its application to CDM will be exposed. Before, it is necessary to mention few points about the estimation procedure in CDM. The estimation implemented in the CDM package (Robitzsch, Kiefer, George, & Uenlue, 2015) of R (R Core Team, 2014) is based on an EM algorithm as described in de la Torre (2011). The parameters estimates of the G-DINA model are estimated using marginalized maximum likelihood estimation (MMLE). The conditional likelihood of the observed data **X** is

$$L(\boldsymbol{X}_i|\boldsymbol{\alpha}_l) = \prod_{j=1}^{J} P(\boldsymbol{\alpha}_{lj})^{X_{ij}} [1 - P(\boldsymbol{\alpha}_{lj})]^{1-X_{ij}} . \quad (6)$$

Following this, the log-marginalized likelihood of the response data can be written as

$$l(\boldsymbol{X}) = \log[L(\boldsymbol{X})] = \log \prod_{i=1}^{I} \sum_{l=1}^{L} L(\boldsymbol{X}_i|\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)\ , \qquad (7)$$

where $p(\boldsymbol{\alpha}_l)$ is the prior probability of $\boldsymbol{\alpha}_l$. In the G-DINA model the probability of a correct response on item j, $P(\boldsymbol{\alpha}_{lj})$, can be written as $P(\boldsymbol{\alpha}_{lj}^*)$ which represents the reduced attribute vector from of $P(\boldsymbol{\alpha}_{lj})$. By taking the derivative of $l(\boldsymbol{X})$ with respect to $P(\boldsymbol{\alpha}_{lj}^*)$, the maximization of $\partial l(\boldsymbol{X})$ with respect to $P(\boldsymbol{\alpha}_{lj}^*)$ is the so-called score function in the LM context

$$S(\theta) = \frac{\partial \log L}{\partial P(\boldsymbol{\alpha}_{lj}^*)} = \left[\frac{1}{P(\boldsymbol{\alpha}_{lj}^*)\left(1 - P(\boldsymbol{\alpha}_{lj}^*)\right)}\right][R_{jl} - P(\boldsymbol{\alpha}_{lj}^*)I_{jl}]\ , \qquad (8)$$

where $I_{\boldsymbol{\alpha}_{lj}^*}$ is the number of respondents expected to be in the latent group $\boldsymbol{\alpha}_{lj}^*$, $R_{\boldsymbol{\alpha}_{lj}^*}$ is the number of respondents in the latent group $\boldsymbol{\alpha}_{lj}^*$ expected to answer the item j correctly, and $p(\boldsymbol{\alpha}_{lj}^*|\boldsymbol{X}_i)$ represents the posterior probability that examinee *i* is in latent group $\boldsymbol{\alpha}_{lj}^*$. Thus, the MMLE estimate of $P(\boldsymbol{\alpha}_{lj}^*)$ is given by $\hat{P}(\boldsymbol{\alpha}_{lj}^*) = {R_{\boldsymbol{\alpha}_{lj}^*}}/{I_{\boldsymbol{\alpha}_{lj}^*}}$.

The second derivative of the log-marginalized likelihood with respect to $P(\boldsymbol{\alpha}_{lj}^*)$ and $P(\boldsymbol{\alpha}_{l'j}^*)$ can be shown to be (de la Torre, 2011)

$$-\sum_{i=1}^{I}\left\{p(\boldsymbol{\alpha}_{lj}^*|\boldsymbol{X}_i)\frac{X_{ij} - P(\boldsymbol{\alpha}_{lj}^*)}{P(\boldsymbol{\alpha}_{lj}^*)[1 - P(\boldsymbol{\alpha}_{lj}^*)]}\right\}\left\{p\left(\boldsymbol{\alpha}_{l'j}^*|\boldsymbol{X}_i\right)\frac{X_{ij} - P\left(\boldsymbol{\alpha}_{l'j}^*\right)}{P(\boldsymbol{\alpha}_{l'j}^*)[1 - P(\boldsymbol{\alpha}_{l'j}^*)]}\right\}. \qquad (9)$$

Using $\hat{P}(\boldsymbol{\alpha}_{lj}^*)$ and the observed $\boldsymbol{X}$ to evaluate (9), the appropriate information matrix for the parameters of item j, $\boldsymbol{I}(\widehat{\boldsymbol{P}}_j^*)$, where $\widehat{\boldsymbol{P}}_j^* = \{P(\boldsymbol{\alpha}_{lj}^*)\}$, can be obtained. The square roots of the diagonal elements of $\boldsymbol{I}^{-1}(\widehat{\boldsymbol{P}}_j^*)$ represent the standard errors $SE[\hat{P}(\boldsymbol{\alpha}_{lj}^*)]$.

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** (i.e. the data) given the hypothesis. The likelihood function is defined as $L(\hat{\theta})$ for the null hypothesis and $L(\tilde{\theta})$ for the alternative. The likelihood of the null hypothesis over the alternative is

$$LR = 2\left(\log L(\hat{\theta}) - \log L(\tilde{\theta})\right) \sim \chi^2(g). \quad (10)$$

In the CDM context the LR test can be conducted to determine if the unrestricted model fits the data significant better than the reduced model. Having a test composed of J items, the application of the LR test at the item level implies that J_K comparisons will be made, where J_K is the number of items measuring at least $K = 2$ attributes. For each of the J_K comparisons, a reduced model (i.e. DINA or *A*-CDM) is fitted to a target item, whereas the unrestricted model (i.e. G-DINA) is fitted to the rest of the items. This model is compared to a model where the unrestricted model is estimated for all items.

