



Universidad Autónoma  
de Madrid

**Biblos-e Archivo**  
Repositorio Institucional UAM

**Repositorio Institucional de la Universidad Autónoma de Madrid**

<https://repositorio.uam.es>

Esta es la **versión de autor** del artículo publicado en:  
This is an **author produced version** of a paper published in:

Journal Of Educational And Behavioral Statistics 48.6 (2023): 719-749

**DOI:** <https://doi.org/10.3102/10769986231158829>

**Copyright:** © 2023 AERA

El acceso a la versión del editor puede requerir la suscripción del recurso

Access to the published version may require subscription

“This is the accepted version of the article may be posted in the author's  
institutional repository and reuse is restricted to non-commercial and no derivative  
uses”

**The Restricted DINA Model: A Comprehensive Cognitive Diagnostic Model  
for Classroom-Level Assessments**

Pablo Nájera<sup>1</sup>, Francisco J. Abad<sup>1</sup>, Chia-Yi Chiu<sup>2</sup>, and Miguel A. Sorrel<sup>1</sup>

<sup>1</sup> Department of Social Psychology and Methodology, Universidad Autónoma de Madrid

<sup>2</sup> Educational Psychology, University of Minnesota Twin Cities

The official citation that should be used for this material is:

Nájera, P., Abad, F. J., Chiu, C.-Y., & Sorrel, M. A. (2023). The restricted DINA model: A comprehensive cognitive diagnostic model for classroom-level assessments. *Journal of Educational and Behavioral Statistics*. Advance online publication.

<https://doi.org/10.3102/10769986231158829>

This paper is not the copy of record and may not exactly replicate the authoritative document published in the journal. The final article is available, upon publication, at

<https://doi.org/10.3102/10769986231158829>

**Author Note**

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

Francisco J. Abad  <https://orcid.org/0000-0001-6728-2709>

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

This research was partially supported by Ministerio de Ciencia, Innovación y Universidades of Spain (Grant PSI2017-85022-P), European Social Fund, the Community of Madrid in Spain (Grant SI3/PJI/2021-00258), and Cátedra de Modelos y Aplicaciones Psicométricas (Instituto de Ingeniería del Conocimiento and Universidad Autónoma de Madrid). Portions of these findings were presented at the XVII Congreso de Metodología de las Ciencias Sociales y de la Salud (July 2022; Teruel, Spain). The simulation codes of this research are publicly available at <https://osf.io/deps9/>. This study was not preregistered. We have no conflict of interest to disclose.

Corresponding concerning this article should be addressed to Pablo Nájera, Faculty of Psychology, 6 Iván Pavlov St, Cantoblanco Campus, Madrid, Spain, 28049. Email: [pablo.najera@uam.es](mailto:pablo.najera@uam.es)

### **Abstract**

The nonparametric classification (NPC) method has been proven to be a suitable procedure for cognitive diagnostic assessments at a classroom level. However, its nonparametric nature impedes the obtention of a model likelihood, hindering the exploration of crucial psychometric aspects such as model fit or reliability. Reporting the reliability and validity of scores is imperative in any applied context. The present study proposes the restricted DINA (R-DINA) model, a parametric cognitive diagnosis model based on the NPC method that provides the same attribute profile classifications as the nonparametric method while allowing to derive a model likelihood and, subsequently, to compute fit and reliability indices. The suitability of the new proposal is examined by means of an exhaustive simulation study and a real data illustration. The results show that the R-DINA model properly recovers the posterior probabilities of attribute mastery, thus becoming a suitable alternative for comprehensive small-scale diagnostic assessments.

*Keywords:* cognitive diagnosis, nonparametric classification, DINA model, classification accuracy, relative fit.

**The Restricted DINA Model: A Comprehensive Cognitive Diagnostic Model  
for Classroom-Level Assessments**

Cognitive diagnostic assessments (CDAs) have received increasing attention in educational research and practice due to the detailed information they provide regarding students' mastery or non-mastery of a series of attributes, be it skills, cognitive processes, or competences. In contrast with the more traditional summative assessment, which rank-orders students based on a single continuous score, CDAs are particularly useful in school settings, where teachers can use the diagnostic information to better guide remedial instruction ([de la Torre & Minchen, 2014](#); [Paulsen & Svetina, 2021](#)).

CDAs rely on cognitive diagnostic models (CDMs) to estimate the students' attribute mastery profile. CDMs are restricted latent class models, that is, multidimensional models in which the latent variables (i.e., attributes) are discrete, usually dichotomous. One of the main inputs of these models is the so-called Q-matrix ([Tatsuoka, 1983](#)), a binary specification matrix of dimensions  $J$  items  $\times$   $K$  attributes in which each q-entry takes a value of  $q_{jk} = 1$  or 0 depending on whether item  $j$  measures attribute  $k$  or not, respectively.

As a family of statistical models, CDMs can be organized based on different criteria. Firstly, it can be distinguished between parametric and nonparametric CDMs. The former are stochastic models that, under some specific assumptions, provide consistent estimates of both item and person parameters via marginalized maximum likelihood estimation (e.g., [de la Torre, 2009](#); [de la Torre, 2011](#)) or Markov chain Monte Carlo algorithms (e.g., [C.-W. Liu et al., 2020](#); [Xu et al., 2020](#)). On the other hand, nonparametric CDMs are deterministic methods that classify students without relying on model parameter estimation. Instead, as will be explained below, classifications are done by comparing observed response patterns with ideal response patterns.

Thus, these likelihood-free procedures are usually more parsimonious and computationally efficient (Chiu & Douglas, 2013). Both parametric and nonparametric attribute profile estimators are related: previous work indicates that the likelihood of parametric models is maximized when discrepancies between observed and ideal patterns are minimized (Chiu et al., 2018).

Secondly, CDMs can be divided into reduced and general models. The defining feature of reduced models is that they entail a specific item response process. For instance, conjunctive models imply that an examinee must master all the attributes involved in an item to correctly answer it. A widely-used parametric conjunctive CDM is the *deterministic input, noisy “and” gate* (DINA) model (Junker & Sijtsma, 2001), whose item response function is expressed as

$$P_j(\alpha_l) = g_j^{1-\eta_{lj}^{(c)}} (1 - s_j)^{\eta_{lj}^{(c)}}, \quad (1)$$

where

$$\eta_{lj}^{(c)} = \prod_{k=1}^K \alpha_{lk}^{q_{jk}} \quad (2)$$

is the conjunctive ideal response,  $\alpha_{lk}$  is attribute  $k$  for latent class  $l$ ,  $q_{jk}$  is the  $q$ -entry concerning item  $j$  and attribute  $k$ ,  $g_j$  is the guessing parameter (i.e., probability of correctly answering item  $j$  when  $\eta_{lj}^{(c)} = 0$ ), and  $s_j$  is the slip parameter (i.e., probability of incorrectly answering item  $j$  when  $\eta_{lj}^{(c)} = 1$ ). Thus, the conjunctive ideal response,  $\eta_{lj}^{(c)}$ , will take a value of 1 or 0 depending on whether latent class  $l$  masters all the required attributes or not, respectively. On the opposite side, disjunctive models imply that it is enough to master only one of the required attributes to correctly answer the item. The *deterministic input, noisy “or” gate* (DINO) model (Templin & Henson, 2006) is a popular disjunctive parametric CDM. The item response function of the DINO model is the same as the one in Equation 1, with the important difference that the ideal response patterns are disjunctive:

$$\eta_{lj}^{(d)} = 1 - \prod_{k=1}^K (1 - \alpha_{lk})^{q_{jk}}, \quad (3)$$

which takes a value of 1 or 0 depending on whether latent class  $l$  masters at least one of the required attributes or nor, respectively. Beyond parametric models, the *nonparametric classification* (NPC) method (Chiu & Douglas, 2013) can accommodate either conjunctive or disjunctive responses.

General models, by contrast, are not restricted to a specific response process. They are saturated models in the sense that a different probability of success is estimated for each latent class. The *general diagnostic model* (GDM; von Davier, 2008), the *log-linear CDM* (Henson et al., 2009), and the *generalized DINA* (G-DINA) model (de la Torre, 2011) are examples of parametric general models. In the latter, the probability of correctly answering item  $j$  is modelled as the sum of all main and interaction effects involving the required attributes:

$$P_j(\alpha_l) = \delta_{j0} + \sum_{k=1}^{K_j} \delta_{jk} \alpha_{lk} + \sum_{k'=1}^{K_j} \sum_{k=1}^{K_j-1} \delta_{jkk'} \alpha_{lk} \alpha_{lk'} + \delta_{j12\dots K_j} \prod_{k=1}^{K_j} \alpha_{lk}, \quad (4)$$

where  $\alpha_l$  is the attribute profile of latent class  $l$ ,  $K_j$  is the number of attributes involved in item  $j$ ,  $\delta_{j0}$  is the intercept for item  $j$ ,  $\delta_{jk}$  is the main effect due to  $\alpha_k$ ,  $\delta_{jkk'}$  is the interaction effect due to  $\alpha_k$  and  $\alpha_{k'}$ , and  $\delta_{j12\dots K_j}$  is the interaction effect due to  $\alpha_1, \alpha_2, \dots, \alpha_{K_j}$ . It should be noted that most reduced CDMs are subsumed in the general models. For instance, the DINA model can be easily derived by removing all the effects in Equation 4 except for the intercept ( $\delta_{j0}$ ) and the highest-order interaction ( $\delta_{j12\dots K_j}$ ; de la Torre, 2011). The G-DINA model has a non-parametric counterpart: the *general NPC* (GNPC) method (Chiu et al., 2018), which extends the NPC method by combining both the conjunctive and disjunctive response processes, thus accommodating more complex data generation processes.

