anales de psicología / annals of psychology
2026, vol. 42, nº 2 (may-august), e16, 1-14
https://doi.org/10.6018/analesps.650971

Published by Servicio de Publicaciones, Universidad de Murcia (Spain), in https://revistas.um.es/analesps

ISSN online: 1695-2294.

Issues and possible solutions in cognitive diagnosis modeling applications: The case of a large-scale educational assessment in Mexico

Scarlett Escudero[1], Ramsés Vázquez-Lira[2], Iwin Leenen[2], and Miguel A. Sorrel[3,*]

1 Department of Educational Psychology, University of Minnesota, United States
2 Faculty of Psychology, Universidad Nacional Autónoma de México, Mexico
3 Faculty of Psychology, Universidad Autónoma de Madrid, Spain

Título: Problemas y posibles soluciones en aplicaciones de modelos de diagnóstico cognitivo: El caso de una evaluación educativa a gran escala en México.

Resumen: Los modelos de diagnóstico cognitivo (CDM, por sus siglas en inglés) son un marco propuesto para la creación y análisis de instrumentos de medición que se originó en el campo de la educación y se ha extendido a otras áreas de interés en la psicología. Estos modelos han recibido una gran atención en los últimos años, lo que ha dado lugar a numerosas contribuciones teóricas. Sin embargo, aún existe una escasez de estudios que apliquen estas metodologías a datos empíricos. Es fundamental evaluar los procedimientos en su contexto natural de aplicación y proporcionar soluciones adaptadas a los problemas que puedan surgir en la práctica. El objetivo de este estudio es aplicar CDM a una evaluación a gran escala diseñada bajo este marco, la cual tiene como objetivo evaluar a docentes de nivel medio superior en México. En el análisis de estos datos, se identificaron cinco problemas y se presentan directrices sobre cómo explorarlos y encontrar soluciones utilizando R. Además de discutir soluciones psicométricas, este trabajo enfatiza la importancia de la consulta con expertos en contenido. Al destacar los posibles desafíos que pueden surgir incluso cuando se cuentan con todos los elementos necesarios para una evaluación educativa a gran escala basada en CDM, este estudio busca orientar futuras aplicaciones empíricas y propuestas metodológicas.

Palabras clave: Modelos de diagnóstico cognitivo. Evaluación educativa. Validez. Lenguaje R.

Abstract: Cognitive diagnosis modeling (CDM) is a proposed framework for the creation and analysis of measurement tools that originated in the field of education and have extended to other areas of interest in psychology. These models have received a lot of attention in recent years, resulting in an abundance of theoretical contributions. However, there is still a shortage of studies applying these methodologies to real data. It is essential to evaluate the procedures in their natural context of application and to provide solutions tailored to the problems that may arise in practice. The purpose of this study is to apply CDM to a large-scale evaluation that assesses high school teachers in Mexico, which was designed based on a CDM framework. Five issues are identified that arose in the analysis of these data and provide guidelines on how to explore and find solutions to these issues with implementation in R. Besides discussing psychometric solutions, this paper also emphasizes the importance of consulting with content experts. By highlighting potential challenges that can arise even when all necessary elements for a large-scale CDM educational assessment are in place, this study aims to guide future empirical applications and methodological proposals.

Keywords: Cognitive diagnosis modeling. Educational assessment. Validity. R language.

Cognitive diagnosis models (CDMs) are latent class models that are used to classify individuals in latent classes through the presence or absence of attributes based on their response pattern. Each item in the test requires mastery of different latent skills, usually referred to as *attributes*. The output of CDMs has the potential to facilitate formative assessment in educational settings (Ren et al., 2021; Sanz et al., 2023), but also to guide clinical interventions (Tan et al., 2022; Templin & Henson, 2006) and support applications in organizational contexts (García et al., 2014; Sorrel et al., 2016). Despite this, there is a notable lack of empirical applications that have been devised from the CDM framework. Consequently, in many cases, the empirical illustrations included in studies do not constitute real examples of application. This means that the models have not been sufficiently tested, and studies addressing this matter are necessary to identify potential problems and determine if the developed solutions can solve these problems, thereby guiding the development of new methodologies.

*** Correspondence address [Dirección para correspondencia]:**
Miguel A. Sorrel. Faculty of Psychology, 6 Iván Pavlov St, Cantoblanco Campus, Madrid (Spain). E-mail: miguel.sorrel@uam.es
(Article received: 21-02-2025; revised: 4-08-2025; accepted: 7-10-2025)

This is the purpose of the present study, which uses empirical data from a large-scale assessment originated from CDMs and enumerates issues and possible solutions. The structure of the article is as follows. First, CDMs are briefly introduced. Second, empirical applications to date are reviewed. Then, in the method section, the assessment and data analysis procedure are described. Finally, the results are discussed, drawing conclusions to guide future research.

Cognitive Diagnosis Modeling

CDM attributes measured in a test, denoted by *K*, are typically classified into two states: mastery and non-mastery for each *i*th examinee ($\boldsymbol{\alpha}_i \in \{0,1\}^K$). Each attribute represents a specific knowledge, skill, ability, or other characteristic (KSAO). Cognitive diagnostic models estimate the probability that a person has mastered each attribute. For example, in a test, one attribute might be 'solving linear equations' and another 'working with fractions.' The model then classifies each attribute for each person as mastered or not mastered.

Different models pose a specific item response function that accounts for the relationship between the assessed attributes and the items. In addition to the responses of individuals to the items, most models will require content ex-

perts to generate a matrix specifying which attribute k is involved in each item j ($J \times K$), commonly known as a Q-matrix, for their estimation (Tatsuoka, 1983). In most cases, this Q-matrix is constructed a priori based on expert judgment, clinical theory, or empirical research findings, although there are alternatives to estimate it from the data (Chen et al., 2023). An example of a Q-matrix is shown in Table 1, which is an extract of the initial Q-matrix used in this study. The attributes are shown in Table 2. The items may vary in complexity. For example, Item 1 evaluates two attributes (9 and 11) out of the 12 attributes assessed by the test. The attribute specification for item j is given by the K-length vector, $\boldsymbol{q}_j$. For example, for Item 1, $\boldsymbol{q}_j = \{000000001010\}$. Each item has an associated q-vector ($\boldsymbol{q}_j$) of length K, where each element (q_{jk}) is 1 if the attribute k is measured in item j and 0 if k is not measured.

We can also establish the attribute vector that indicates mastery or non-mastery of attributes by an individual, $\boldsymbol{\alpha}_l$, where $l = 1, \dots, L = 2^K$, L being the number of possible latent classes. All CDMs can be expressed as $P(X_j = 1|\boldsymbol{\alpha}_l) = P_j(\boldsymbol{\alpha}_l)$, where the probability of success is conditional to latent class (l). The latent classes define the possible combinations of attributes that can occur and represent the outcome of interest.

Most commonly, each item measures only a subset of the attributes assessed in the test. We can identify $\boldsymbol{\alpha}^*_{lj}$ as the reduced attribute vector that includes only the K^*_j attributes measured in item j, thus having a K^*_j-length. For example, in Item 1, $K^*_j = 2$, and $\boldsymbol{\alpha}^*_{lj}$ would be $\boldsymbol{\alpha}^*_{11} = \{00\}$, $\boldsymbol{\alpha}^*_{21} = \{10\}$, $\boldsymbol{\alpha}^*_{31} = \{01\}$, and $\boldsymbol{\alpha}^*_{41} = \{11\}$. As a partition of latent classes, we can establish $2^{K^*_j}$ latent groups, in this case, four for Item 1.