Wald test

The W test takes into account the curvature of the log-likelihood function, which is denoted by $C(\hat{\theta})$ and defined by the absolute value of $d^2 \log L / d\theta^2$ evaluated at $\theta = \hat{\theta}$. If $r(\theta) = 0$ is a vector of g functional restrictions imposed by H_0 on the *k*-vector $\boldsymbol{\theta}$ ($k > g$), then asymptotically (Buse, 1982)

$$W = [r(\hat{\theta})]' \left[RI(\hat{\theta})^{-1}R'\right]^{-1} [r(\hat{\theta})] \sim \chi^2(g), \quad (11)$$

where R is the g x k matrix of partial derivatives $\partial r(\theta)/\partial\theta$, evaluated at $\hat{\theta}$.

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 reduced model p, this test requires setting up $\boldsymbol{R}_{jp}$, a $\left(2^{K_j^*} - p\right) \times 2^{K_j^*}$ restriction matrix which includes the specific constraints that make the saturated model to be equivalent to the reduced model of interest (the same that were exposed above). The Wald statistic is then computed as

$$W_j = [R \times P_j]'[R \times Var(P_j) \times R]^{-1}[R \times P_j] \sim \chi^2(g), \quad (12)$$

where $P_j = \{P(\boldsymbol{\alpha}_{lj}^*)\}$ are the probability estimates under unrestricted model and $Var(P_j)$ is the inverse of the information matrix.

Lagrange multiplier test

The LM test is based on the slope of the log-marginalized likelihood, which is called score function, $S(\theta) = d \log L \,/\, d\theta$. By definition, $S(\theta)$ is equal to zero when evaluated at the unrestricted MMLE of θ (i.e. $\hat{\theta}$), but not when evaluated at $\tilde{\theta}$. If the constraints were true, we would expect $S(\tilde{\theta})$ to be small, so that the rejection of the null hypothesis is associated with large values of LM. This score function should be weighted by the information matrix. The LM statistic is then defined as

$$LM = S(\tilde{\theta})' \, Var(P_j) \, S(\tilde{\theta}), \qquad LM \sim \chi^2(g). \quad (13)$$

Following the parameter estimation of these models under the G-DINA framework, de la Torre (personal communication, October, 2014) noted that the score function could be assumed to be (8).Then, if we estimate the reduced models (i.e. DINA and *A*-CDM) in their saturated form, we also have an estimation of the information matrix for these reduced models. Thus, we can implement the LM test in CDM research.

The current study

The mayor purpose of item-level evaluation is to find a parsimonious model to fit the sample data while maintaining theoretical meaningfulness. Although general CDMs might provide better model–data fit, as pointed out by de la Torre & Lee (2013) there are several reasons that make specific models preferable to the saturated model. First, in comparison to reduced CDMs, general CDMs require larger sample sizes for item parameter calibration. Second, they have parameters with less straightforward and meaningful interpretations. Third, appropriate reduced models lead to better attribute classification accuracy than the saturated model, particularly when the sample size is small (Rojas, Olea, & de la Torre, 2012).

Hypothesis testing concerns the question of whether data appear to favor or disfavor the null hypothesis. We either reject or fail to reject the null hypothesis. There are two ways to make incorrect inferences. *Type I* errors denoted by α are committed when the null hypothesis is falsely rejected. *Type II* errors denoted by β occur when the null hypothesis is incorrectly accepted. The *power* of a test is the probability of rejecting the null hypothesis when is false (1- β). One test is said to be better than other if it has the maximum power among all test with Type I error less than or equal to some particular level. As far as we know, the statistical properties of RMSEA, S-X2, LR test, and LM test are still unknown. The Wald statistic has already been studied in terms of Type I error and power rates (de la Torre & Lee, 2013) and it has been found that it has a relative accurate Type I error rate, particularly with large samples and small number of parameters. It also has a high power to detect when a reduced model is not appropriate. The effect that other factors not considered in de la Torre and Lee's (2013) study (e.g. test length, item quality) may have on the performance of this statistic remains unclear. Thus, the main purpose of this study is to systematically examine the Type I error and power rates of the item fit statistics described above and provide information about the usefulness of these indexes on different plausible scenarios.

Method

A simulation study was conducted to investigate the performance of the item fit statistics. The following factors varied: a) generating model (MOD): DINA model and *A*-CDM model; b) test length (J): 12 and 24 items; c) sample size (N): 500 and 1,000; d) item quality (IQ) or discrimination (defined as the difference between the maximum and the minimum probabilities of correct response according to the attribute latent profile): .4 and .8. The probabilities of success for individuals who mastered none of the required attributes were fixed to .10 and .30 for the high quality and low quality conditions respectively. The probabilities of success for individuals who mastered all of the required attributes were fixed to .90 and .70 for the high quality and low quality conditions respectively. For the additive model, an increment of $.80/K_j^*$ and $.40/K_j^*$ was associated with each attribute mastery for the high quality and low quality conditions respectively (see Figure 5); e) dimensional magnitude of correlational structure (DIM): one unidimensional scenario (all the attributes correlated at .5) and two bi-dimensional, varying the between-dimensions correlation (r = .5 so that each attribute correlates at .5 with the other attribute measuring the same dimension and .25 with the other attribute) or 0 so that each attribute only correlates at .5 with the other attribute measuring the same dimension. The levels for the data factors were chosen so that they were representative of the range of values that are encountered in applied settings. A summary of the simulation design can be found in Table 4.

Table 4.