Selecting one CDM among the different possibilities (e.g., reduced or general, parametric or nonparametric) for a particular application might seem a difficult task. Accordingly, a number of studies have been recently conducted with the aim of shedding some light on this question (Chiu et al., 2018; C. Ma et al., 2022; W. Ma & Jiang, 2021; Oka & Okada, 2021; Paulsen & Svetina, 2021; Sen & Cohen, 2021; Sorrel et al., 2021). As it would be expected in any kind of statistical model, these studies consistently found that the greater the complexity of the CDM, the more challenging the estimation process. Generally, parametric CDMs demand large samples ( $N > 500$ ) and high-quality items (i.e., with power to discriminate between latent classes) to obtain accurate parameter estimates (C. Ma et al., 2022; W. Ma & Jiang, 2021; Oka & Okada, 2021; Sen & Cohen, 2021; Sorrel et al., 2021). This is especially true for general models, which, in addition, are greatly impacted by the complexity of the Q-matrix (i.e., the number of attributes being measured by the items), since the number of item parameters exponentially increases as the Q-matrix becomes more complex (Sorrel et al., 2021). In this line, fitting a general CDM to data generated by a reduced model is suboptimal in terms of parsimony: it might lead to less accurate item parameter estimates and attribute profile classifications compared to fitting the correct reduced model (Sorrel et al., 2021). On another note, even though small sample sizes usually lead to a large item parameter estimation bias, some studies have found that the impact on the attribute profile classification accuracy is not as relevant (C. Ma et al., 2022; Paulsen & Svetina, 2021; Sen & Cohen, 2021). This might be explained by the fact that, even though item parameter estimates might be biased, the correct attribute profile will lead to the maximum likelihood as long as the estimates point in the right direction (e.g., an estimated probability of success higher than 0.5 for an examinee that masters the required attributes; C. Ma et al., 2022; S. Wang & Douglas, 2015).

The likelihood-free nature of nonparametric CDMs makes them remarkably robust to those conditions that hinder parameter estimation: small sample sizes, low-quality items, and complex Q-matrices. In these situations, both the NPC and the GNPC method outperform their parametric counterparts (i.e., DINA/DINO, G-DINA) in terms of classification accuracy ([Chiu et al., 2018](#); [C. Ma et al., 2022](#); [Oka & Okada, 2021](#)). These results highlight the suitability of nonparametric CDMs for classroom-level educational assessments, where non-ideal conditions are expected (e.g., very small sample size). However, nonparametric methods suffer from a nontrivial drawback: because they do not contemplate the computation of likelihoods, it is not possible to derive fit indices nor classification accuracy measures from them. Thus, the application of nonparametric CDMs in real settings cannot fulfil the conventional and fundamental psychometric criteria that concerns the interpretation and reporting of test scores ([American Educational Research Association \[AERA\] et al., 2014](#)).

In light of the above, practitioners that want to implement CDAs in classroom-level settings must commit, as of today, to either of two suboptimal approaches: (a) fitting a parametric CDM to their data, fully aware that the challenging conditions (e.g.,  $N < 50$ ) will likely lead to biased parameter estimates, or (b) applying a nonparametric CDM to their data and faithfully believe that the interpretation of the scores (e.g., reliability, fit) is valid. To address this issue, the present paper proposes the restricted DINA (R-DINA) model, a parametric CDM based on the NPC method that, while retaining its accurate attribute profile classifications and desirable properties (i.e., parsimony, robustness to challenging conditions), allows for the computation of likelihoods and, subsequently, of posterior probabilities, fit indices, and classification accuracy measures.



The remaining of the paper is laid out as follows. First, the NPC method is described. Second, the relationship between the likelihood, posterior probabilities, fit indices, and classification accuracy measures is discussed. Third, the rationale and definition of the R-DINA model is presented. Fourth, the performance of the R-DINA model is systematically evaluated by means of an exhaustive simulation study. Fifth, two real datasets are used to illustrate the proposed procedure. Finally, practical implications, limitations, and future research lines are discussed.

## The Nonparametric Classification Method

The NPC method was developed by [Chiu and Douglas \(2013\)](#) as a parsimonious and computationally efficient cognitive diagnostic procedure that could obtain accurate classifications without relying on item parameter estimation. The rationale of the NPC method consists in comparing an examinee's observed response pattern ( $\mathbf{y}_i$ ) with the ideal response patterns from all possible latent classes ( $\boldsymbol{\eta}_l$ ). As explained above, these ideal responses can be either conjunctive ( $\boldsymbol{\eta}_l^{(c)}$ ; see [Equation 2](#)) or disjunctive ( $\boldsymbol{\eta}_l^{(d)}$ ; see [Equation 3](#)), and it is the researcher's labor to predetermine which response process is more appropriate for their data. Note that the calculation of the ideal response pattern for all latent classes is straightforward based on a selected response process (i.e., conjunctive or disjunctive) and a given Q-matrix. Then, the distance between the ideal response patterns and the observed response patterns is computed. The simplest distance measure is the Hamming distance, which refers to the number of discrepancies between both response vectors:

$$d_h(\mathbf{y}_i, \boldsymbol{\eta}_l) = \sum_{j=1}^J |y_{ij} - \eta_{lj}|, \quad (5)$$

where  $\eta_{lj}$  can be either  $\eta_{lj}^{(c)}$  or  $\eta_{lj}^{(d)}$ . Based on this information, the NPC's attribute profile estimate for examinee  $i$  is the one that minimizes the Hamming distance:

$$\hat{\alpha}_i = \arg \min_l d_h(\mathbf{y}_i, \boldsymbol{\eta}_l). \quad (6)$$

S. Wang and Douglas (2015) showed that Equation 6 is a consistent estimator as test length goes to infinity ( $J \rightarrow \infty$ ) under the assumptions of conditional independence, item discriminatory power (i.e.,  $g_j < 0.5$  and  $s_j < 0.5$ ), and Q-matrix completeness. A Q-matrix is complete when the ideal responses of all attribute profiles are distinct; introducing an identity matrix in the Q-matrix (i.e., there is at least one item measuring each attribute in isolation) is a sufficient condition for its completeness (Chiu et al., 2009). It should be noted that the NPC procedure is independent from the sample size, since it does not rely on any parameter estimate. Also, it is important to note that, even though the NPC method's estimator is consistent whenever  $g_j < 0.5$ ,  $s_j < 0.5$  and  $J \rightarrow \infty$  (S. Wang & Douglas, 2015), it adopts  $g_j = s_j \forall j$ , that is, the method operates as if there were the same probability of guessing and slipping all items. The reason for this is embedded in the rationale of the Hamming distances, which count the number of discrepancies between the observed and ideal responses, giving the same weight to both types of discrepancies (i.e.,  $y_{ij} = 1$  and  $\eta_{lj} = 0$ ;  $y_{ij} = 0$  and  $\eta_{lj} = 1$ ) and to all items, regardless of their difficulty or discrimination. Despite this, the NPC method has shown a satisfactory robustness to challenging conditions such as 20% of misspecified q-entries, dissimilar guessing, and slip parameters generated from a  $U(0, 0.5)$ , or even data generated from the G-DINA model (Chiu et al., 2018; Chiu & Douglas, 2013). The importance of  $g_j = s_j \forall j$  being truthful will decrease as  $J \rightarrow \infty$ , where the NPC method's estimator will be consistent under the softer condition of  $g_j < 0.5$  and  $s_j < 0.5$ .

### Relative Fit and Estimated Classification Accuracy

In parametric CDMs, different fit and classification accuracy indices can be derived from the likelihoods and posterior probabilities. Namely, the likelihood of observing a response pattern  $\mathbf{y}_i$  given attribute profile  $\boldsymbol{\alpha}_l$  is computed as

$$\mathcal{L}(\mathbf{y}_i|\boldsymbol{\alpha}_l) = \prod_{j=1}^J P_j(\boldsymbol{\alpha}_l)^{y_{ij}} [1 - P_j(\boldsymbol{\alpha}_l)]^{1-y_{ij}}, \quad (7)$$

and the likelihood of the model is then obtained as

$$\mathcal{L} = \sum_{i=1}^N \sum_{l=1}^L \mathcal{L}(\mathbf{y}_i|\boldsymbol{\alpha}_l) p(\boldsymbol{\alpha}_l), \quad (8)$$

where  $N$  is the sample size,  $L$  is the total number of possible latent classes, and  $p(\boldsymbol{\alpha}_l)$  is the prior distribution of attribute profiles. Note that, since we are focusing on dichotomous attributes in the current paper,  $L = 2^K$ . From here, several well-known relative fit indices can be computed, such as the Akaike information criterion (AIC; [Akaike, 1974](#)), Bayesian information criterion (BIC; [Schwarz, 1978](#)), consistent AIC (CAIC; [Bozdogan, 1987](#)), or sample size-adjusted BIC (SABIC; [Sclove, 1987](#)):

$$AIC = -2 \ln \mathcal{L} + 2p, \quad (9)$$

$$BIC = -2 \ln \mathcal{L} + p \ln N, \quad (10)$$

$$CAIC = -2 \ln \mathcal{L} + p[1 + \ln N], \quad (11)$$

and

$$SABIC = -2 \ln \mathcal{L} + p \ln[(N + 2)/24], \quad (12)$$

where  $p$  is the number of model parameters. In the CDM context, the AIC, BIC, and CAIC indices have been shown to adequately perform at selecting the generating model under the presence of response process or Q-matrix misspecifications ([Chen et al., 2013](#); [Gao et al., 2021](#)).

The posterior probability of examinee  $i$  belonging to latent class  $l$  can be derived from the likelihood:

$$P(\alpha_l | \mathbf{y}_i) = \frac{\mathcal{L}(\mathbf{y}_i | \alpha_l) p(\alpha_l)}{\sum_{l=1}^L \mathcal{L}(\mathbf{y}_i | \alpha_l) p(\alpha_l)}. \quad (13)$$

Marginal probabilities of mastering each of the attributes independently can be obtained from the posterior probabilities:

$$P(\alpha_k | \mathbf{y}_i) = \sum_{l=1}^L P(\alpha_l | \mathbf{y}_i) \alpha_{lk}. \quad (14)$$

In parametric CDM, attribute profile estimates are usually made either by using the maximum a posteriori (MAP) estimation, which corresponds to the attribute profile with the maximum posterior probability, or by the expected a posterior (EAP) estimation, which consists in dichotomizing the marginal probabilities (usually with a .50 cutoff). Moreover, since posterior probabilities are a measure of certainty about attribute profile classifications, they are often used for the computation of many reliability indices. The interested reader is referred to [Johnson and Sinharay \(2018, 2020\)](#) for a comprehensive review of different classification accuracy and consistency measures in the CDM framework. One common classification accuracy measure is the  $\tau$  index ([W. Wang et al., 2015](#)), which is an estimation of the proportion of attribute profiles that are correctly classified:

$$\tau = \frac{\sum_{i=1}^N \sum_{l=1}^L P(\alpha_l | \mathbf{y}_i) I(\hat{\alpha}_i = \alpha_l)}{N}, \quad (15)$$

where  $I(\cdot)$  is the indicator function. It follows from this equation that an accurate estimate of the classification accuracy depends on an accurate estimate of the posterior probabilities. This, in turn, depends on the calculation of the likelihoods, which are based on the item parameter estimates. Thus, the  $\tau$  index (as well as any other reliability measure based on posterior

probabilities) is expected to obtain biased classification accuracy estimations under those conditions that hinder item parameter estimation (e.g., small sample sizes, low-quality items, complex Q-matrices). For instance, a scarce number of examinees in certain latent classes, which is likely to occur under these challenging scenarios, might lead to extreme probability of success estimations (i.e., close to 0 or 1; [Chiu et al., 2018](#); [W. Ma & Jiang, 2021](#)). These will lead to very sharpened posterior probabilities, thus resulting in a false sensation of certainty as indicated by an overestimated  $\tau$  index ([Kreitchmann et al., 2022](#)).