Table 1
Extract of the Initial Q-Matrix used in the Empirical Study.

Item	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12
1	0	0	0	0	0	0	0	0	1	0	1	0
2	0	0	0	0	0	1	0	0	0	0	1	1
…	…	…	…	…	…	…	…	…	…	…	…	…
99	0	0	0	0	1	0	0	0	0	0	0	0
100	0	0	0	0	1	0	0	0	0	0	0	0

Note. A= Attribute. Only the first and last two items in the test are shown.

Thus, both the number of latent classes and their interpretation are established a priori in the case of CDMs. There are several constraints that can be imposed regarding how the underlying attributes interact to produce the observed responses. These interactions include conjunctive, disjunctive, and additive processes (see de la Torre & Sorrel, 2023 for an introduction to these models). The deterministic input noisy "and" gate model (DINA; Junker & Sijtsma, 2001) attests to a conjunctive interaction. Meanwhile, the deterministic input noisy "or" gate model (DINO; Templin & Henson, 2006) has a disjunctive approach.

Table 2
Attributes Measured in the Test.

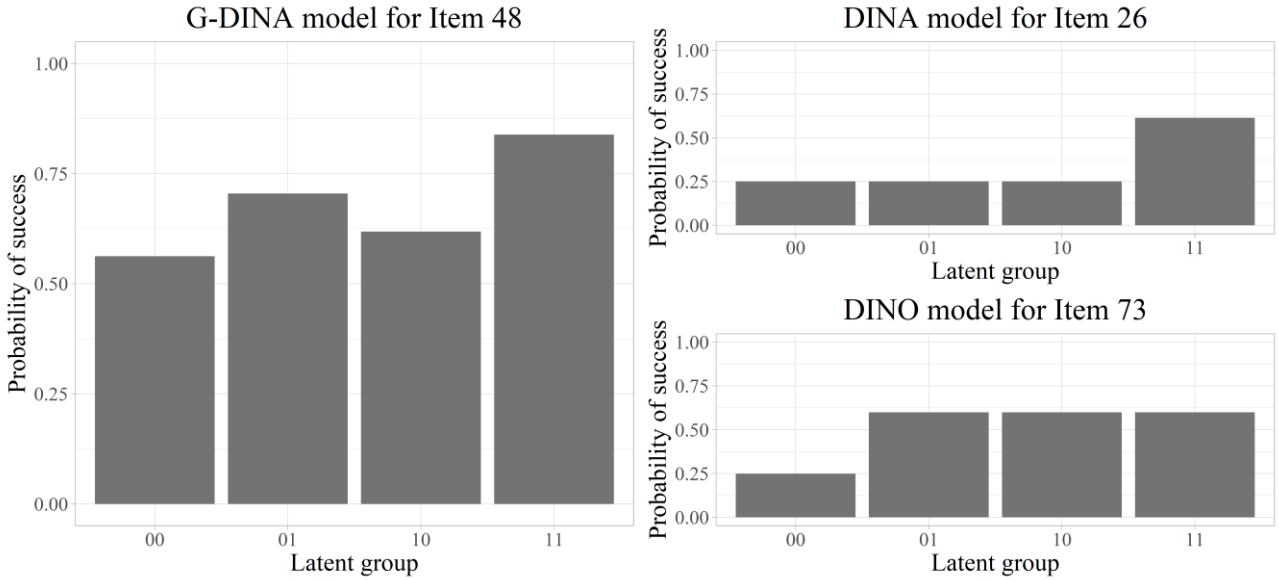
Attribute	Description
A1	Understand and apply the philosophical and ethical principles of human rights, the New Mexican School, and the constitutional legislation and secondary laws of the 2019 Educational Reform.
A2	Recognize the components of the National Educational System and the aspects, contents, and characteristics of the current educational model.
A3	Understand and apply current educational regulations, manuals, protocols, and operational agreements for schools and zones.
A4	Recognize the delineation and scope of the functions and capacities of teachers as public servants and educational agents.
A5	Understand and apply theoretical principles and pedagogical methods, intervention, and dialogue with the school community to improve learning achievement in the classroom for all students, considering their interests and characteristics.
A6	Design, plan, and implement strategies for information gathering, educational assessment, and feedback to detect and address special needs and improve students' educational achievement.
A7	Understand the characteristics of students' cognitive, emotional, moral, physical, social, and cultural development, and the importance of considering them in educational work.
A8	Understand the basic principles and concepts of inclusion, diversity, interculturality, and educational equity, as well as design and implement strategies to address learning barriers and participation.
A9	Understand the principles and conceptual notions of the classroom climate, school coexistence, and psychosocial risk factors in the classroom, and design and implement strategies for effective prevention and intervention.
A10	Understand the conditions, causes, and circumstances of child sexual abuse, bullying, and mistreatment in schools, as well as any situation that puts students' integrity at risk, and design and implement strategies for effective prevention and intervention.
A11	Autonomously manage the detection of professional development needs and evaluation of teaching practice and know and implement the continuous training resources and strategies available for their attention.
A12	Promote dialogue, reflection, and collaboration among students, parents, peers, members of the Technical and Zone Councils, and other members of the school community for professional development and the improvement of teaching practice, as well as for the achievement of learning and educational quality.

The DINA model establishes a non-compensatory process, where the probability of a correct response is high only when the examinee has mastered all the attributes measured by the particular item. Specifically, regardless of K^*_j, the DINA model estimates only two parameters per item: the probability of an incorrect response for those who have mastered the required attributes, slip parameter (s_j), and the probability of a correct response for those who have not

mastered the required attributes, guessing parameter (g_i). Due to its simplicity and applicability in the field of optimal performance measurement in educational settings, the DINA model is the most popular in available studies (Sessoms & Henson, 2018). The DINO model is equally parsimonious but reflects a "reverse" compensatory process, where mastering at least one of the attributes is sufficient to achieve a high probability of a correct response. Thus, only individuals in the group $\boldsymbol{\alpha}_{ij}^{*} = \{00\}$ will have a low probability of a correct response. An illustration of these two models is provided in Figure 1.

Figure 1

Example of the G-DINA, DINA, and DINO Item Response Function for Items 48, 26, and 73 using the Final 56x9 Q-Matrix.



It is possible to relax these constraints significantly by estimating a different probability of a correct response for each latent group. This results in a saturated model known as the generalized DINA model (G-DINA; de la Torre, 2011), which can better fit the item response function to the data, provided that the sample size is sufficient (Sorrel et al., 2021). The G-DINA model explains the success probabilities of each latent group using parameters that capture both the additive effects of individual attributes and their interactions. As shown in Figure 1, each latent group can have a different probability of success, reflecting both the main effect of mastering each attribute and the interaction effects. In some cases, these parameters can be constrained to produce the DINA model, in which only individuals who have mastered all attributes have a high probability of success, or the DINO model, in which mastering just one attribute is sufficient for a high probability of success, also illustrated in Figure 1. In this example, only two parameters are estimated rather than four: guessing, the probability of success for individuals who do not meet the condition, and slip, the probability of failure for individuals who do meet the condition.