Independent Variables According to the Research Design

Independent variables	Levels				
	L1	L2	L3	L4	L5
Data factors					
Generating model	DINA	*A*-CDM			
Test length	12	24			
Sample size	500	1,000			
Item quality	High item discrimination	Low item discrimination			
Dimensional magnitude of correlational structure	Unidimensional	Bidimensional (r =.5)	Bidimensional (r =.0)		
Method factors					
Model fitted	DINA	*A*-CDM	G-DINA		
Item fit statistic	S-X2	RMSEA	LR test	Wald test	LM test

Table 4 shows a 2×2×2×2×3 (MOD × J × N × IQ × DIM) between-subjects design that produces a total of 48 factor combinations. The number of attributes was fixed to $K = 4$. The *Q*-matrix used in simulating the response data and fitting the models is given in Table 5. Please note that the item fit statistics for relative fit are only necessary for items with more than one required attribute. For each condition, 200 data sets were generated and DINA, *A*-CDM and G-DINA models were fitted.

Type I error rate is computed as the proportion of times that we reject H_0 when the reduced model is the generating model. We did not simulated data under the G-DINA model. However, both reduced models are nested in the G-DINA model. That allows us to examine the power rate of the each reduced model when data is generated under the other reduced model. For example, in the case of the DINA model power rate is computed as the proportion of time that we fail to reject H_0 (e.g. DINA is the generating model) when the generating model is *A*-CDM. Type I error and power of the χ^2distributed item fit statistics (i.e., S-X2, LR, W, and LM) were investigated using .05 as significance level. The RMSEA item fit statistic is bounded below 0, with lower values indicating a better fit to the data. Kunina-Habernicht et al. (2009) suggest that RMSEA values lower than .10 and .05 are indicative of moderate and good fit to the data, respectively, so we assess the Type I error rate and power considering these values. A Type I error rate of .05 and a test power of at least .80 will be considered adequate.

Table 5.

Simulation study Q-Matrix for the J=12 condition. For the J=24 condition the number of item measuring K_j^* =1,2, and 3 attributes is doubled.

	Attribute			
Item	α_1	α_2	α_3	α_4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
5	1	1	0	0
6	1	0	1	0
7	1	0	0	1
8	0	1	1	0
9	1	1	1	0
10	1	1	0	1
11	1	0	1	1
12	0	1	1	1

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 that met the Type I error and power criteria (i.e. Type I error ≅ .05 and power > .80). Dependent variable in the ANOVAs was the Type I error rate associated to each statistical test for all items with the five data factors as between subjects factors. Due to large sample size, most effects were significant. For this reason, omega squared ($\widehat{\omega}^2$) measure of effect size was chosen to establish the impact of the independent variables. Since it is less biased, $\widehat{\omega}^2$ is preferable to others effect size measures as $\hat{\eta}^2$ (Fowler, 1985). It should be taken into account that it is possible to obtain a negative value of $\widehat{\omega}^2$ for F<1. According to Cohen (1988), values near to .01 represent small effects, .06 medium effects, and .14 or greater large effects. With regard to the interaction effects, a cut-off of $\widehat{\omega}^2$ > .06 was used to establish the most salient interactions. We also checked that the estimates of observed power in the ANOVA were greater than .80.

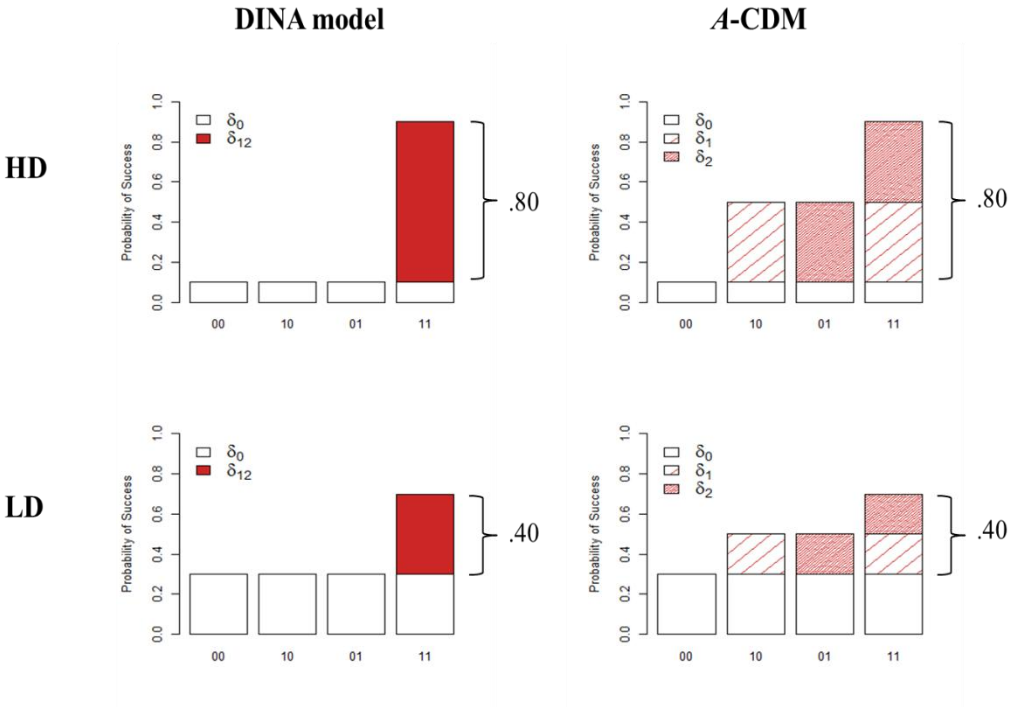


Figure 5. Representation of the generating model (DINA vs *A*-CDM) and item quality (high item discrimination vs low item discrimination) data factors for items measuring two attributes ($K_j^* = 2$).