### **The restricted DINA model**

In the present paper we propose the restricted DINA (R-DINA) model, a parametric CDM based on the NPC method that allows to derive the model likelihood from the nonparametric classifications and, consequently, to compute relative fit indices and classification accuracy measures. As will be shown below, the R-DINA establishes a bridge between parametric (i.e., DINA model) and nonparametric (i.e., NPC method) cognitive diagnosis by holding the following properties: (a) it is aligned with the rationale of the NPC method, (b) it provides the same exact attribute profile classifications as the NPC method, and (c) it allows to derive the model likelihoods and posterior probabilities from such classifications.

Namely, the R-DINA model is built upon the same rationale of the Hamming distances which, as stated above, operates giving the same weight to all discrepancies between the observed and ideal responses, regardless of the type of mistake (guess or slip) and the items' discrimination. Following this, we reformulate those equally weighted discrepancies into a single parameter  $\varphi$  that represents the probability of providing an observed response (i.e.,  $y_{lj}$ ) different from the ideal one (i.e.,  $\eta_{lj}$ ). Making a connection with the DINA model, this parameter can be

understood as  $\varphi = g_j = s_j \forall j$ . Similar to Equation 1, the item response function for the R-DINA model is defined as

$$P_j(\alpha_l) = \varphi^{1-\eta_{lj}}(1 - \varphi)^{\eta_{lj}}, \quad (16)$$

where  $\eta_{lj}$  is the ideal response for latent class  $l$  in item  $j$ . Note that the ideal response can be either conjunctive ( $\eta_{ij}^{(c)}$ ; see Equation 2) or disjunctive ( $\eta_{ij}^{(d)}$ ; see Equation 3). Thus, the R-DINA terminology is reserved to the conjunctive case, while the restricted DINO (R-DINO) model can be used to refer to the disjunctive one. Even though we refer to the R-DINA model throughout the manuscript for the sake of simplicity, all explanations regarding the R-DINA model are extensible to the R-DINO model given the duality between the DINA and DINO models (Köhn & Chiu, 2016). From here, likelihoods can be computed as in

$$\mathcal{L}(\mathbf{y}_i | \alpha_l) = \prod_{j=1}^J [\varphi^{1-\eta_{lj}}(1 - \varphi)^{\eta_{lj}}]^{y_{ij}} \{1 - [\varphi^{1-\eta_{lj}}(1 - \varphi)^{\eta_{lj}}]\}^{1-y_{ij}}. \quad (17)$$

Since both the observed and ideal responses are dichotomous, it is straightforward to derive from Equation 17 that the term inside the product will be equal to  $1 - \varphi$  whenever  $\eta_{lj} = y_{ij}$ , and equal to  $\varphi$  whenever  $\eta_{lj} \neq y_{ij}$ . Recalling that the Hamming distance,  $d_h(\mathbf{y}_i, \boldsymbol{\eta}_l)$ , counts the number of discrepancies between the observed and ideal responses (i.e.,  $\eta_{lj} \neq y_{ij}$ ) throughout a test of length  $J$ , the number of agreements is equal to  $J - d_h(\mathbf{y}_i, \boldsymbol{\eta}_l)$ . And, because all discrepancies (or agreements) lead to the same probability of success for all items, Equation 17 can be reformulated as

$$\mathcal{L}(\mathbf{y}_i | \alpha_l) = \varphi^{d_h(\mathbf{y}_i, \boldsymbol{\eta}_l)}(1 - \varphi)^{J-d_h(\mathbf{y}_i, \boldsymbol{\eta}_l)}. \quad (18)$$

Thus, by means of the model parameter  $\varphi$ , a correspondence between the parametric likelihood and the NPC's Hamming distances is established. From here, the computation of posterior probabilities, relative fit indices, and classification accuracy measures is straightforward. The  $\varphi$

parameter can be estimated via marginalized maximum likelihood using the expectation-maximization algorithm. Namely, following the notation of [de la Torre \(2009\)](#), the estimator of  $\varphi$  in the M-step is

$$\hat{\varphi} = \frac{\sum_{j=1}^J [R_j^{(0)} + I_j^{(1)} - R_j^{(1)}]}{\sum_{j=1}^J [I_j^{(0)} + I_j^{(1)}]}, \quad (19)$$

where  $I_j^{(0)}$  denotes the expected number of examinees with  $\eta_{lj} = 0$ ,  $R_j^{(0)}$  is the expected number of correct responses among  $I_j^{(0)}$ , and  $I_j^{(1)}$  and  $R_j^{(1)}$  have an equivalent interpretation but for those examinees with  $\eta_{lj} = 1$ . However, taking advantage of the simplicity of [Equation 18](#), where  $\varphi$  is the only unknown quantity, a numerical method can be used to estimate the  $\varphi$  parameter much faster. In the present paper we use the Brent's method ([Brent, 2002](#)) to estimate  $\hat{\varphi}$ .

A few additional considerations should be noted regarding the relationship between the R-DINA model and the NPC method. First, as implied by [Equation 18](#), there is a univocal inverse exponential relation between Hamming distances and model likelihoods. It stems from here that, for a specific examinee, the latent class with the minimum Hamming distance will be also the one with the highest likelihood. Accordingly, the attribute profile classifications from the R-DINA model will be exactly the same as those from the NPC method whenever maximum likelihood or MAP estimation with a uniform prior distribution is used for the former. As stated before, these classifications have been shown to outperform those from the DINA model under a wide range of conditions, particularly with small sample sizes ([Chiu et al., 2018](#); [Chiu & Douglas, 2013](#); [C. Ma et al., 2022](#); [Oka & Okada, 2021](#)). Second, and related to the previous point, the R-DINA model makes it possible for the researcher to specify a prior distribution for the latent classes whenever using EAP or MAP estimation, which can be helpful when attribute hierarchies are expected, and some latent classes are not deemed as possible (e.g., [Tu et al.,](#)

2019). Third, unlike the NPC method, the R-DINA model allows to explore the psychometric properties (e.g., model fit, reliability) related to the attribute profile classifications, which is a crucial requirement for all valid assessments (AERA et al., 2014). Fourth, the restrictive assumption  $\varphi = g_j = s_j \forall j$  of the R-DINA model, which stem from the rationale of the NPC method, seems difficult to be fulfilled in real applications. However, we claim that it might be preferred over more complex parametric CDMs in those conditions in which the amount of information is limited, such as when working with small sample sizes. Following the notation used in Equation 19, in the DINA and DINO models the guessing and slip parameters are estimated as  $\hat{g}_j = R_j^{(0)} / I_j^{(0)}$  and  $\hat{s}_j = [I_j^{(1)} - R_j^{(1)}] / I_j^{(1)}$ , respectively. It can be noted that, if either  $I_j^{(0)}$  or  $I_j^{(1)}$  is small, parameter estimation might not be accurate, even potentially leading to boundary problems in which the estimate is equal to 0 or 1 (Kreitchmann et al., 2022; W. Ma & Jiang, 2021). Hence, under these challenging scenarios, estimating a single parameter for the whole model that summarizes the total amount of probabilistic error might be more adequate than poorly estimating the parameters of specific items. We explore this in a systematic fashion in the following section by means of an exhaustive simulation study in which different sources of model error are considered.

## Simulation Study

### Diagnostic Methods

The R-DINA model was systematically evaluated in the present study. Namely, the Hamming distances using conjunctive ideal responses were obtained using the NPCD package version 1.0-11 (Zheng & Chiu, 2019) of R software (R Core Team, 2021). The *optimize* function from the `stats` R package was used to find the maximum-likelihood estimate of the  $\varphi$  parameter with the Brent's method. The complete implementation of the R-DINA model has



been included in the `cdmTools` package (Nájera et al., 2023) for public usage. The performance of the R-DINA model was compared to that of the DINA model, which was fitted using the `GDINA` package version 2.8.7 (W. Ma & de la Torre, 2020). The MAP estimator was used to make attribute profile classifications in both the R-DINA and DINA models using a uniform prior for the latent classes. This means that the attribute profile classifications made by the R-DINA model are equivalent to those of the NPC method.

### Design and Data Generation

In order to evaluate the performance of the R-DINA model under the presence of model misspecification, data were generated under the DINA model using the `GDINA` package version 2.8.7 and the `cdmTools` package version 1.0.3 (Nájera et al., 2023). Moreover, nine simulation factors were manipulated: sample size ( $N = 25, 50, 100, 200$ ), number of attributes ( $K = 4, 6$ ), number of items per attribute ( $JK = 5, 10$ ), Q-matrix complexity ( $QC = 0.3, 0.4, 0.5$ ), Q-matrix misspecification rate ( $QM = 0, 0.2$ ), average item quality ( $IQ = 0.4, 0.6, 0.8, \text{mixed}$ ), item quality range ( $IQ_R = 0.05, 0.10$ ), attribute thresholds ( $AT = 0, 1$ ), and attribute correlations ( $AC = 0, 0.4, 0.8$ ). One hundred datasets were generated for each of the 4608 conditions resulting from combining the different simulation factor levels.