Review of the Empirical Applications

The term *retrofitting* (Gierl & Cui, 2008; Liu et al., 2018) has been used to describe the application of CDMs to assessments originally developed using frameworks such as classical test theory or item response theory. Despite the growing interest in CDMs, empirical applications without retrofitting remain rare. According to a review by Sessoms and Henson (2018) of CDM literature from 2009 to the that time, out of all the articles found, over half (51% of 74 papers) were simulation-based. The majority evaluated a new development (e.g., a new model or index) and included an applied data analysis only as an illustration of the application of the proposed development. Thus, less than half of the CDM literature of the time were *real* empirical applications. The authors pointed out that only 8% of the studies used the results obtained from CDMs. Moreover, they are often supplementary to simulation studies and typically "retrofitted" to unidimensional tests, even though it is documented that retrofitting can severely impact model fit (Gierl et al., 2010).

It is worth emphasizing that teachers are an essential element in education, as research shows that variables related to teachers are best predictors of students' outcomes (García-Bacete & Rosel-Remírez, 2021; Moreno-Murcia et al., 2024). In recent years, there has been a growing interest in evaluating and centering teachers as important roles in education in general in Spanish speaking contexts (Alonso-Tapia et al., 2020). Thus, this study continues the growing international importance of teacher assessment for educational admission and improvement.

The utility of CDM can only be reliably determined if they are applied in real-world contexts and if any issues arising in those contexts are explored. Although a few years have passed since Sessoms and Henson (2018) and interesting empirical applications, such as Ren et al. (2021), have begun to emerge, they are still anecdotal. This study presents results for a test designed within the CDM framework to assess high school teachers in Mexico on a large scale. This allows for the continuation of this line of work by utilizing these high-stakes data in an educational context.

The Present Study

The purpose of this paper is to illustrate how advanced methodologies developed in recent years can help solve problems that arise in real application contexts. Demonstrations of the R (R Core Team, 2024) functions are included in a didactic manner, so that this paper serves as a complement to previous tutorials on CDMs (Shi et al., 2021), but omitting general aspects covered in these tutorials and highlighting those uniquely addressed in this work. This paper aims to serve as a guide on the key considerations involved in developing and applying CDMs from the ground up. One aspect to consider is that CDMs have traditionally been treated as predominantly confirmatory; that is, once the Q-matrix generated by experts is available, it is typically applied to the data without further scrutiny. In the proposed workflow, assessing the quality of the data and the fit of the models serves as a necessary prerequisite. Only then do we examine the relationships between items and attributes, with the final step considering that, if the attributes are highly correlated, a higher-order model may account for the multidimensional structure. This paper also emphasizes the importance of consulting with experts. To ensure clarity, the exposition of the methods is presented alongside their practical applications, thus, the issues and possible solutions are presented in both the Method and Results sections that follow.

Method

Participants

The sample used in this study was composed of 8,321 participants who were teacher candidates to the Mexican public national education system at the high school level. The sample was distributed among all 32 Mexican states. Because the sample was made up of applicant candidates to the educational system, it was a convenience sample. The test was applied in the summer of 2021. 54.09% of participants were women and 45.90% were men. Participants had a mean age of 34.13 years (SD = 6.91). We filtered out the examinees who had more than 4 items with atypical time responses. The cut-off criteria established to clean the database focused on the latency of answers that were less than 15 seconds or more than 250 seconds per item.

Instrument

The test consisted of 100 multiple-choice items, each with four answer options. The initial Q-matrix of the test measured 12 attributes as shown in Table 2. A CDM framework was used to design, apply, evaluate, and analyze this high school teacher assessment in Mexico. This assessment was developed through collaboration between educational measurement and psychometric experts and high school teachers. The test design was carried out by the relevant stakeholders, not by the authors of the present study. Although administered in 2021 and originally designed with the CDM framework in mind, the test was scored according to classical test theory, using a 0-to-100 scale after recoding the multiple-choice items. As will be shown, the CDM framework enables examinees to be classified as either possessing or not possessing each attribute, while also allowing for rank-ordering based on a continuous theta score standardized as is common in traditional item response theory.

The table of specifications was based on the content specifics adjacent to the professional teacher profile, which encompasses different domains. These domains address the professional identity, curriculum mastery, instructional planning and implementation, collaborative participation within the school community, and the ongoing development necessary for enhancing teaching performance. Each domain is composed of criteria and indicators that outline what teachers of the New Mexican School need to know and be able to do (Secretaría de Educación Pública & Unidad del Sistema para la Carrera de las Maestras y los Maestros, 2019).

Procedure

The test was administered online with a 3-hour time limit, allowing participants to complete it from any location on any computer. Instructions required participants to keep their cameras on, with audio recorded. They were informed that suspicious behavior, such as cheating or having another person in the frame, would result in sanctions, including potential cancellation of the exam. A human team evaluated flagged behaviors to determine if they were fraudulent. In total, 21 participants were detected engaging in fraudulent behavior. The Unidad del Sistema para la Carrera de las Maestras y Maestros oversaw the recruitment and application of the test and provided the data.

Data analysis

The analysis was conducted using R; the packages used were *cdmTools* (Nájera et al., 2024) and *GDINA* (Ma & de la Torre, 2020). To facilitate a more didactic understanding, the explanation of the R functions used for each issue is presented alongside the corresponding results. Specifically, the results will be presented in six sections. The first section will discuss overall results, while the remaining five sections will be labeled as issues. The issues are presented in the order of

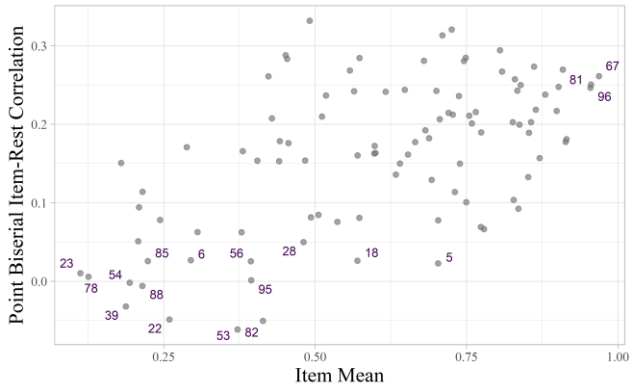
appearance within the usual analysis sequence, as indicated, for example, by Shi et al. (2021). These issues might be commonly encountered in empirical applications; thus methodological solutions to these issues will be proposed. The data were analyzed using the DINA model, following the guidance of the experts who participated in the test development, which indicated that mastering all the attributes involved was necessary to correctly answer the items. Additionally, the fit of the G-DINA and a proposed mixed model (which refers to the combination of two or more models) will be explored to provide a more comprehensive assessment.

Results

Overall Results

The lowest score was 0 and the highest score was 87, with a mean of 60.3 (*SD* = 9.34). At least 99% of the examinees answered each item. Figure 2 shows information on classical test theory item-level indicators, such as item mean and point biserial item-rest correlation (r_{pbis}) for difficulty and discrimination, respectively. The mean of all items is .60 (*SD* = .23), which is an acceptable mid-range of difficulty. The mean of the item-rest correlation for all items is .16 (*SD* = .09), which is somewhat low.

Figure 2
Classical Test Theory Statistics.



Note. The items marked are low-quality items in both item mean ($.20 < \bar{X}_j < .90$) and/or item-rest correlation ($r_{pbis} = .15$).