A R (R Core Team, 2014) code was written to generate the item responses, calibrate the models, and estimate the item fit statistics. In doing so, the *gdina*, *gdina.wald*, *itemfit.sx2*,and *sim.gdina* included in the CDM package (Robitzsch et al., 2015) were employed. Another R code was written to compute the LM statistic as it has not been implemented yet for CDM.

Results

As will be exposed, the effect sizes for dimensional magnitude of correlational structure were not relevant for any of the statistics (i.e. $\widehat{\omega}^2 < .001$). Taking this into consideration and due to space limits, only results regarding the one of the levels of the factor correlation among the attributes are presented. The chosen level was the unidimensional scenario. The results in their entirety are shown in Annex 2 (type I error) and 3 (power).

Type I error. Type I error rates are shown in Table 6. When we employ .05 as cut-off point of RMSEA the Type I error reaches unacceptable levels, especially when the number of items is high (J = 24) and the items have a high discriminative power, reaching values of .813 and .878 for DINA and *A*-CDM models respectively. These values improves when the number of subjects is larger (N = 1000). When we consider .10 as cut-off point, the Type I error drops to approximately 0, with the only exception of the *A*-CDM model with a high number of highly discriminative items (J = 24) and a small sample size (N = 500), where the value is .112. Regarding the S-X2 statistic, the Type I error is between .037 and .096. With regard to LR test, Wald test, and LM test, Type I error rate is close to the nominal significance level when items are highly discriminative. On the contrary, values are much greater than the nominal value when the discriminative power is low, with the only exception of the values reached for LM test when the true model is the *A*-CDM. No much variability regarding the sample size is observed.

Table 6.

Type I Error of the item fit statistics (RMSEA, S-X2, LR, Wald, and LM) for the Two Reduced Models: unidimensional scenario.

			RMSEA				S-X2		LR		Wald		LM	
	Factors		cut-off point				cut-off point							
			.05		.10		.05							
Model	IQ	J	N											
			500	1000	500	1000	500	1000	500	1000	500	1000	500	1000
DINA	HD	12	.396	.030	.000	.000	.037	.040	.059	.054	.031	.052	.159	.093
		24	.813	.201	.006	.000	.054	.052	.071	.064	.033	.064	.158	.091
	LD	12	.019	.000	.000	.000	.052	.055	.343	.288	.449	.357	.180	.182
		24	.389	.025	.000	.000	.073	.062	.263	.170	.299	.172	.200	.184
A-CDM	HD	12	.558	.168	.006	.000	.085	.088	.136	.090	.017	.011	.029	.028
		24	.878	.526	.112	.000	.078	.074	.099	.086	.005	.004	.051	.032
	LD	12	.055	.000	.000	.000	.096	.081	.314	.292	.246	.221	.086	.042
		24	.164	.001	.000	.000	.083	.087	.295	.268	.250	.197	.021	.002

Note. LR = Likelihood ratio test; W = Wald test; LM = Lagrange multiplier test; IQ = Item quality; J = Test length; N = Sample size. HD = High item discrimination; LD = Low item discrimination. Shaded cells correspond to values in the [.02, .08] interval.

Power. Power rates are shown in Table 7. Both cutt-off points of RMSEA values leads to acceptable values when the generating model is *A*-CDM and the items are highly discriminative. Power rate does not increase with a large sample size (N = 1,000). Regarding the S-X2 statistic, we see that inacceptable values are found across all conditions, being the average power .31. Power values are greater than .80 only when the *A*-CDM is the generating model, the items are highly discriminative, and the sample size is high. With regard to LR test, Wald test, and LM test, power is generally 1.00 in the high discriminative items conditions. When items are not highly discriminative, values drop to an average value of .53 and .54 in the case of LR and Wald tests respectively. A higher power is obtained when the sample size is high (N = 1,000). Power rates for the LM test are close to 0 when the quality of the items is poor.

The results exposed above lead us not to consider RMSEA and S-X2 for further analysis, as the Type I error (e.g. RMSEA, .05 as cut-off point) and power (e.g. S-X2) are far from reaching acceptable values.

ANOVA results. $\widehat{\omega}^2$values associated to each effect are shown in Table 8. As noted above, the effect sizes for dimensionality were not relevant for any of the statistics $\widehat{\omega}^2{}_{\mathrm{DIM}}{}^2 < .001$ in all cases. None of the interactions had a salient effect (i.e. $\widehat{\omega}^2 < .06$). Regarding the main effects, item quality has a large effect on LR and Wald tests (. $\widehat{\omega}^2{}_{\mathrm{IQ}}$ = .295 and .527 respectively), but small on LM test ($\widehat{\omega}^2{}_{\mathrm{IQ}}$ = .025). Also notable was that Wald and LM attained a similar pattern in its performance across the simulated conditions. Besides item quality, both are affected by the sample size ($\widehat{\omega}^2{}_{\mathrm{N}}$ = .022 and .026 respectively), and the generating model. The effect of the generating model had a medium effect on the Wald test, but a large effect on the LM test ($\widehat{\omega}^2{}_{\mathrm{MOD}}$ = .059 and .316 respectively). In addition, test length has a small effect on the Wald test ($\widehat{\omega}^2{}_{\mathrm{J}}$ = .041). Marginal means for the main effects are given in Table 9. Type I error rate was well kept around .05 in the high item discrimination conditions. As described above, item discrimination has the most salient effect for LR and W test: when the item discrimination is low Type I error is much higher than the nominal level (.28, .27, and .11 for the LR, W, and LM tests respectively). This makes it difficult to interpret the marginal means for all the others factors, because conditions with high item discrimination and low item discrimination are mixed. That is why the marginal means are higher than .05 in almost all the cases. This is also true but to a lesser degree for LM test. Regarding sample size and test length, marginal means are more close to the nominal level as the sample size and the number of items increase. There were no relevant effect of the dimensional correlational structure: marginal means are almost equal across the levels of this factor. A different pattern of results for the statistics is obtained with respect to the effect of the generating model. In the case of the LR test, the marginal mean for the DINA model is more close to the nominal level than the one for the *A*-CDM (.16 vs .20), but the effect is small considering the effect size. In the case of W and LM tests, the effect is much salient and the marginal mean is closer to the nominal level in the case of *A*-CDM (.18 vs .12 and .16 vs. 04, respectively). The power observed for all the effects described above is equal to 1.000, except for the dimensional of correlational structure factor were the power observed is equal to .218, .860, and 363 for the LR, W, and LM tests respectively.