The Q-matrix complexity refers to the number of cells in the Q-matrix that are equal to 1. In the simulation, the Q-matrices were randomly generated with the only constraints of having, at least, one identity matrix to ensure its completeness (Chiu et al., 2009), and ensuring that each item measured one attribute at the least. The Q-matrix misspecification rate was defined as the proportion of cells in the fitted Q-matrix (i.e., the one used to fit the models) that differed from their corresponding cell in the generating Q-matrix (i.e., the one that was used to generate the data). We considered correctly specified Q-matrices ( $QM = 0$ ) and incorrectly specified Q-

matrices with 20% of misspecified cells ( $QM = 0.2$ ) in the simulation study. Furthermore, the average item quality was computed as  $\sum_{j=1}^J (1 - g_j - s_j) / J$ . Item parameters were drawn from uniform distributions; [Table 1](#) summarizes the correspondence between item parameters, average item quality, and item quality range. The mixed item quality condition was added as a challenging factor for the R-DINA model, as it violates its assumptions. In this condition, the guessing and slip parameters were not similar to each other, but they rotated between small guessing and large slip for odd items (i.e.,  $j = 1, 3, \dots, J - 1$ ) and large guessing and small slip for even items (i.e.,  $j = 2, 4, \dots, J$ ). Thus, even though the average item quality for  $IQ = mixed$  was 0.6, which coincided with the medium item quality condition, this condition was expected to be more demanding for the R-DINA model. Moreover, the item quality range reflected the degree of deviation from the average guessing or slip parameters. A higher item quality range (i.e.,  $IQ_R = 0.10$ ) increases the degree of violation of the R-DINA model's assumption of equally discriminative items.

Attribute profiles were generated following the multivariate normal threshold model ([Chiu et al., 2009](#)). That is,  $K$  continuous latent variables were drawn from a multivariate normal distribution with mean equal to 0 for all variables and correlations equal to  $AC$  (i.e., 0, 0.4, or 0.8) between all variables. Then, the continuous latent variables were discretized into dichotomous attributes by applying a threshold. In the condition of  $AT = 0$ , the threshold was equal to 0 for all variables, and thus all attributes were generated as being equally prevalent in the population. In the condition of  $AT = 1$ , attribute thresholds varied. Namely, they were equal to  $\{-1, -0.33, 0.33, 1\}$  for each of the respective attributes under  $K = 4$ , and equal to  $\{-1, -0.6, -0.2, 0.2, 0.6, 1\}$  under  $K = 6$ . Hence, attributes had a different prevalence in the population.

**Table 1.** *Item Parameter Generation*

Average item quality ( $IQ$ )	Guessing ( $g_j$ ) and slip ( $s_j$ ) parameters
Low quality ( $IQ = 0.4$ )	$g_j, s_j \sim U(0.3 - IQ_R, 0.3 + IQ_R)$
Medium quality ( $IQ = 0.6$ )	$g_j, s_j \sim U(0.2 - IQ_R, 0.2 + IQ_R)$
High quality ( $IQ = 0.8$ )	$g_j, s_j \sim U(0.1 - IQ_R, 0.1 + IQ_R)$
Mixed quality ( $IQ = mixed$ )	Odd items:
	$g_j \sim U(0.1 - IQ_R, 0.1 + IQ_R); s_j \sim U(0.3 - IQ_R, 0.3 + IQ_R)$
	Even items:
	$g_j \sim U(0.3 - IQ_R, 0.3 + IQ_R); s_j \sim U(0.1 - IQ_R, 0.1 + IQ_R)$

*Note.*  $IQ_R$  = item quality range (either 0.05 or 0.10).

The conditions included in the simulation study have been previously used in several simulation studies (e.g., [W. Ma & Jiang, 2021](#); [Nájera et al., 2021](#)) as they are regarded as representative from applied studies ([Sessoms & Henson, 2018](#)). Notably, small sample sizes were considered, which is aligned with the main purpose of the present study of proposing a method that is especially aimed at classroom-level assessments. Previous studies with a similar purpose have employed sample sizes as low as  $N = 20$  ([Oka & Okada, 2021](#)), and applied research with school samples have used between  $N = 44$  and 105 (e.g., [Jang et al., 2015](#); [Ren et al., 2021](#)). Furthermore, Q-matrix complexity is a widely disregarded simulation factor in CDM simulation studies. However, a few works have highlighted the impact that this component might have on parameter estimation and results stability (e.g., [Sorrel et al., 2021](#)).

### Performance measures

The performance of the R-DINA and DINA models was evaluated in terms of posterior probabilities recovery, true classification accuracy, estimated classification accuracy, and relative fit. Note that parameter estimation accuracy is not evaluated because, since the data is generated by the DINA model, there is not a true value for the  $\varphi$  parameter to which  $\hat{\varphi}$  can be compared. However, the item parameters are used in the calculation of posterior probabilities which are our

variable of interest as they are the starting point for the rest of the calculations (e.g., person parameters, reliability, and model fit).

The recovery of the posterior probabilities was evaluated in terms of root-mean-square error (RMSE):

$$RMSE[P(\alpha_l|y_i)] = \sqrt{\frac{1}{L \cdot N} \sum_{l=1}^L \sum_{i=1}^N \left( \hat{P}(\alpha_l|y_i) - P(\alpha_l|y_i) \right)^2}, \quad (19)$$

where  $\hat{P}(\alpha_l|y_i)$  and  $P(\alpha_l|y_i)$  are the estimated and true posterior probabilities of latent class  $l$  for examinee  $i$ , respectively.

True classification accuracy was computed as the proportion of correctly classified attribute profiles (PCP), defined as

$$PCP = \frac{\sum_{i=1}^N I(\hat{\alpha}_i = \alpha_i)}{N}. \quad (20)$$

Estimated classification accuracy was calculated by means of the  $\tau$  index (see [Equation 15](#)) which, as explained above, is an estimation of the PCP. Namely, in order to evaluate the recovery of the estimated classification accuracy, the bias was computed:

$$Bias(\tau) = \tau - PCP. \quad (21)$$

Finally, relative fit was measured by the AIC, BIC, CAIC, and SABIC, as defined in [Equations from 9 to 12](#). Note that the DINA model contains  $2 \times J$  item parameters (guessing and slip parameters for each item) plus  $L - 1$  structural parameters, where  $L$  is the number of latent classes ( $L = 2^K$ ). On the contrary, the R-DINA model contains the same  $L - 1$  structural parameters, but only 1 model parameter (i.e.,  $\varphi$ ). Thus,  $p = L$  for the R-DINA model.

The results are presented in a structured fashion for each of the performance measures. First, in order to summarize the large number of simulation conditions, a repeated measures

ANOVA was conducted on each performance measure. Whenever useful, these results are complemented with a univariate ANOVA to better understand the performance of each model under certain conditions. A partial eta squared higher than 0.14 was considered as a heuristic to identify relevant effect in the ANOVAs (Cohen, 2013). The partial eta squared values for the repeated measures ANOVAs can be consulted on Table A1 of the Appendix. Second, for each performance measure, the average and standard deviation across the levels of each relevant simulation factor (i.e., those with a relevant effect on the ANOVA) is presented for each procedure. Third, graphical visualizations are displayed whenever useful to complement the explanation of the results. Plots were generated with the `ggplot2` package version 3.3.5 (Wickham, 2016) and the `ggpubr` package version 0.4.0 (Kassambara, 2020). All R codes can be accessed at <https://osf.io/deps9/>.

## Results

### *Recovery of Posterior Probabilities*

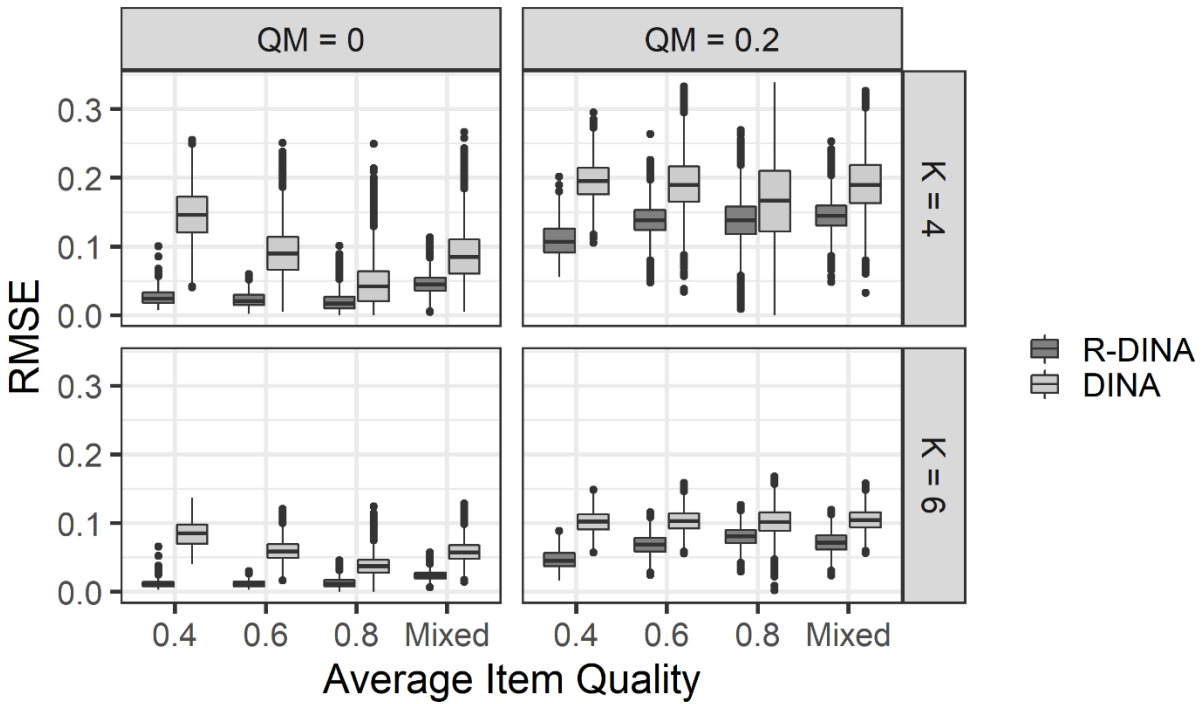
The RMSE of posterior probabilities for the R-DINA and DINA models differed according to the ANOVA ( $\eta_p^2 = .30$ ). The within-factor also interacted with the average item quality ( $\eta_p^2 = .25$ ). Furthermore, the between-factor effects of the Q-matrix misspecification rate ( $\eta_p^2 = .79$ ), the number of attributes ( $\eta_p^2 = .63$ ), their interaction ( $\eta_p^2 = .32$ ), and the interaction between the Q-matrix misspecification rate and the average item quality ( $\eta_p^2 = .17$ ) were also relevant. Table 2 shows the mean and standard deviation of the posterior probabilities RMSE for these relevant simulation factors. The R-DINA model consistently obtained more accurate posterior probabilities ( $.022 \leq \text{Mean RMSE} \leq .100$ ) than the DINA model ( $.076 \leq \text{Mean RMSE} \leq .144$ ). These differences were particularly notable under low-quality items, where the R-DINA model was much more accurate ( $\text{Mean RMSE} = .048$ ) than the DINA ( $\text{Mean RMSE} = .132$ ).