Before proceeding with the issues, a depuration of low-quality items was conducted to ensure a high-quality test. We filtered out items with r_{pbis} lower than .15 and item mean lower than .20 or higher than .90. In total, 44 items were eliminated, resulting in a Q-matrix of 56x12. Thus, the average item mean of all 56 items is .65 (*SD* = .15), resulting in a slightly easier test. The mean of the item-rest correlation for all 56 items is .22 (*SD* = .05). Consistency of these results was verified against an exploratory factor analysis conducted with oblimin rotation, extracting the number of factors recommended by the parallel analysis (10).

Issue 1: Q-Matrix Not Identified

The Q-matrix, including its associated model parameters, must be identified to ensure a reliable and valid estimation and inference. A statistical model is said to be identified if its parameters can be uniquely determined from the data. This means that each set of parameter values leads to a distinct pattern in the observed data. When a model is not identified, different parameter combinations can produce the same results, making it impossible to obtain unique and reliable estimates. Gu and Xu (2021) describe two types of joint identifiability in cognitive diagnostic models: strict and generic. Strict identifiability requires strong conditions, summarized as completeness, distinctness, and repetition, to ensure that each parameter can be estimated uniquely. Completeness means that each attribute must be measured individually by at least one item. Distinctness requires that every combination of attributes represented in the Q-matrix is different. Repetition implies that each attribute must appear in at least three items. When these requirements are too restrictive for practical applications, a weaker notion called generic identifiability can be used. Generic identifiability relaxes the conditions, requiring that each attribute is measured by at least two items (generic completeness) and appears at least once outside these main items (generic repetition). The main difference is that strict identifiability guarantees uniqueness in all possible scenarios, while generic identifiability ensures it in almost all practical cases, allowing for some rare exceptions. The cdmTools::is.Qid function can be used to check if a Q-matrix fulfills these conditions for identifiability, as shown in Figure 3.

As shown in the code output, the initial 100x12 Q-matrix failed to meet the completeness condition and lacked both strict and generic identifiability. The high-quality 56-item matrix also lacks most conditions. To address this issue, attributes 1, 3, and 11 were removed because they did not meet the completeness condition, as none had a single-attribute item. Therefore, after resolving Issue 1, we obtained a 56x9 Q-matrix, referred to as the modified Q-matrix, which is identified.

It is important to mention that it is normal for there to be a gap between specialized new research and empirical applications. This Q-matrix was developed in 2021, as was Gu and Xu (2021)'s study. This may help to understand why their guidelines were not met in the initial Q-matrix. Although the removal of attributes and items is not ideal in terms of content validity, we have found that it has indeed allowed us to meet the necessary identifiability requirements to proceed with the remaining analyses. We strongly recommend future applications to consider this identifiability requirement since the initial construction of the Q-matrix in conjunction with content of the test to avoid this issue.

Figure 3
Identifiability Code and Results.

```
>R Qid_initial <- cdmTools::is.Qid(Q_100x12, model = "DINA")
>R print(Qid_initial)
Model = DINA

Identifiability conditions:
A) Completeness        = FALSE
B) Distinctiveness     = TRUE
C) Repetition          = TRUE

Strict identifiabilty  = FALSE
Generic identifiabilty = FALSE

>R Qid_56x12 <- cdmTools::is.Qid(Q_56x12, model = "DINA")
>R print(Qid_56x12)
Model = DINA

Identifiability conditions:
A) Completeness        = FALSE
B) Distinctiveness     = TRUE
C) Repetition          = FALSE

Strict identifiabilty  = FALSE
Generic identifiabilty = FALSE

>R Qid_56x9 <- cdmTools::is.Qid(Q_56x9, model = "DINA")
>R print(Qid_56x9)
Model = DINA

Identifiability conditions:
A) Completeness        = TRUE
B) Distinctiveness     = TRUE
C) Repetition          = TRUE

Strict identifiabilty  = TRUE
Generic identifiabilty = TRUE
```

Note. In this figure, as well as in the subsequent ones that involve code, the R prompt '>R' is included to distinguish between code and output. To copy and execute the code in R, the symbol '>R' should be removed.

Issue 2: Q-Matrix Misspecification and Ambiguous Attribute Definitions

Classification problems arise largely because of Q-matrix misspecifications, as they negatively affect the estimation of the model parameters, potentially leading to classifying examinees into inaccurate latent classes (Gao et al., 2017; Rupp & Templin, 2008). We explored two approaches to examine this.

The first approach was to conduct a revalidation of the Q-matrix by forming an expert committee. This committee should specialize in measurement and evaluation, contain experts in the field being evaluated in the test, and an expert in CDMs, or at least someone with a basic understanding of the models. For this study, a committee of 6 experts in educational measurement was formed. Although one member dropped out in the second phase of the revalidation, it did not impact the process. The committee evaluated the Q-matrix, contrasting their opinions with the initial Q-matrix, and justifying any discrepancies. Their goal was to ensure that the Q-matrix specification accurately measured the intended attributes for each item. In total, they proposed 23 changes in 14 items to the Q-matrix. This 56x9 Q-matrix will be referred to as modified Q-matrix.

They also observed and commented on the ambiguity of the description of each attribute. For example, they found that attributes 8 and 9 had overlapping content, making it difficult to determine which attribute certain items measured. Additionally, attribute 12 had "smaller attributes" within it, so they recommended re-writing it to improve the understanding of the evaluative focus. Specifically, attribute 12 can be decomposed into: 1) promote dialogue, reflection, and collaboration among educational agents for professional development and the improvement of teaching practice and 2) promote dialogue, reflection, and collaboration among educational agents for the achievement of learning and educational quality. However, the committee did not provide detailed suggestions to improve the Q-matrix or items.

The second solution we implemented was to conduct an empirical Q-matrix validation using the Hull method in conjunction with the proportion of variance accounted for (PVAF) and an attribute-test-level iterative implementation following results of Nájera et al. (2021). This empirical method aims to correct the potential misspecifications a Q-matrix might have. This is achieved by accounting for the complexity of each candidate q-vector for each item. This method should always be used hand in hand with the opinion of the expert committee, complementing each other. The Q-matrix validation can be run using the cdmTools::valQ function as presented in Figure 4. The input was the G-DINA model estimation with the modified Q-matrix (56x9).

Figure 4
Empirical Q-Matrix Validation Code and Results.

```
>R valQ_9att <- cdmTools::valQ(gdina_9att_modified, index =
"PVAF", iterative = "test.att")
Iteration = 001 | Item/s modified = 17, 30, 31, 39, 51

>R valQ_9att$n.iter
[1] 1
>R mean(valQ_9att$sug.Q == Q_56x9_mod)
[1] 0.9900794
```

The Hull method proposed changes to only five items, resulting in a Q-matrix that was 99% like the initial one. Overall, 19 changes to the initial Q-matrix were agreed upon by both the expert committee and the Hull method, comprising 6 additions and 13 eliminations. Whenever the two sources were different, it was up to the authors to integrate both proposals. Table 3 provides examples of these changes. For Item 49, modifications were made to align with recommendations from both the expert committee and the Hull method. In contrast, for Item 34, both sources agreed on removing Attribute 8 but differed on whether to add Attribute 5. Ultimately, only Attribute 12 was assigned, based on theoretical considerations. For Item 91, the expert commit-

tee's suggestions aligned with the Hull method, and the final integration incorporated two attributes. Therefore, after solving Issues 1 and 2, we have a Q-matrix of 56x9, which we will refer to as final Q-matrix.