Table 7.

Power the item fit statistics (RMSEA, S-X2, LR, Wald, and LM) for the DINA Model (A-CDM generated data) and A-CDM (DINA generated data): *unidimensional scenario.*

Factors			RMSEA				S-X2		LR		Wald		LM	
			cut-off point				cut-off point							
			.05		.1		.05							
Model	IQ	J	N											
			500	1000	500	1000	500	1000	500	1000	500	1000	500	1000
DINA	HD	12	.715	.695	.303	.213	.112	.225	.985	1.000	.967	1.000	.909	.998
		24	.845	.695	.660	.658	.258	.405	.995	1.000	1.000	1.000	1.000	1.000
	LD	12	.023	.000	.000	.000	.065	.058	.491	.482	.489	.533	.253	.251
		24	.200	.015	.000	.000	.062	.063	.468	.576	.488	.525	.337	.412
A-CDM	HD	12	.954	.942	.644	.534	.710	.892	1.000	1.000	.999	1.000	.826	.899
		24	.999	1.000	.922	.895	.772	.938	1.000	1.000	.999	1.000	.954	.999
	LD	12	.045	.002	.000	.000	.083	.100	.324	.447	.380	.477	.086	.042
		24	.313	.078	.011	.004	.133	.142	.634	.845	.640	.810	.019	.014

Note. LR = Likelihood ratio test; W = Wald test; LM = Lagrange multiplier test; IQ = Item quality; J = Test length; N = Sample size. Shaded cells correspond to values over .80.

Table 8.

Anova Effect Sizes.

Effect Type /	Item fit statistic		
Data factors	LR	W	LM
Main Effects			
N	0.015	**0.022**	**0.026**
DIM	$\widehat{\omega}^2 < 0$	0.001	0.000
J	0.011	**0.041**	0.004
IQ	**0.295**	**0.527**	**0.025**
MOD	0.016	**0.059**	**0.316**
Two-Way Interactions			
N * DIM	$\widehat{\omega}^2 < 0$	$\widehat{\omega}^2 < 0$	$\widehat{\omega}^2 < 0$
N * J	$\widehat{\omega}^2 < 0$	0.000	0.000
DIM * J	0.005	0.037	0.001
N * IQ	0.001	0.001	0.002
DIM * IQ	$\widehat{\omega}^2 < 0$	$\widehat{\omega}^2 < 0$	0.000
J * IQ	$\widehat{\omega}^2 < 0$	0.001	0.002
N * MOD	0.000	0.000	0.001
DIM * MOD	0.005	0.037	0.003
J * MOD	0.002	0.022	0.004
IQ * MOD	$\widehat{\omega}^2 < 0$	0.007	0.029
Three-Way Interactions			
N * DIM * J	0.000	$\widehat{\omega}^2 < 0$	0.001
N * DIM * IQ	$\widehat{\omega}^2 < 0$	$\widehat{\omega}^2 < 0$	0.000
N * J * IQ	0.000	0.000	$\widehat{\omega}^2 < 0$
DIM * J * IQ	0.000	0.000	$\widehat{\omega}^2 < 0$
N *DIM * MOD	0.000	$\widehat{\omega}^2 < 0$	0.000
N * J * MOD	0.003	0.009	0.008
DIM * J * MOD	0.000	$\widehat{\omega}^2 < 0$	0.000
N * IQ * MOD	0.000	0.000	$\widehat{\omega}^2 < 0$
DIM * IQ * MOD	0.000	0.000	0.000
J * IQ * MOD	0.009	0.027	0.006
Four-Way Interactions			
N * DIM * J * IQ	0.000	0.000	$\widehat{\omega}^2 < 0$
N * DIM * J * MOD	0.001	$\widehat{\omega}^2 < 0$	$\widehat{\omega}^2 < 0$
N * DIM * IQ * MOD	0.000	0.000	0.000
N * J * IQ * MOD	0.001	0.000	0.001
DIM * J * IQ * MOD	$\widehat{\omega}^2 < 0$	0.001	0.000
Five-Way Interaction			
N * DIM * J * IQ * MOD	0.000	0.000	$\widehat{\omega}^2 < 0$

Note. LR = Likelihood ratio test; W = Wald test; LM = Lagrange multiplier test. MOD = Model, J = Test length; N = Sample size; IQ = Item quality; DIM = Dimensional magnitude of correlational structure. Main effects with η_p^2 values greater than .02 are shown in bold.

Table 9.

Marginal means Type I error rates.