**Table 2.** *Recovery of Posterior Probabilities*

	<i>K</i>		<i>QM</i>			<i>IQ</i>			<i>Total</i>
	4	6	0	0.2	0.4	0.6	0.8	<i>Mixed</i>	
R-DINA	.081 (.057)	.041 (.029)	.022 (.014)	.100 (.040)	.048 (.040)	.061 (.052)	.063 (.063)	.072 (.057)	.061 (.049)
DINA	.139 (.066)	.081 (.029)	.076 (.042)	.144 (.054)	.132 (.050)	.112 (.056)	.087 (.063)	.110 (.057)	.110 (.059)

*Note.* *K* = number of attributes; *QM* = Q-matrix misspecification rate; *IQ* = average item quality. The mean (and standard deviation) is shown in each cell.

Figure 1 depicts the interaction between the four relevant simulation factors concerning the recovery of posterior probabilities. First, it shows that the R-DINA model was robust to the average item quality, especially when the Q-matrix was correctly specified. Second, the R-DINA model provided accurate posterior probabilities under all conditions ( $Mean RMSE \leq .080$ ), except for incorrectly specified Q-matrices and low number of attributes ( $Mean RMSE \geq .109$ ). Third, the recovery of posterior probabilities for the DINA model was largely dependent on the average item quality, particularly under the condition of low number of attributes.



**Figure 1.** Posterior probabilities RMSE as a function of the model (i.e., R-DINA, DINA), average item quality, Q-matrix misspecification rate ( $QM$ ), and number of attributes ( $K$ ).

### *True and Estimated Classification Accuracy*

The R-DINA and DINA models provided a similar true classification accuracy (i.e., PCP) across all conditions, as judged by the fact that the within-factor did not obtain any relevant effect size in the repeated measures ANOVA. However, the PCP of both models was influenced

by the average item quality ( $\eta_p^2 = .71$ ), the Q-matrix misspecification rate ( $\eta_p^2 = .63$ ), the number of attributes ( $\eta_p^2 = .44$ ), the number of items per attribute ( $\eta_p^2 = .37$ ), and the attribute correlations ( $\eta_p^2 = .16$ ). [Table 3](#) summarizes the mean and standard deviations for the PCP across the levels of these factors. In general, both models obtained more accurate classifications with less attributes (*Mean PCP*  $\geq .632$ ), longer tests (*Mean PCP*  $\geq .625$ ), correctly specified Q-matrices (*Mean PCP*  $\geq .672$ ), high-quality items (*Mean PCP*  $\geq .744$ ), and highly correlated attributes (*Mean PCP*  $\geq .597$ ). In general, the R-DINA model provided more stable PCP within factors ( $.158 \geq PCP SD \geq .252$ ) than the DINA model ( $.179 \geq PCP SD \geq .278$ ).

Regarding the estimated classification accuracy bias (i.e.,  $\tau - PCP$ ), the repeated measures ANOVA detected large differences between the R-DINA and the DINA models ( $\eta_p^2 = .42$ ). The within-factor also showed relevant interactions with the average item quality ( $\eta_p^2 = .27$ ) and the Q-matrix misspecification rate ( $\eta_p^2 = .19$ ). [Table 4](#) shows the estimated classification accuracy bias for these simulation factors. The sample size was also included here because, even though it obtained a marginally large interaction effect with the model ( $\eta_p^2 = .13$ ), it is a substantially relevant effect. The main finding that stems from [Table 4](#) is that the R-DINA model obtained a very low estimated classification accuracy bias, which was very consistent between factors ( $.003 \leq Mean Bias \leq .046$ ) and within factors ( $.052 \leq Bias SD \leq .092$ ). On the other hand, the DINA model consistently showed largely overestimated classification accuracy estimates ( $.194 \leq Mean Bias \leq .490$ ), which was particularly large under small sample sizes (*Mean Bias* = .419), misspecified Q-matrices (*Mean Bias* = .431), and low-quality items (*Mean Bias* = .490). The difference between the R-DINA and DINA models in terms of their classification accuracy estimated can be clearly noted in [Figure 2](#). This figure shows, in a logarithmic scale, the frequency (i.e., number of conditions) under which a specific  $\tau$  and PCP were obtained. Thus,



# CDM FOR CLASSROOM-LEVEL ASSESSMENTS

**Table 3.** *True Classification Accuracy*

	<i>K</i>		<i>JK</i>		<i>QM</i>		<i>IQ</i>				<i>AC</i>			
	4	6	5	10	0	0.2	0.4	0.6	0.8	<i>Mixed</i>	0	0.4	0.8	<i>Total</i>
R-DINA	.643 (.208)	.477 (.225)	.493 (.214)	.627 (.229)	.678 (.209)	.443 (.190)	.362 (.158)	.569 (.199)	.744 (.217)	.564 (.201)	.524 (.252)	.559 (.229)	.597 (.207)	.560 (.232)
DINA	.632 (.244)	.470 (.252)	.477 (.243)	.625 (.258)	.672 (.238)	.430 (.225)	.322 (.179)	.560 (.224)	.747 (.217)	.576 (.229)	.494 (.278)	.538 (.255)	.622 (.233)	.551 (.261)

*Note.* *K* = number of attributes; *JK* = number of items per attribute; *QM* = Q-matrix misspecification rate; *IQ* = average item quality;

*AC* = attribute correlations. The mean (and standard deviation) is shown in each cell.

**Table 4.** *Estimated Classification Accuracy Bias*

	<i>N</i>				<i>QM</i>		<i>IQ</i>				<i>Total</i>
	25	50	100	200	0	0.2	0.4	0.6	0.8	<i>Mixed</i>	
R-DINA	.024 (.092)	.025 (.072)	.025 (.059)	.025 (.052)	.003 (.054)	.046 (.078)	.018 (.070)	.028 (.069)	.022 (.071)	.030 (.072)	.025 (.070)
DINA	.419 (.247)	.358 (.223)	.288 (.199)	.230 (.181)	.216 (.198)	.431 (.198)	.490 (.202)	.311 (.201)	.194 (.190)	.300 (.203)	.324 (.226)

*Note.* *N* = sample size; *QM* = Q-matrix misspecification rate; *IQ* = average item quality. The mean (and standard deviation) is shown in

each cell.

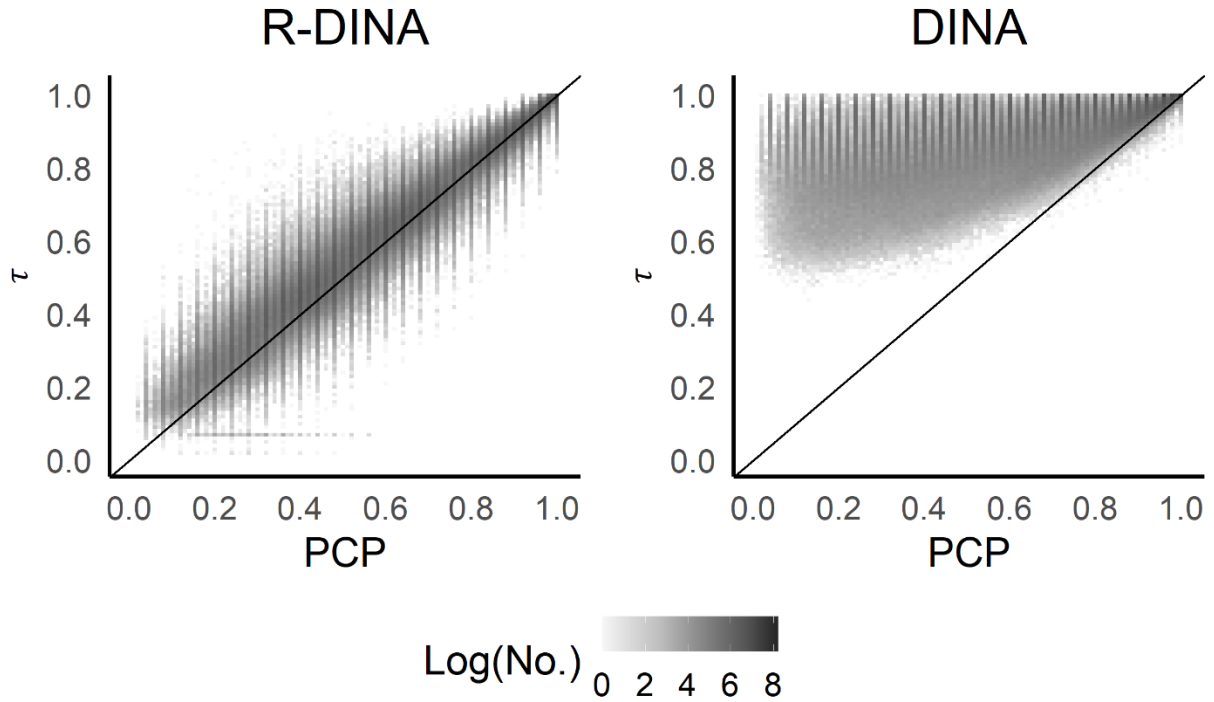
**Table 5.** *Relative Fit Difference between the DINA and the R-DINA*

	<i>N</i>				<i>K</i>		<i>JK</i>		<i>QM</i>		<i>IQ</i>			
	25	50	100	200	4	6	5	10	0	0.2	0.4	0.6	0.8	<i>Mixed</i>
AIC	30	108	257	553	199	275	179	295	155	319	71	168	332	378
BIC	−60	−33	65	309	66	74	69	71	−12	152	−96	1	165	210
CAIC	−134	−107	−9	235	7	−15	20	−28	−86	78	−170	−73	91	136
SABIC	169	199	298	544	251	354	223	382	220	385	136	233	398	443

*Note.* *N* = sample size; *K* = number of attributes; *JK* = number of items per attribute; *QM* = Q-matrix misspecification rate; *IQ* =

average item quality. Positive and negative values indicate better fit for the DINA and the R-DINA, respectively.

the R-DINA model showed a large correlation between the true and estimated classification accuracy ( $r = .953$ ), while the DINA model only showed a moderate relation ( $r = .528$ ). In the most extreme case, the DINA model estimated a classification accuracy of  $\tau = .999$  when, in reality, it was not accurately classifying any examinee ( $PCP = 0$ ).