Table 3
Examples of Q-Matrix Suggestions.

Item	Initial Q-matrix	Expert committee	Hull method	Integration (final Q-matrix)
34	A8, A12	A5*, A12	A12	A12
49	A6, A8	A2*, A5*, A8	A2, A5, A8	A2, A5, A8
91	A5	A5, A7	A5, A7	A5, A7

Note. *At least one expert expressed uncertainty about the necessity of the attribute.

Issue 3: Good Fit Always

Another issue encountered was that the model demonstrated good fit regardless of the Q-matrix or model used, which makes it difficult to clearly identify the underlying response process of the test. The model fit can be obtained with the GDINA::modelfit function, as shown in Figure 5 for the final Q-matrix and DINA model. Note that to obtain fit indices of the other models, the GDINA.obj argument must be each model to evaluate.

Figure 5
Model Fit Code and Results.

```
>R fit_dina_9att_final <- GDINA::modelfit(dina_9att_final)
>R print(fit_dina_9att_final)
Test-level Model Fit Evaluation

Relative fit statistics:
 -2 log likelihood =  527049.9  (number of parameters =  623)
 AIC  =  528295.9  BIC =  532673.3
 CAIC =  533296.3  SABIC =  530693.6

Absolute fit statistics:
 M2 =  2705.253  df =  973  p =  0
 RMSEA2 =  0.0146  with  90 % CI: [ 0.014 , 0.0153 ]
 SRMSR =  0.0229
```

As seen in Table 4, the initial 100x12 Q-matrix had a good fit under the G-DINA and DINA models. This should not happen because, according to the revalidation process, this Q-matrix is misspecified and not identified. The final 56x9 Q-matrix also presents good fit, regardless of the model. The model configurations including mixed and higher-order models will be discussed later in Issue 4. The only statistic indicating poor fit is the M_2 statistic. While this statistic has shown reliable results for assessing fit in Liu et al. (2016), it is noteworthy that this study involves many items, variables, and a large sample size, all factors documented to increase the likelihood of model rejection (Maydeu-Olivares & Joe, 2006).

Table 4
Goodness of Fit of All Models and Q-Matrices Tested.

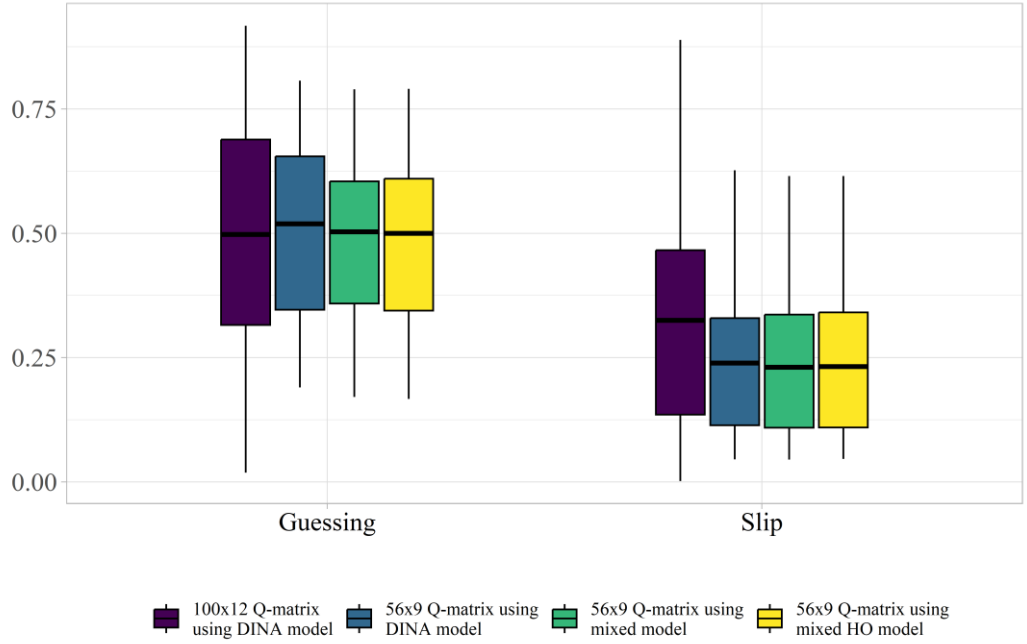
	Relative fit		Absolute fit		
	AIC	*BIC*	M_2	$RMSEA_2$	*SRMSR*
	Initial 100x12 Q-matrix				
G-DINA	**901877.2**	**932898.8**	1316.62**	**.011**	**.020**
DINA	903457.8	933581.7	1510.75**	**.011**	.021
	Modified 56x9 Q-matrix				
G-DINA	**527613.1**	**532440.2**	2640.26**	**.015**	**.022**
DINA	528324.5	532701.9	2854.452**	**.015**	.023
	Final 56x9 Q-matrix				
G-DINA	527474.6	532273.7	2543.058**	**.015**	**.022**
DINA	528295.9	532673.3	2705.253**	**.015**	.023
Mixed	527468.5	532239.4	2586.184**	**.015**	**.022**
Higher-order Mixed	**527112.4**	**528419.3**	3969.599**	**.015**	.023

Note. **$p < .001$. The model with the better fit for each Q-matrix is highlighted in bold. AIC = Akaike information criterion; BIC = Bayesian information criterion; $RMSEA_2$ = Root mean square error of approximation. SRMSR = Standardized root mean squared residual.

Another aspect to mention is the quality of the items according to the guessing and slip parameters. To explore this, the distribution of g_i and s_j for each of the model configurations discussed in this article is shown in Figure 6. It can be observed that the initial configuration (DINA model with 100x12 Q-matrix) exhibits the highest guessing and slip values, indicating poorer-quality items. Since the test was refined by eliminating low-quality items, many of which were difficult, the average guessing probability remained similar, although the slip was successfully reduced.

What the other model configurations (DINA with final 56x9 Q-matrix, mixed model, or mixed model with higher order) show is that item quality remains consistent regardless of the specific models applied. It is worth noting in this case, as discussed later, that mixed models and higher-order models aim to simplify the parameter space, making the stability of item quality a desirable outcome.

Figure 6
Guessing and Slip Parameters Across Models.



Note. The middle line in the boxplot here represents the mean, not the median. HO = Higher-order.

Issue 4: Multiple Models for the Test: Underlying Cognitive Process

One aspect of validity that has historically received somewhat less attention is validity evidence based on the response process. Under CDM, this can be empirically addressed using item-level model comparison statistics (Ma et al., 2016; Ravand & Robitzsch, 2018). As previously mentioned, the experts involved in the test construction indicated that the response process was non-compensatory, meaning that all attributes required by an item must be mastered to answer it correctly. Although it is reasonable to expect, based on the global assessment, that the DINA model would be suitable for each item, using a single model for all test items generally does not accurately reflect reality (Sorrel et al., 2017a). It is important to note that for items measuring a single attribute, all models are equivalent. Differentiating between models is only relevant when an item measures more than one attribute. In this test, 22 items are multi-attribute.