Item fit statistic	*Data factor / Level* *N*		*DIM*			*J*		*IQ*		*MOD*	
	500	1000	Uni.	Bi. (r =.5)	Bi. (r =.0)	12	24	HD	LD	DINA	A-CDM
LR	.20	.16	.18	.19	.19	.20	.17	.08	.28	.16	.20
W	.17	.13	.15	.16	.15	.17	.13	.03	.27	.18	.12
LM	.11	.08	.10	.10	.10	.10	.09	.08	.11	.16	.04

Note. LR = Likelihood ratio test; W = Wald test; LM = Lagrange multiplier test. MOD = Model, J = Test length; N = Sample size; IQ = Item quality; DIM = Dimensional magnitude of correlational structure. Uni. = Unidimensional; Bi. = Bidimensional. HD: High item discrimination; LD: Low item discriminaiton.

Discussion and conclusions

The proper application of a statistical model requires the assessment of model-data fit. The process of model selection involves checking the model-data fit, which can be examined at the test, at the item or at the person level. While 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) and at the person-level (e.g. Liu et al., 2013; Cui & Leighton, 2009), in the field of cognitive diagnosis modeling, the statistical properties of the majority of the item fit statistics proposed (e.g. RMSEA, S-X2, LR test, and LM test) remains unknown or need further investigation (e.g. Wald test; de la Torre & Lee, 2013). Taking the above into account, this study focused on how model fit can be evaluated at item level.

Our study contributes to two domains. First, we evaluate the potential usefulness of some statistics proposed to assess absolute fit (RMSEA and S-X2) and relative fit (LR test, Wald test, and LM test) at the item level. In order to employ these statistics in practical use, it is necessary that its Type I error rates are close to the nominal value and that they have a great power to reject false models. Our findings show that, in general, RMSEA, regardless the cut-off point, and S-X2 do not reach an accurate power. This problem is compounded when the generating model is the DINA. Such a great limitation allows us not to consider these two statistics for practical use. With regard to the three procedures for assessing relative fit at the item level, when the quality of items is good accurate rates of Type I error and power are reached, even when the sample size is small (N = 500). Considering the LM test, there are two major limitations. First, Type I error rates are slightly higher when DINA model is the generating model. Second, when data were generated with the DINA model and items had not a high discriminative power, the LM test was not able to consistently flag each item as not conforming to the *A*-CDM model.

The overall message regarding the performance of these item fit statistics is that the empirical significance level of the LR test and the Wald test conforms closely to the nominal significance level. In addition, power comparisons also favour these two test over the LM test. The LR test was found to be more robust to the data factors than the Wald test. However, both tests are highly affected by the item discrimination. When items are not discriminative, the power rate tend to be low. It implies that we could not differentiate between DINA and *A*-CDM models in practical settings when the item discriminations is poor. The choice between the LR test and the Wald test has to do with its implementation requirements. The LR requires J_K + 1 models to be estimated, where J_K is the number of items measuring at least K = 2 attributes. A model where the unrestricted model (i.e. G-DINA) is applied to all items is compared with J_K models each of them having one item estimated under a reduced model. On the contrary, the Wald test requires only the unrestricted model to be estimated.

When presenting these conclusions, several important caveats are in order. First, we expected that the dimensional magnitude of the correlational structure to be a determining factor; however, its effect was negligible. A second caveat relates to the salient effects of the generating model. Alternative models also nested in the G-DINA, such as DINO (Templin & Henson, 2006) and the *reduced reparameterized unified model* (R-RUM, DiBello, Roussos, & Stout, 2007; Hart, 2002), could also be employed. In addition, it would be interesting to use different model combinations (e.g. a model derived from the combination of the DINA and the *A*-CDM ; de la Torre & Lee, 2013). Third, the Type I errors for items measuring a different number of attributes could be documented separately (i.e. $K_j^* = 2$ and $K_j^* = 3$) as it is done in de la Torre and Lee (2013). Fourth, in the *cognitively diagnostic educational assessment* field the number of attributes tends to be high. For example, Lee, Park, and Taylan (2011) applied the DINA model to the responses of 25 mathematics items included in the TIMSS 2007. Based on the 2007 TIMSS Framework for Fourth Grade Mathematics, they specifically constructed a *Q*-Matrix composed of 15 attributes. The same applies to the application of CDMs to the widely analyzed Tatsuoka's (1984) fraction subtraction data set (e.g. DeCarlo, 2011; Tatsuoka, 2002), where all test items are based on 8 attributes. It is recommendable that future research examine the effect of the number of attributes. Finally, all items were simulated to have the same discriminative power (i.e. high or low discriminative power). In a more realistic scenario, discriminative and non-discriminative items are mixed. Clearly, more research is needed along these lines.

References

Adam, R. J., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23. doi: 10.1177/0146621697211001

Bartram, D. (2005). The Great Eight Competencies: A criterion-centric approach to validation. *Journal of Applied Psychology, 90,* 1185-1203. doi:10.1037/0021-9010.90.6.1185

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange Multiplier tests: an expository note. *American Statistician, 36*, 153-157. doi: 10.1080/00031305.1982.10482817

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in Cognitive Diagnosis Modeling. *Journal of Educational Measurement, 50*(2), 123-140. doi:10.1111/j.1745-3984.2012.00185.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cui, Y., & Leighton, J.P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46*, 429-449. doi: 10.1111/j.1745-3984.2009.00091.x

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. doi: 10.1111/jedm.12022

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115-130. doi: 10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179-199. doi:10.1007/s11336-011-9207-7

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47*, 227-249. doi:10.1111/j.1745-3984.2010.00110.x

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA Model, classification, latent class sizes, and the Q-Matrix. *Applied Psychological Measurement*, *35*, 8–26. doi: 10.1177/0146621610377081

DiBello, L. V., Stout, W., & Roussos, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-390). Hillsdale, NJ: Lawrence Erlbaum.