**Figure 2.** Estimated (i.e.,  $\tau$ ) and true (i.e., PCP) classification accuracy for each of the 460800 simulated data sets.  $\text{Log}(\text{No.})$  = the number of data sets with a particular combination of  $\tau$  and PCP in a logarithmic scale.

### ***Relative Fit***

The R-DINA and the DINA models showed a similar relative fit according to repeated measures ANOVA on each of the four fit indices (i.e., AIC, BIC, CAIC, SABIC). [Table A1](#) in the [Appendix](#) shows all the relevant between-factor effects, which were very similar for the four fit indices. The sample size ( $\eta_p^2 = .99$ ), the number of items per attribute ( $.94 \geq \eta_p^2 \geq .95$ ), the number of attributes ( $.86 \geq \eta_p^2 \geq .89$ ), the average item quality ( $.73 \geq \eta_p^2 \geq .74$ ), and the Q-matrix

misspecification rate ( $\eta_p^2 = .25$ ) were the most influential factors. Table 5 shows, for these factors, the difference on relative fit between the R-DINA and the DINA models. For instance, in the case of the AIC, the table shows the values of  $AIC(R-DINA) - AIC(DINA)$ , where positive and negative outcomes represent a better relative fit for the DINA and the R-DINA, respectively. It can be seen from Table 5 that the AIC and SABIC consistently preferred the DINA model, especially under larger sample sizes and mixed-quality items. This is expected given that the DINA was the generated model. However, the BIC and CAIC preferred the R-DINA model under some conditions. Namely, under sample sizes equal or lower to 50 (for the BIC) or 100 (for the CAIC), under correctly specified Q-matrix, or low-quality items. The CAIC was also lower for the R-DINA model under a large number of attributes, long tests, and medium-quality items. Please note that, as stated before, the repeated measures ANOVA did not reveal large differences between the R-DINA and the DINA for any fit index and simulation factor, so these results should be interpreted with caution.

In order to better understand the correspondence between the best-fitting procedure and its actual performance (i.e., posterior probabilities recovery, true classification accuracy), we conducted an additional analysis, which is summarized in Table 6. Namely, for each fit index and performance measure, we constructed a cross-table that contains: (a) the proportion of conditions in which a procedure obtains the best fit, and (b) the proportion of conditions in which a procedure provides a better performance. Note that, for the PCP, the sum of the proportions included in each cross-table is equal to .901 instead of 1, because both models provided the same PCP in 9.90% of the conditions. In the table, the proportion of correct ( $PC$ ) identification represents the proportion of conditions in which a relative fit index preferred the best performing procedure. The fit indices have been ordered from the one that prefers the DINA model under a

higher proportion of conditions (i.e., SABIC), to the one that prefers the R-DINA model under a higher proportion of conditions (i.e., CAIC). Thus, Table 6 shows that the SABIC and AIC indices preferred the DINA over the R-DINA in the large majority of conditions, even though the DINA model led to a worse overall performance in terms of posterior probabilities recovery and true classification accuracy. On the contrary, the CAIC and BIC indices, which preferred the R-DINA model under a larger proportion of conditions, were able to identify the best performing procedure in more than 50% of the conditions. In any case, no fit index was able to identify the best performing procedure in more than 70% of the conditions, which reflects that there is a mild relation between relative fit and posterior probabilities recovery or true classification accuracy.

**Table 6.** *Correspondence between Relative Fit and Performance*

		Lower PP RMSE			Higher PCP		
		R-DINA	DINA	PC	R-DINA	DINA	PC
SABIC	R-DINA	0.006	0.000	0.060	0.005	0.001	0.443
	DINA	0.939	0.054		0.457	0.438	
AIC	R-DINA	0.113	0.002	0.166	0.074	0.023	0.490
	DINA	0.832	0.053		0.389	0.416	
BIC	R-DINA	0.544	0.014	0.585	0.295	0.189	<b>0.545</b>
	DINA	0.402	0.041		0.167	0.250	
CAIC	R-DINA	0.664	0.021	<b>0.698</b>	0.345	0.253	0.530
	DINA	0.282	0.034		0.117	0.185	
		0.946	0.054		0.463	0.439	

*Note.* Columns reflect the proportion of times that the R-DINA or the DINA model achieved better results in terms of posterior probabilities RMSE (Lower PP RMSE) or classification accuracy (Higher PCP). Rows reflect the proportion of times that each relative fit index preferred the R-DINA or the DINA model. PC reflects the proportion of conditions that a relative fit index selected the best performing procedure for a given performance measure. The best PC for each measure is shown in bold.

### Real Data Illustration

In this section the performance of the R-DINA is illustrated by using two different real data sets. The first data set was the fraction subtraction data (FRAC), which contains the responses of 536 examinees to a test formed by 20 items measuring 8 attributes. The data is publicly available at the `GDINA` package and has been previously analyzed with the DINA model by [de la Torre \(2009\)](#). The second data set corresponds to the responses of 504 examinees to an elementary probability test theory assessment (PTT) formed by 12 items measuring 4 attributes. The data is publicly available at the `edmdata` package ([Balamuta et al., 2021](#)), and has been previously analyzed with the DINA model by [Y. Chen et al., \(2021\)](#). To mimic the simulation study design, 100 subsamples were randomly drawn from each data set with sizes of  $N = 25, 50, 100$ , and  $200$ . The R-DINA and the DINA models were fitted to each subsample, and the following information was extracted: parameter estimates ( $\hat{\phi}$  for the R-DINA model, and average  $\hat{g}_j$  and  $\hat{s}_j$  for the DINA model); classification congruency ( $CC$ ), computed as the proportion of examinees for a given sample size that are classified in the same latent class as with the complete data set; estimated classification accuracy ( $\tau$  index); and relative fit. For the latter, only the CAIC and SABIC are reported, since they showed the highest preference for the R-DINA and the DINA in the simulation study, respectively.

The mean and standard deviation for the parameter estimates across the 100 replicates for each sample size is presented in [Table 7](#). First, it should be noted that the R-DINA model provided very consistent parameter estimates across the different sample sizes and within the 100 replicates for each sample size, both for the FRAC ( $.109 \geq \hat{\phi} \geq .113$ ) and PPT ( $.123 \geq \hat{\phi} \geq .126$ ) data sets. The parameter estimates of the DINA model were less stable across and within sample sizes. The average guessing and slip parameters were more similar in magnitude in the FRAC

data set ( $.103 \geq \hat{g}_j \geq .127$ ;  $.107 \geq \hat{s}_j \geq .142$ ), although the difference between both parameter estimates was not very large either for the PTT data ( $.204 \geq \hat{g}_j \geq .270$ ;  $.094 \geq \hat{s}_j \geq .118$ ). The higher similarity between the guessing and slip estimates in the FRAC data suggests that the two models should perform more similarly for these data compared to the PTT.

**Table 7.** *Parameter Estimates*

	$N = 25$	$N = 50$	$N = 100$	$N = 200$	$N = N$
FRAC					
$\hat{\phi}$	.109 (.017)	.113 (.011)	.111 (.006)	.112 (.005)	.112
$\hat{g}_j$	.127 (.041)	.120 (.031)	.109 (.020)	.103 (.016)	.103
$\hat{s}_j$	.107 (.027)	.121 (.022)	.129 (.015)	.138 (.010)	.142
PTT					
$\hat{\phi}$	.126 (.023)	.123 (.015)	.123 (.009)	.124 (.006)	.124
$\hat{g}_j$	.270 (.089)	.223 (.075)	.205 (.056)	.204 (.048)	.224
$\hat{s}_j$	.094 (.029)	.106 (.023)	.115 (.014)	.117 (.009)	.118

*Note.* Average estimates (and standard deviations) across the 100 replicates are presented in each

cell.  $N = N$  represents the estimation with the complete dataset.  $\hat{\phi}$  refers to the R-DINA parameter, while  $\hat{g}_j$  and  $\hat{s}_j$  refer to the guessing and slip parameters of the DINA model, respectively.

Regarding the classification congruency between the subsamples and the whole data sets, the R-DINA model always provided the same attribute profile classifications regardless of the sample size (note that these classifications are the same as those provided by the NPC method). That is, the average  $CC$  was equal to 1 for all sample sizes, for both the FRAC and PTT data sets. This was not the case for the DINA model, which obtained average  $CC = .458, .479, .514$ , and  $.560$  for  $N = 25, 50, 100$ , and  $200$ , respectively, with the FRAC data. The  $CC$  was higher for the PTT data ( $CC = .803, .883, .920$ , and  $.939$ ), which is explained by the fact that the PTT test involved 4 attributes, instead of the 8 attributes measured with the FRAC. Thus, it is easier to

have a full match in the attribute profiles. It is important to recall that the *CC* only measures the consistency within a method, and thus it does not reflect classification accuracy.

[Table 8](#) shows the estimated classification accuracy. First, note that the R-DINA model always reported a lower classification accuracy than the DINA model, which is in line with the simulation study, in which the DINA model largely overestimated classification accuracy. Second, the estimated classification accuracy by the DINA model decreased as the sample size increased; this is also aligned with the previous results, since the classification accuracy bias is expected to be milder as item parameter estimation becomes more accurate (by increasing the sample size). On the contrary, the estimated classification accuracy was very consistent across sample sizes for the R-DINA, as expected by the abovementioned stability in the parameter estimates. Finally, the estimated classification accuracy was higher for the PTT than for the FRAC data set, which is also explained by the former having half the attributes than the latter.

**Table 8.** *Estimated Classification Accuracy*

	$N = 25$	$N = 50$	$N = 100$	$N = 200$	$N = N$
			FRAC		
R-DINA	.403 (.064)	.402 (.054)	.406 (.035)	.404 (.021)	.402
DINA	.567 (.075)	.573 (.063)	.562 (.045)	.556 (.028)	.543
			PTT		
R-DINA	.864 (.053)	.868 (.028)	.869 (.020)	.869 (.014)	.867
DINA	.977 (.022)	.967 (.022)	.961 (.017)	.950 (.016)	.937

*Note.* Average values (and standard deviations) across the 100 replicates are presented in each

cell.  $N = N$  represents the estimation with the whole sample size.

Finally, [Table 9](#) shows the CAIC and SABIC of both models. Both fit indices reported a better relative fit for the DINA over the R-DINA across all sample sizes. The difference between the fit of both procedures was smaller for the FRAC data, which was expected by the fact that the average guessing and slip parameters were more similar in these data set. Moreover, the CAIC reported a more similar fit between the procedures than the SABIC, which is in line with

the simulation results that showed that the SABIC had a great tendency towards the DINA model, while the CAIC tended to prefer the R-DINA model under some conditions. Finally, the fit was also more similar for both procedures with smaller sample sizes, which was also a simulation study finding. Nevertheless, these indices should be interpreted with caution as noted by the low correspondence between better fit and performance (i.e., posterior probabilities recovery, classification accuracy) shown in the simulation study.