We propose two approaches to this issue. Firstly, we applied the two-step likelihood ratio (2LR) test to determine which CDM model is the most appropriate at the item level (Sorrel et al., 2017b). This test provides model suggestions for the items using the largest p-value rule at the .05 alpha level. The R code shown in Figure 7 demonstrates how to implement this test. The arguments specify which reduced CDMs are possible for each item. In this case, only DINA and DINO were included to replicate the options presented by the expert committee, as described below.

Figure 7
Model Comparison Code and Results.

```
>R mc <- GDINA::modelcomp(gdina_9att_final, method = "LR",
models = c("DINA", "DINO"), LR.args = list(LR.approx =
TRUE))
>R mc$selected.model[which(mc$selected.model$models !=
"GDINA"), ]
       models pvalues adj.pvalues
Item 3    DINA 0.1822      0.3645
Item 41   DINO 0.2142      0.3645
```

The second solution involved having an expert committee assess the cognitive processes underlying each multi-attribute item during the revalidation phase. It is important to emphasize that these two solutions should always complement each other. The experts assessed whether each item was based on a compensatory process (where mastering only one required attribute was sufficient to answer correctly, as in the DINO model), a non-compensatory process (where mastering all required attributes was necessary to answer correctly, as in the DINA model), or another mechanism (more general, as in the G-DINA model).

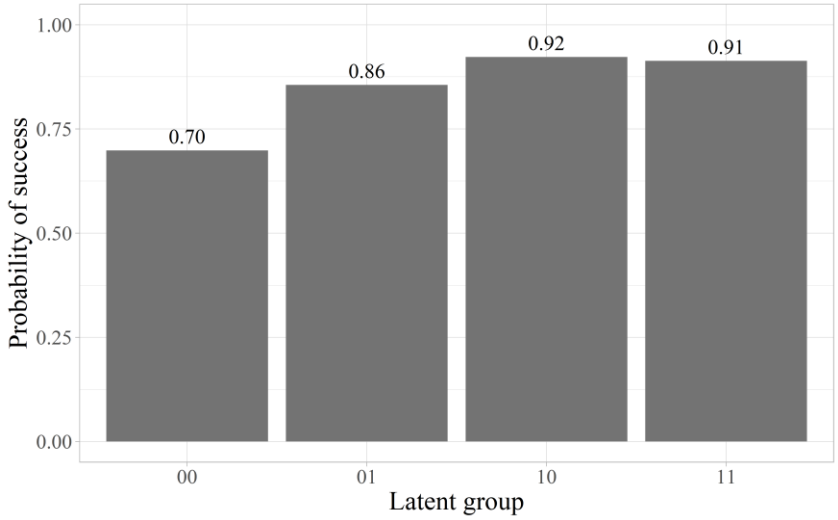
Table 5 illustrates the results for both solutions, the 2LR test and the expert committee. Overall, the 2LR test suggested one item used the DINO model and another the DINA model. Whenever the two sources were different, it was up to the authors to integrate both proposals, guided by the estimated item success probabilities. For example, examining the item probabilities of Item 74, as in Figure 8, we can see the similarity between the bars in latent groups $\{01\}, \{10\}$ and $\{11\}$, which explains why the LR test suggests the DI-

NO model. It is to note that the probability of answering correctly is always more than .50, necessitating caution. This case demonstrates why the LR test should always go hand in hand with the suggestions by the expert team. The integration of the suggestions will be referred to as mixed model, which contains the final 56x9 Q-matrix obtained in Issues 1-4, with the models differing by item. In this mixed model, all other items are modeled with the G-DINA model.

Table 5
Item-Level Model Suggestions in the 2LR Test.

Item	2LR test	Expert committee	Integration (Mixed model)
9	DINA	DINA	DINA
74	DINO	DINA	DINO

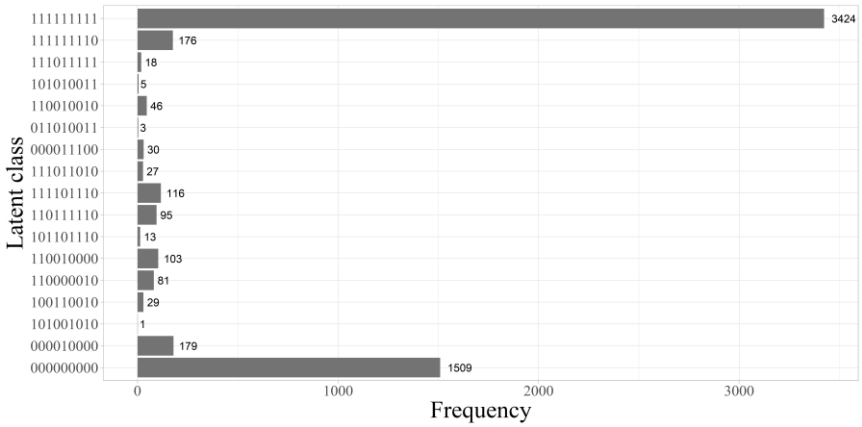
Figure 8
Item Success Probabilities for Item 74 (Final 56x9 Q-Matrix).



Issue 5: Extreme Latent Class Categorization

Another issue we encountered was that the latent classes in which there had been the most examinees classified were the extremes, {000000000} and {111111111}, while the intermediate latent classes had significantly fewer examinees (see Figure 9). This is an issue reported to happen when retrofitting data to a CDM model. Moreover, attributes are highly correlated with each other. These high correlations could indicate that the attributes may not be statistically distinct enough to require diagnostic information at the attribute level and may instead reflect a continuous trait (Ma et al., 2020). These results suggest that the test may not measure all 9 attributes, may assess them poorly, or that CDM may not be the best approach, despite not being a retrofitting study. To address this issue, we propose an exploratory approach using parallel analysis to assess whether the test is unidimensional or not.

Figure 9
Mixed Model Classification of Examinees (Final 56x9 Q-Matrix).



Note. Only some notable latent classes are shown, since there are hundreds of classes ($2^K = 512$).

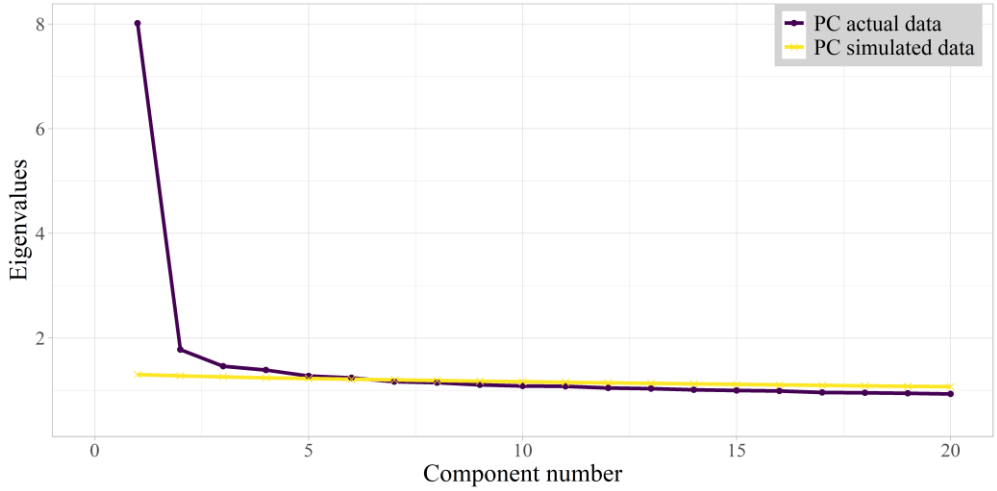
Figure 10
Dimensionality Analyses Code and Results.

```
>R out_pa <- cdmTools::paK(dat = data_56, fa = "pc")
>R out_pa$sug.K
[1] 6

>R mcK <- cdmTools::modelcompK(dat = data_56, exploreK =
1:9, stop = "AIC", val.Q = FALSE, verbose = TRUE)
>R mcK.56$sug.K[c("AIC", "BIC")]
AIC BIC
 8   5
```