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.

Embretson, S.E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika*, *64*, 407-433. doi:10.1037/a0014877

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). Elsevier. pp. 796–801. ISBN 978-0-444-86185-6.

Fischer, G.H. (1995b). The linear logistic test model. En G.H. Fischer & I.W. Molenaar (Eds.). *Rasch Models: Foundations, recent developments and aplications.* (pp.131-155). New York: Springer-Verlag.

Fowler, R.L. (1985). Point estimates and confidence intervals in measures of association. *Psychological Bulletin*, *98*, 160-165. doi: 10.1037/0033-2909.98.1.160

García, P. E., Olea, J., & de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema*, *3*, 372-377. doi:10.7334/psicothema2013.322

Haagenars, J., & McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.

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., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29*, 262–277. doi:10.1177/0146621604272623

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191-210. doi:10,1007: s11336-008-9089-5

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25,* 258-272. doi:10.1177/01466210122032064

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation, 35*, 64–70. doi:10.1016/j.stueduc.2009.10.003

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing*, *11*(2)*,* 144-177. doi:10.1080/15305058.2010.534571

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement,* *33*(8)*,* 579-598. doi: 10.1177/0146621609331960

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50-64. doi: 10.1177/01466216000241003

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. doi: 10.1177/0146621603027004004

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2015). CDM: Cognitive Diagnosis Modeling. R package version 4.2-12. 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. (2007). The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models. *International Journal of Testing*, *7*, 95–125. doi:10.1080/15305050701193454

Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response theory models. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment in education: Theory and practice* (pp. 205–241). Cambridge: Cambridge University Press.

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York: The Guilford Press.

Rupp. A.A., & Templin, J.L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-ofthe-art. *Measurement*, *6*, 219-262. doi:10.1080/15366360802490866

Sinharay, S., & Almond, R. G. (2007). Assessing Fit of Cognitively Diagnostic Models - A Case Study. *Educational and Psychological Measurement*, *67*(2), 239-257.

Sorrel, M.A., Olea, J., Abad, F.J., Aguado, D., & Lievens, F. (2015). Validity and reliability of Situational Judgement Test scores: A new approach through Cognitive Diagnosis Models. *Organizational Research Methods*. (Revised and Resubmitted)

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society, Series C, Applied Statistics*, *51*, 337–350.

Tatsuoka, K. (1984). *Analysis of errors in fraction addition and subtraction problems*. Final Report for NIE-G-81-0002, University of Illinois, Urbana-Champaign.

Tatsuoka, K.K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Templin J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305. doi:10.1037/1082-989X.11.3.287

Van der Linden, W. & Hambleton, R. (1997). *Handbook of modern item response theory*. New York: Springer.

von Davier, M. (2005). *A general diagnostic model applied to language testing data*. ETS Research Report.No. RR-05-16. Princeton, NJ: Educational Testing Service.

Weekley, J.A., Hawkes, B., Guenole, N. & Ployhart, R.E. (2015). Low-fidelity simulations. *Annual Review of Organizational Psychology and Organizational Behavior, 2*, 295-322. doi: 10.1146/annurev-orgpsych-032414-111304

Annex 1

Attribute descriptions based on test specifications.

Attribute	Definition	Typical behavioral patterns for people mastering the attribute in the educational environment
Study habits.	Study habits refers to the pattern of behavior adopted by students in the pursuit of their studies that serves as the vehicle of learning. It is the degree to which the student engages in regular acts of studying that are characterized by appropriate studying routines occurring in an environment that is conducive to studying.	Reviews of material, study every day, take practice tests, efficiently organize his/her work, etc.
Study attitudes.	Study attitudes refers to a student's positive attitude toward the specific act of studying and the student's acceptance and approval of the broader goals of education.	Think education is relevant to their future, persist with enthusiasm or effort, have a good opinion of their teachers, etc.
Helping others.	Helping others refers to voluntary actions that help another person with a problem. These helping behaviors can both be directed within or outside the organization.	Carry out volunteer actions that do not directly benefit them, share notes with their peers, help peers who are in troubles, etc.
Generalized compliance.	Generalized compliance refers to following rules and procedures, complying with organizational values and policies, conscientiousness, and meeting deadlines.	Stick with the existing timetable, be always punctual, do not defy the teacher, etc.

Annex 2

Type I error of the item fit statistics (RMSEA, S-X2, LR, Wald, and LM) for the two reduced models.