**Table 9.** *Relative Fit*

		$N = 25$	$N = 50$	$N = 100$	$N = 200$	$N = N$
		FRAC				
CAIC	R-DINA	1594 (33)	2305 (42)	3516 (51)	5789 (72)	13039
	DINA	1588 (34)	2211 (42)	3233 (52)	5096 (68)	10953
SABIC	R-DINA	544 (33)	1245 (42)	2452 (51)	4722 (72)	11971
	DINA	378 (34)	990 (42)	2007 (52)	3866 (68)	9722
		PTT				
CAIC	R-DINA	411 (20)	761 (29)	1457 (34)	2844 (44)	7036
	DINA	380 (21)	651 (33)	1168 (43)	2187 (61)	5239
SABIC	R-DINA	345 (20)	694 (29)	1391 (34)	2777 (44)	6969
	DINA	220 (21)	489 (33)	1005 (43)	2025 (61)	5077

*Note.* Average values (and standard deviations) across the 100 replicates are presented in each cell.  $N = N$  represents the estimation with the whole sample size.

### Discussion

Cognitive diagnosis modeling (CDM) has a great potential to be used in educational assessments at a classroom level, given that they provide detailed information about students' strengths and weaknesses that can guide remedial instruction and teachers' efforts (Chiu et al., 2018; Paulsen & Valdivia, 2021). Despite this promising area of application, few studies have been conducted with small sample sizes (Sessoms & Henson, 2018). This is mainly due to the lack of optimal diagnostic procedures for these challenging settings. Parametric CDM requires large sample sizes in order to obtain accurate item parameter estimates. A disrupted estimation will lead to biased parameters, less precise attribute classifications, and overestimated



classification accuracy ([Kreitchmann et al., 2022](#); [W. Ma & Jiang, 2021](#); [Sen & Cohen, 2021](#)).

To address this, nonparametric CDM was proposed as a suitable alternative to provide accurate attribute profile classifications under those challenging conditions that disrupt parameter estimation (e.g., small sample size, low-quality items, complex Q-matrices; [Chiu et al., 2018](#); [C. Ma et al., 2022](#); [Oka & Okada, 2021](#)). However, up until today it was not possible to derive the likelihood nor posterior probabilities from these methods, which prevented from calculating fit and reliability indices. Thus, the validity of nonparametric CDM could not be explored, and hence its application in real settings entailed a leap of faith. This is an unacceptable practice according to the standards ([AERA et al., 2014](#)) since it compromises the validity of the measurement. The main purpose of the present paper was to present the R-DINA model, a parametric CDM that allows to compute the likelihoods (and fit indices) and posterior probabilities (and classification accuracy measures) from the attribute profile classifications made by the NPC method. By doing this, a comprehensive CDM procedure for classroom-level assessments can be achieved, combining the accurate classifications from nonparametric CDM method and the possibility of evaluating the properties of the method as in a parametric CDM.

The performance of the R-DINA model was evaluated and compared to that of the DINA model by means of an exhaustive simulation study. First, it was observed that the  $\varphi$  parameter provided a better recovery of posterior probabilities; in fact, it provided more accurate posterior probabilities than the DINA model even under those conditions that do not follow the assumptions of the R-DINA model, such as a mixed-quality items. Second, even though the true classification accuracy (i.e., PCP) was very similar for both procedures under most conditions, the poor recovery of posterior probabilities for the DINA model resulted in a largely overestimated classification accuracy. That is, classification accuracy measures will report very

high values for the DINA model regardless of its true precision, giving a false sense of confidence in the attribute profile classifications. On the contrary, the R-DINA model reported very accurate classification accuracy estimates, meaning that researchers can trust its empirical reliability. Finally, both procedures led to similar relative fit, with the BIC and CAIC (that heavily penalize complexity) showing better fit for the R-DINA model under challenging conditions (e.g., small sample size, low-quality items), and the AIC and SABIC (that penalize complexity to a lower extent) preferring the DINA model. This was expected, since the DINA was the generating model. However, the low correspondence between the value of these fit indices and the actual performance of the procedures suggests that practitioners should be cautious when basing their decisions on the relative fit. In this study, the BIC and CAIC were the indices that correctly identified the best performing procedure in more conditions.

The simulation results were also illustrated using real data. In general, the findings were congruent with those of the simulation study. Namely, parameter estimation for the R-DINA model was very stable across sample sizes and replicates; the DINA model showed different estimates with sample sizes of 25 and 200 for both data sets. The estimated classification accuracy was also higher for the DINA model, particularly under very small sample sizes, which reflects the overestimation tendency found in the simulation study due to a poor item parameter recovery. The lower dependency of the R-DINA model on sample size was also reflected by the fact that it made the same attribute profile classifications regardless of sample size ( $CC = 1$ ), while the DINA model obtained more consistent attribute profile estimates as sample size increased. Finally, relative fit was always at least slightly better for the DINA model. As stated before, these results should be interpreted with caution, since this might not necessarily need that

the DINA model should be preferred in terms of parameter estimation accuracy or true classification accuracy.

Despite the promising performance of the R-DINA model in small sample size scenarios, the reader might wonder whether it is realistic to assume that all items have the same guessing and slip parameters. We are fully aware that this is a very restrictive assumption, and that it will be very unlikely to be fulfilled in real applications. However, statistical models should not only be judged by their phenomenological truthfulness, since all models are wrong to some extent, but by their practical usefulness in those scenarios for which they have been developed (Box, 1976). In that regard, the higher parsimony of the R-DINA model compared to the DINA model makes from the former a less plausible but more stable model. In the present paper, the R-DINA model has shown a more satisfactory performance than the DINA model with small sample sizes, even when the degree of model misspecification was moderate (e.g., 20% of misspecified q-entries, dissimilar guessing and slip parameters). This is also in line with previous studies, where the NPC method showed a satisfactory performance even when the data was generated by the G-DINA model (Chiu et al., 2018), which implies a large degree of model misspecification (recall that the R-DINA model provides the same classifications as the NPC method). All this evidence points out in the direction that, whenever the amount of information in the data is scarce, fitting a more restrictive model might be more beneficial (e.g., better parameter recovery and classification accuracy) than wrongly estimating a complex one (e.g., Sorrel et al., 2021). Note that the opposite is also true and that, whenever there is enough information to reliably estimate a more complex model, fitting a more restrictive method will not provide any added value.

Given the above, a question might arise about how to identify the preferred CDM to use in a particular classroom-level assessment, especially since the simulation study revealed that

relative fit indices do not always lead to the method that provided the more desirable results. Two possible tentative approaches could be taken. The first one consists in applying a bootstrapping procedure, in which the DINA and the R-DINA models are systematically applied to several data sets formed by randomly resampling with a replacement from the original data set. Thus, an empirical distribution of item parameter estimates and attribute profile classifications could be constructed as a measure of the stability of the results. A similar approach has been already implemented in other psychometric frameworks, such as exploratory graph analysis (Christensen & Golino, 2021). The second approach is based on the evaluation of absolute fit for both the DINA and R-DINA models, both at the item (J. Chen et al., 2013) and test level (Hansen et al., 2016). Here, the challenge might be related to the low statistical power derived from the very small sample sizes. Related to this, model comparison at the item level could be also extended to the R-DINA model to differentiate whether an item conforms to a conjunctive, disjunctive, or general response (Sorrel, Abad, et al., 2017; Sorrel, de la Torre, et al., 2017). On another note, the R-DINA model could be also extended to cognitive diagnostic computerized adaptive testing (CD-CAT), thus allowing for variable length applications with very small sample sizes, with a similar rationale than the one used by the *nonparametric item selection* rule (Chang et al., 2019). Finally, evaluating the performance of the R-DINA model using a prior attribute distribution (in comparison to maximum likelihood estimation or the NPC method) might be worth considering given the often encountered highly correlated attributes in applied studies (Sessoms & Henson, 2018).

We believe that the R-DINA model fills an important gap in the cognitive diagnosis framework by providing both accurate and robust attribute profile classifications (as the NPC method) and measures of model fit and reliability (as any other parametric CDM) in small-scale

## CDM FOR CLASSROOM-LEVEL ASSESSMENTS

assessments. By making progress in these directions, CDM could be applied not only in large-scale assessments ([Sessoms & Henson, 2018](#)), but also at a classroom level, where the diagnostic information can be used to better guide remedial instruction.

## References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.
- Balamuta, J. J., Culpepper, S. A., & Douglas, J. A. (2021). *Edmdata: Data Sets for Psychometric Modeling*. R package version 1.2.0. <https://CRAN.R-project.org/package=edmdata>
- Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, 71(356), 791–799. <https://doi.org/10.1080/01621459.1976.10480949>
- Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3), 345–370. <https://doi.org/10.1007/BF02294361>
- Brent, R. P. (2002). *Algorithms for minimization without derivatives*. Dover Publications.
- Chang, Y.-P., Chiu, C.-Y., & Tsai, R.-C. (2019). Nonparametric CAT for CD in Educational Settings With Small Samples. *Applied Psychological Measurement*, 43(7), 543–561. <https://doi.org/10.1177/0146621618813113>
- Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and Absolute Fit Evaluation in Cognitive Diagnosis Modeling: Relative and Absolute Fit Evaluation in CDM. *Journal of Educational Measurement*, 50(2), 123–140. <https://doi.org/10.1111/j.1745-3984.2012.00185.x>