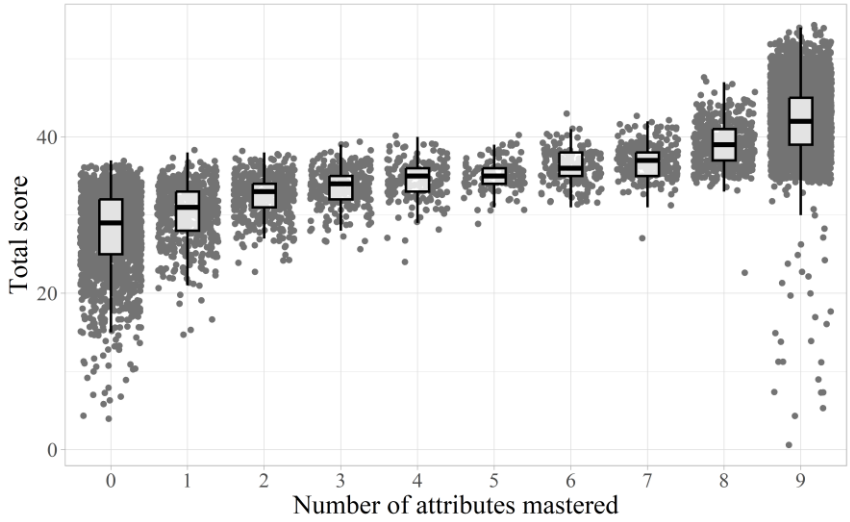
The parallel analysis can be implemented using the function cdmTools::paK, as shown in Figure 10. The parallel analysis conducted using the reduced dataset with 56 items suggests 6 components underlying the test. Furthermore, when we take a closer look at the scree plot in Figure 11, one predominant component arises clearly. There is a large gap between the first and second component. This suggests that the test may be unidimensional, or a higher-order CDM can be a better approximation. Another option available is the cdmTools::modelcompK function, which determines the number of attributes underlying the model using model fit comparison, as shown in Figure 10. The analysis using the reduced dataset with 56 items suggests 8 and 5 attributes underlying the test (according to AIC and BIC, respectively). Integrating the results from both functions, it is suggested that the 56-item dataset contains between 5 and 8 components, with one particularly clear component that accounts for a significant portion of the variance.

Figure 11
Parallel Analysis Solution for 56-Item Test.



To test a higher-order approximation, a higher-order model (de la Torre & Douglas, 2004) was fitted, with a mixed model approach, as seen in Issue 4. Under this model, nine discrete attributes were estimated, and a continuous latent trait was added to explain the correlations between attributes, representing general ability through a two-parameter logistic model. The correlation between estimates of higher-order ability and total scores in the test was .91. As seen in Figure 12, in general, as the total score increased, so did the number of attributes mastered. The extremes of 0 and 9 attributes mastered have the most clustered individuals. These results support the idea of a strong underlying component. However, despite the high correlations, we can clearly observe that the distribution of the total score and number of attributes mastered does not follow a clear and perfect relation, indicating that diagnostic information at the attribute level may be necessary for this test. The examinees who master all attributes overlap with many who have zero attributes in the test score. That is, two individuals with the same test score may have very different attribute patterns. This overlap cannot be explained with classical test theory or unidimensional item response theory alone, and this is where CDMs can help us obtain more detailed diagnostic information.

Figure 12
Distribution of Test Total Score and Number of Attributes Mastered using the Higher-Order Mixed Model (Final 56x9 Q-Matrix).



This higher-order mixed model addresses all five issues and is considered the final model, endorsed by the authors. This is consistent with the model fit results in Table 4, where the higher-order mixed model demonstrates the best fit, based on AIC and BIC.

Discussion

Artificial data commonly used in simulation studies can differ substantially from real data. Therefore, it is essential to study the performance of statistical methods on empirical data. In response to the growing field of the use of CDM, this paper analyzes what can go wrong in a CDM application from the start and provides a source of evidence of potential issues and solutions to consider during the construction of a CDM application. This paper reports five issues encountered and provided possible solutions or explanations, including the instructions to apply these analyses in R. The aim is to complement previous applications in a way that facilitates the emergence of more empirical studies (Ma et al., 2020; Ravand & Robitzsch, 2018).

The first issue was the lack of identifiability in the initial Q-matrix, resolved by meeting the conditions for strict and generic identifiability (Gu & Xu, 2021). As a result of this analysis, three of the twelve initial attributes were removed. During the revision of this manuscript, we explored whether the factor structure obtained through exploratory factor analysis, considering that parallel analysis suggested fewer components, could provide insights into potential attribute merging. Although this investigation did not yield conclusive results in the present case, it underscores an interesting avenue for future research, particularly the use of metrics such as Tucker's congruence coefficients (Lorenzo-Seva & Ten Berge, 2006) to guide attribute merging. More importantly, in this regard, it is crucial to take into account the identification conditions detailed here (completeness, distinctness, and repetition; Gu & Xu, 2021) to avoid model identifiability issues. Test developers can check these requirements for preliminary Q-matrices using the R package cdmTools (Nájera et al., 2024). The second issue involved Q-matrix misspecification and ambiguous attribute definitions, which were addressed by revising the Q-matrix with an expert committee and validation techniques. The third issue showed good fit across model configurations, but closer inspection revealed that the final model, a combination of different models including a higher-order distribution, had better fit, though caution was needed due to low-quality items. The fourth issue focused on identifying the best model at the item level through expert evaluation of cognitive processes and the two-step likelihood ratio test. We emphasize that, in line with other empirical applications and the available research (de la Torre et al., 2018; Ravand & Robitzsch, 2018; Sorrel et al., 2021), it becomes evident that a single model is unlikely to fit all items. Therefore, item-level model selection, combined with expert evaluation, emerges as a solution to this challenge. The fifth issue addressed dimensionality, where extreme latent categorization suggested one dimension, later integrated into a higher-order model.

The results indicate that the test lacks a strong conceptual framework, leading to poor-quality attributes, ambiguous content, and Q-matrix misspecification. It is possible the test tried to force something mnemonic to more complex competencies and skills, as, for example, memorizing the General Law of Education does not count as an attribute. In other words, the test might be measuring knowledge instead of proficiency. We recommend that test developers using CDM frameworks in future research establish a solid theoretical foundation to avoid such issues. Although in the reduced dataset of 56 items the rank-ordering of examinees based on classical test theory (sum scores) and the higher-order trait estimated under the CDM shows a high degree of similarity

(r = 0.945), they do not match completely. Moreover, the CDM offers the added advantage of allowing personalized interventions tailored to individual attribute profiles, targeting areas that have not yet been mastered (Ren et al., 2021).