				RMSEA				S-X2		LR		Wald		LM	
Factors				RMSEA cut-off point				chi-square distribution cut-off point							
				0.05		0.10		0.05							
Model	IQ	J	Dim.	*N*											
				500	1000	500	1000	500	1000	500	1000	500	1000	500	1000
DINA	HD	12	Unidimensional	.396	.030	.000	.000	.037	.04	.059	.054	.031	.052	.159	.093
			.5	.403	.030	.000	.000	.049	.042	.064	.058	.031	.064	.146	.097
			0	.387	.030	.000	.000	.048	.044	.072	.062	.036	.063	.184	.102
		24	Unidimensional	.813	.201	.006	.000	.054	.052	.071	.064	.033	.064	.158	.091
			.5	.793	.205	.005	.000	.045	.053	.066	.058	.039	.058	.139	.092
			0	.797	.226	.006	.000	.062	.053	.07	.062	.038	.053	.166	.116
	LD	12	Unidimensional	.019	.000	.000	.000	.052	.055	.343	.288	.449	.357	.18	.182
			.5	.025	.000	.000	.000	.046	.047	.401	.258	.482	.349	.205	.182
			0	.030	.001	.000	.000	.055	.052	.339	.257	.428	.318	.197	.173
		24	Unidimensional	.389	.025	.000	.000	.073	.062	.263	.17	.299	.172	.2	.184
			.5	.380	.034	.000	.000	.064	.064	.245	.174	.287	.169	.202	.175
			0	.433	.034	.000	.000	.071	.061	.251	.185	.27	.174	.199	.185
A-CDM	HD	12	Unidimensional	.558	.168	.006	.000	.085	.088	.136	.09	.017	.011	.029	.028
			.5	.564	.167	.006	.000	.1	.094	.131	.099	.011	.011	.037	.033
			0	.575	.150	.007	.000	.122	.094	.131	.104	.009	.006	.051	.033
		24	Unidimensional	.878	.526	.112	.000	.078	.074	.099	.086	.005	.004	.051	.032
			.5	.880	.510	.119	.001	.093	.074	.101	.082	.007	.003	.04	.031
			0	.882	.502	.113	.001	.114	.087	.099	.078	.006	.004	.035	.035
	LD	12	Unidimensional	.055	.000	.000	.000	.096	.081	.314	.292	.246	.221	.086	.042
			.5	.062	.002	.000	.000	.099	.084	.317	.299	.27	.233	.083	.042
			0	.052	.001	.000	.000	.097	.083	.322	.306	.271	.226	.071	.024
		24	Unidimensional	.164	.001	.000	.000	.083	.087	.295	.268	.25	.197	.021	.002
			.5	.172	.002	.000	.000	.088	.082	.318	.298	.261	.211	.023	.001
			0	.192	.001	.000	.000	.092	.095	.365	.256	.258	.172	.015	.002

Note. LR = Likelihood ratio test; W = Wald test; LM = Lagrange multiplier test; IQ = Item quality; J = Test length; Dim. = Dimensional magnitude of correlational structure; N = Sample size. HD = High item discrimination; LD = Low item discrimination. Shaded cells correspond to values in the [.02, .08] interval.

Annex 3

Power of the item fit statistics (RMSEA, S-X2, LR, Wald, and LM) for the DINA model (A-CDM generated data) and A-CDM (DINA generated data).

				RMSEA				S-X2		LR		Wald		LM	
Factors				RMSEA cut-off point				chi-square distribution cut-off point							
				0.05		0.10		0.05							
				N											
Model	IQ	J	Dim.	500	1000	500	1000	500	1000	500	1000	500	1000	500	1000
DINA	HD	12	Unidimensional	.715	.695	.303	.213	.112	.225	.985	1.000	.967	1.000	.909	.998
			.5	.705	.670	.323	.218	.082	.140	.988	1.000	.964	1.000	.914	.998
			0	.703	.659	.271	.165	.083	.099	.974	1.000	.937	.998	.926	.996
		24	Unidimensional	.845	.695	.660	.658	.258	.405	.995	1.000	1.000	1.000	1.000	1.000
			.5	.831	.690	.661	.655	.194	.319	1.000	1.000	.999	1.000	.999	1.000
			0	.825	.682	.641	.654	.106	.249	.990	1.000	.969	1.000	.986	1.000
	LD	12	Unidimensional	.023	.000	.000	.000	.065	.058	.491	.482	.489	.533	.253	.251
			.5	.027	.000	.000	.000	.063	.058	.468	.481	.474	.506	.259	.276
			0	.028	.000	.000	.000	.073	.056	.462	.455	.438	.468	.264	.271
		24	Unidimensional	.200	.015	.000	.000	.062	.063	.468	.576	.488	.525	.337	.412
			.5	.247	.019	.000	.000	.064	.060	.472	.542	.484	.495	.359	.414
			0	.260	.020	.000	.000	.064	.058	.489	.487	.490	.455	.381	.442
A-CDM	HD	12	Unidimensional	.954	.942	.644	.534	.710	.892	1.000	1.000	.999	1.000	.826	.899
			.5	.948	.953	.592	.618	.663	.885	1.000	1.000	.997	1.000	.812	.963
			0	.923	.934	.680	.654	.677	.877	1.000	1.000	.998	1.000	.879	.969
		24	Unidimensional	.999	1.000	.922	.895	.772	.938	1.000	1.000	.999	1.000	.954	.999
			.5	.999	.999	.912	.852	.716	.877	1.000	1.000	1.000	1.000	.947	.999
			0	.998	.998	.856	.850	.641	.869	1.000	1.000	.999	1.000	.797	.947
	LD	12	Unidimensional	.045	.002	.000	.000	.083	.100	.324	.447	.380	.477	.086	.042
			.5	.050	.004	.000	.000	.081	.091	.387	.485	.432	.544	.053	.029
			0	.036	.007	.000	.000	.095	.108	.393	.477	.435	.533	.042	.013
		24	Unidimensional	.313	.078	.011	.004	.133	.142	.634	.845	.640	.810	.019	.014
			.5	.314	.161	.006	.003	.112	.158	.586	.839	.625	.813	.017	.008
			0	.407	.177	.008	.001	.120	.148	.597	.806	.618	.793	.004	.002

Note. LR = Likelihood ratio test; W = Wald test; LM = Lagrange multiplier test; IQ = Item quality; J = Test length; Dim. = Dimensional magnitude of correlational structure; N = Sample size; HD = High item discrimination; LD = Low item discrimination. Shaded cells correspond to values over .80.