- Chen, Y., Liu, Y., Culpepper, S. A., & Chen, Y. (2021). Inferring the Number of Attributes for the Exploratory DINA Model. *Psychometrika*, 86(1), 30–64.  
<https://doi.org/10.1007/s11336-021-09750-9>
- Chiu, C.-Y., & Douglas, J. (2013). A Nonparametric Approach to Cognitive Diagnosis by Proximity to Ideal Response Patterns. *Journal of Classification*, 30(2), 225–250.  
<https://doi.org/10.1007/s00357-013-9132-9>
- Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster Analysis for Cognitive Diagnosis: Theory and Applications. *Psychometrika*, 74(4), 633–665. <https://doi.org/10.1007/s11336-009-9125-0>
- Chiu, C.-Y., Sun, Y., & Bian, Y. (2018). Cognitive Diagnosis for Small Educational Programs: The General Nonparametric Classification Method. *Psychometrika*, 83(2), 355–375.  
<https://doi.org/10.1007/s11336-017-9595-4>
- Christensen, A. P., & Golino, H. (2021). Estimating the Stability of Psychological Dimensions via Bootstrap Exploratory Graph Analysis: A Monte Carlo Simulation and Tutorial. *Psych*, 3(3), 479–500. <https://doi.org/10.3390/psych3030032>
- Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences* (0 ed.). Routledge.  
<https://doi.org/10.4324/9780203771587>
- de la Torre, J. (2009). DINA Model and Parameter Estimation: A Didactic. *Journal of Educational and Behavioral Statistics*, 34(1), 115–130.  
<https://doi.org/10.3102/1076998607309474>
- de la Torre, J., & Minchen, N. (2014). Cognitively Diagnostic Assessments and the Cognitive Diagnosis Model Framework. *Psicología Educativa*, 20(2), 89–97.  
<https://doi.org/10.1016/j.pse.2014.11.001>

- de la Torre, J. (2011). The Generalized DINA Model Framework. *Psychometrika*, 76(2), 179–199. <https://doi.org/10.1007/s11336-011-9207-7>
- Gao, X., Ma, W., Wang, D., Cai, Y., & Tu, D. (2021). A Class of Cognitive Diagnosis Models for Polytomous Data. *Journal of Educational and Behavioral Statistics*, 46(3), 297–322. <https://doi.org/10.3102/1076998620951986>
- Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology*, 69(3), 225–252. <https://doi.org/10.1111/bmsp.12074>
- 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(2), 191–210. <https://doi.org/10.1007/s11336-008-9089-5>
- Jang, E. E., Dunlop, M., Park, G., & van der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing*, 32(3), 359–383. <https://doi.org/10.1177/0265532215570924>
- Johnson, M. S., & Sinharay, S. (2018). Measures of Agreement to Assess Attribute-Level Classification Accuracy and Consistency for Cognitive Diagnostic Assessments: Measures of Agreement to Assess Attribute-Level Classification Accuracy and Consistency. *Journal of Educational Measurement*, 55(4), 635–664. <https://doi.org/10.1111/jedm.12196>
- Johnson, M. S., & Sinharay, S. (2020). The Reliability of the Posterior Probability of Skill Attainment in Diagnostic Classification Models. *Journal of Educational and Behavioral Statistics*, 45(1), 5–31. <https://doi.org/10.3102/1076998619864550>



- Junker, B. W., & Sijtsma, K. (2001). Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory. *Applied Psychological Measurement*, 25(3), 258–272. <https://doi.org/10.1177/01466210122032064>
- Kassambara, A. (2020). *Ggpubr: “ggplot2” Based Publication Ready Plots*. R package version 0.4.0. <https://CRAN.R-project.org/package=ggpubr>
- Köhn, H.-F., & Chiu, C.-Y. (2016). A Proof of the Duality of the DINA Model and the DINO Model. *Journal of Classification*, 33(2), 171–184. <https://doi.org/10.1007/s00357-016-9202-x>
- Kreitchmann, R. S., de la Torre, J., Sorrel, M. A., Nájera, P., & Abad, F. J. (2022). Improving reliability estimation in cognitive diagnosis modeling. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-022-01967-5>
- Liu, C.-W., Andersson, B., & Skrondal, A. (2020). A Constrained Metropolis–Hastings Robbins–Monro Algorithm for Q Matrix Estimation in DINA Models. *Psychometrika*, 85(2), 322–357. <https://doi.org/10.1007/s11336-020-09707-4>
- Ma, C., de la Torre, J., & Xu, G. (2022). Bridging Parametric and Nonparametric Methods in Cognitive Diagnosis. *Psychometrika*. <https://doi.org/10.1007/s11336-022-09878-2>
- Ma, W., & de la Torre, J. (2020). GDINA: An R Package for Cognitive Diagnosis Modeling. *Journal of Statistical Software*, 93(14). <https://doi.org/10.18637/jss.v093.i14>
- Ma, W., & Jiang, Z. (2021). Estimating Cognitive Diagnosis Models in Small Samples: Bayes Modal Estimation and Monotonic Constraints. *Applied Psychological Measurement*, 45(2), 95–111. <https://doi.org/10.1177/0146621620977681>

- Nájera, P., Abad, F. J., & Sorrel, M. A. (2021). Determining the Number of Attributes in Cognitive Diagnosis Modeling. *Frontiers in Psychology, 12*, 614470.  
<https://doi.org/10.3389/fpsyg.2021.614470>
- Nájera, P., Sorrel, M. A., & Abad, F. J. (2023). *cdmTools: Useful Tools for Cognitive Diagnosis Modeling. R package version 1.0.3*. <https://github.com/Pablo-Najera/cdmTools>
- Oka, M., & Okada, K. (2021). *Assessing the Performance of Diagnostic Classification Models in Small Sample Contexts with Different Estimation Methods*.  
<https://doi.org/10.48550/ARXIV.2104.10975>
- Paulsen, J., & Svetina, D. (2021). Examining cognitive diagnostic modeling in classroom assessment conditions. *The Journal of Experimental Education, 1–18*.  
<https://doi.org/10.1080/00220973.2021.1891008>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ren, H., Xu, N., Lin, Y., Zhang, S., & Yang, T. (2021). Remedial Teaching and Learning From a Cognitive Diagnostic Model Perspective: Taking the Data Distribution Characteristics as an Example. *Frontiers in Psychology, 12*, 628607.  
<https://doi.org/10.3389/fpsyg.2021.628607>
- Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2).  
<https://doi.org/10.1214/aos/1176344136>
- Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*(3), 333–343. <https://doi.org/10.1007/BF02294360>

- Sen, S., & Cohen, A. S. (2021). Sample Size Requirements for Applying Diagnostic Classification Models. *Frontiers in Psychology, 11*, 621251. <https://doi.org/10.3389/fpsyg.2020.621251>
- Sessoms, J., & Henson, R. A. (2018). Applications of Diagnostic Classification Models: A Literature Review and Critical Commentary. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 1–17. <https://doi.org/10.1080/15366367.2018.1435104>
- Sorrel, M. A., Abad, F. J., & Nájera, P. (2021). Improving Accuracy and Usage by Correctly Selecting: The Effects of Model Selection in Cognitive Diagnosis Computerized Adaptive Testing. *Applied Psychological Measurement, 45*(2), 112–129. <https://doi.org/10.1177/0146621620977682>
- Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential Item-Fit Evaluation in Cognitive Diagnosis Modeling. *Applied Psychological Measurement, 41*(8), 614–631. <https://doi.org/10.1177/0146621617707510>
- Sorrel, M. A., de la Torre, J., Abad, F. J., & Olea, J. (2017). Two-Step Likelihood Ratio Test for Item-Level Model Comparison in Cognitive Diagnosis Models. *Methodology, 13*(Supplement 1), 39–47. <https://doi.org/10.1027/1614-2241/a000131>
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with missconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345–354.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305. <https://doi.org/10.1037/1082-989X.11.3.287>
- Tu, D., Wang, S., Cai, Y., Douglas, J., & Chang, H.-H. (2019). Cognitive Diagnostic Models With Attribute Hierarchies: Model Estimation With a Restricted Q-Matrix Design.

- Applied Psychological Measurement*, 43(4), 255–271.  
<https://doi.org/10.1177/0146621618765721>
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, 61(2), 287–307.  
<https://doi.org/10.1348/000711007X193957>
- Wang, S., & Douglas, J. (2015). Consistency of Nonparametric Classification in Cognitive Diagnosis. *Psychometrika*, 80(1), 85–100. <https://doi.org/10.1007/s11336-013-9372-y>
- Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-Level and Pattern-Level Classification Consistency and Accuracy Indices for Cognitive Diagnostic Assessment: Attribute-Level and Pattern-Level Indices for CDA. *Journal of Educational Measurement*, 52(4), 457–476. <https://doi.org/10.1111/jedm.12096>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.  
<https://ggplot2.tidyverse.org>
- Xu, X., de la Torre, J., Zhang, J., Guo, J., & Shi, N. (2020). Estimating CDMs Using the Slice-Within-Gibbs Sampler. *Frontiers in Psychology*, 11, 2260.  
<https://doi.org/10.3389/fpsyg.2020.02260>
- Zheng, Y., & Chiu, C.-Y. (2019). *NPCD: Nonparametric Methods for Cognitive Diagnosis*. *R package version 1.0-11*. <https://CRAN.R-project.org/package=NPCD>

## Appendix

**Table A1.** *Partial Eta Squared for Repeated Measures ANOVAs*

Effect	PP <i>RMSE</i>	PCP	$\tau$ <i>Bias</i>	Relative fit
<i>Within Effects</i>				
<i>M</i>	.30		.42	
<i>M</i> $\times$ <i>N</i>				
<i>M</i> $\times$ <i>JK</i>				
<i>M</i> $\times$ <i>QM</i>			.19	
<i>M</i> $\times$ <i>IQ</i>	.25		.27	
<i>M</i> $\times$ <i>AC</i>				
<i>Between Effects</i>				
<i>N</i>			.13	.99
<i>K</i>	.63	.44		.86 – .89
<i>JK</i>		.37		.94 – .95
<i>QM</i>	.79	.63	.35	.25
<i>IQ</i>		.71	.26	.73 – .74
<i>AC</i>		.16		
<i>N</i> $\times$ <i>JK</i>				.89
<i>N</i> $\times$ <i>K</i>				.75 – .76
<i>N</i> $\times$ <i>QM</i>				.30
<i>N</i> $\times$ <i>IQ</i>				.59
<i>K</i> $\times$ <i>JK</i>				.38 – .40
<i>K</i> $\times$ <i>QM</i>	.32			
<i>JK</i> $\times$ <i>IQ</i>				.30
<i>QM</i> $\times$ <i>IQ</i>	.17			

*Note.* *M* = model; *N* = sample size; *QM* = Q-matrix misspecification rate; *IQ* = average item

quality; *AC* = attribute correlations; *K* = number of attributes; *JK* = number of items per attribute;

PP *RMSE* = RMSE of posterior probabilities; PCP = proportion of correctly classified attribute

profiles;  $\tau$  *Bias* = bias of the  $\tau$  index. Relative fit includes the four fit indices (i.e., AIC, BIC,

CAIC, SABIC), with the minimum and maximum partial eta squared among the four indices

shown in each cell.