This study includes several limitations. First, the expert committee did not modify or redesign problematic items or attributes, which should be addressed in future applications. While the main focus of the article is on exploring solutions through modeling and the analysis of an existing database, it is important to highlight the value of complementing this approach with qualitative analysis, involving experts or discussion groups, to better understand why certain items showed poor quality. In most contexts, understanding why items perform poorly provides the added benefit of allowing them to be reformulated and retained. After eliminating 44 low-quality items and resolving Issues 1 and 2, the Q-matrix was not revised to reincorporate the content of the removed items and attributes. We emphasize that Q-matrix revalidation should include multiple rounds of a feedback process between content experts and methodological experts, and that any modification to the Q-matrix or the instrument should be justified by both empirical evidence and theoretical considerations. The experts could also play a role in integrating the Q-matrix changes at a broader level, for instance, by re-defining the evaluated attributes. Considering that many items were removed, it is appropriate to examine whether item removal substantially affects participant scores. When shortening a test, one way to assess equivalence between the reduced (56 items) and the original scale (100 items) is to calculate the correlation between their total scores. In this dataset it was .945, indicating that the ranking of individuals remains virtually unchanged. This correlation was consistent across sex and age (partial r controlling for age = .945; controlling for sex = .944), suggesting no differential impact of item removal. Additionally, a more comprehensive evaluation of the test's behavior could be conducted, such as a differential item functioning analysis across variables like sex and age. For categorical variables, this can be further explored through modeling with CDM (e.g., Ma et al., 2021). Overall, this study aims to guide the development of better empirical applications, and we recommend more frequent involvement of content experts who follow the statistical analyses presented here.

A key advantage of CDMs is their capacity to provide diagnostic feedback on teachers' specific strengths and weaknesses. When comparing the different models explored (DINA, DINO, mixed, higher-order) for the final Q-matrix with 9 attributes, we found that although there appears to be a high level of agreement in person–attribute classifications (ranging from 89% to 94%), these percentages still translate into thousands of discrepancies: out of 74,880 possible decisions (8,320 persons × 9 attributes), the models differ in approximately 5,000 to 8,000 cells. The comparison between mixed and higher-order mixed shows the greatest similarity, with 93.5% agreement in person–attribute classifications. Nonetheless, this still represents around 4,800 discrepancies out of 74,880 decisions, indicating that, although both models tend to produce highly consistent results, the choice between them can still meaningfully affect the assignment of attributes for a considerable number of individuals. Beyond supporting admission decisions to the public education system, CDMs can offer detailed information on teachers' mastery of individual attributes, not just overall scores, thereby guiding targeted professional development aligned with the priorities of the New Mexican School. Model refinement is essential for reliable attribute assignments and informed educational decisions. This study identifies potential issues that may arise when constructing a test using the CDM framework. Therefore, it is crucial for a CDM-based test to incorporate solid theoretical foundations to prevent model and Q-matrix misspecification. While there are many factors to consider and manage, CDMs offer a promising approach in assessment methodologies. At the same time, the results emphasize the need to further explore the performance of the developed statistics (e.g., fit, Q-matrix validation, person fit, etc.) under conditions that more closely resemble real-world situations.

Complementary information

Disclosure statement.- The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding.- The last author was supported by MICIU/AEI/10.13039/501100011033 and ERDF/EU under the project "Computerized adaptive tests based on new assessment formats" (reference: PID2022-137258NB-I00) and the UAM IIC Chair on Psychometric Models and Applications.

Data Research Availability Statement.- The data used for this study are available upon reasonable request to the corresponding author.

Acknowledgements.- During the writing of this paper, the second and third authors acknowledge the support of the Unidad del Sistema para la Carrera de las Maestras y los Maestros (USICAMM). The authors would also like to thank the expert committee in the revalidation process and Daniela Peralta for helping the organization of the revalidation process.

References

Alonso-Tapia, J., Ruiz, M. Á., & Huertas, J. A. (2020). [Differences in classroom motivational climate: causes, effects and implications for teacher education. A multilevel study] Diferencias en el clima motivacional en el aula: causas, efectos e implicaciones para la formación docente. Un estudio multinivel. *Annals of Psychology*, *36*(1), 122–133. https://doi.org/10.6018/analesps.337911

Chen, Y., Culpepper, S. A., & Chen, Y. (2023). Bayesian inference for an unknown number of attributes in restricted latent class models. *Psychometrika*, *88*(2), 613–635. https://doi.org/10.1007/s11336-022-09900-7

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J., & Sorrel, M. A. (2023). Cognitive diagnosis models. In F. Ashby, H. Colonius, & E. Dzhafarov (Eds.), *New Handbook of Mathematical Psychology* (Cambridge Handbooks in Psychology, pp. 385–420). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108902724.010

de la Torre, J., van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*, *51*(4), 281–296. https://doi.org/10.1080/07481756.2017.1327286

Gao, M., Miller, M. D., & Liu, R. (2017). The impact of Q-matrix misspecification and model misuse on classification accuracy in the generalized DINA model. *Journal of Measurement and Evaluation in Education and Psychology*, *8*, 391–403. https://doi.org/10.21031/epod.332712

García, P. E., Olea, J., & de la Torre, J. (2014). Application of cognitive diagnosis models to competency-based situational judgment tests. *Psicothema*, *26*(3), 372–377. https://doi.org/10.7334/psicothema2013.322

García-Bacete, F. J., & Rosel-Remírez, J. F. (2021). Spanish validation of the Questionnaire on Teacher Interaction in the upper grades of primary education (QTI-P) and how this interaction influences academic performance. *Annals of Psychology*, *37*(1), 101–113. https://doi.org/10.6018/analesps.415111

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research and Perspectives*, *6*(4), 263–268. https://doi.org/10.1080/15366360802497762

Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing*, *10*, 318–341. https://doi.org/10.1080/15305058.2010.509554

Gu, Y., & Xu, G. (2021). Sufficient and necessary conditions for the identifiability of the Q-matrix. *Statistica Sinica*, *31*, 449–472. https://www.jstor.org/stable/26969691

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. https://doi.org/10.1177/01466210122032064

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement*, *78*(3), 357–383. https://doi.org/10.1177/0013164416685599

Liu, Y., Tian, W., & Xin, T. (2016). An Application of M2 Statistic to Evaluate the Fit of Cognitive Diagnostic Models. *Journal of Educational and Behavioral Statistics, 41*, 3–26. https://doi.org/10.3102/1076998615621293

Lorenzo-Seva, U., & Ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*, 57–64. https://doi.org/10.1027/1614-2241.2.2.57

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*(14), 1-26. https://doi.org/10.18637/jss.v093.i14

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40*(3), 200–217. https://doi.org/10.1177/0146621615621717

Ma, W., Minchen, N., & de la Torre, J. (2020). Choosing between CDM and unidimensional IRT: The proportional reasoning test case. *Measurement: Interdisciplinary Research and Perspectives*, *18*(2), 87–96. https://doi.org/10.1080/15366367.2019.1697122

Ma, W., Terzi, R., & de la Torre, J. (2021). Detecting differential item functioning using multiple-group cognitive diagnosis models. *Applied Psychological Measurement*, *45*(1), 37–53. https://doi.org/10.1177/0146621620965745

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713–732. https://doi.org/10.1007/s11336-005-1295-9

Moreno-Murcia, J. A., Huéscar Hernández, E., León, J., Fin, G., Nodari Júnior, R. J., Valero-Valenzuela, A., Tristán, J., Gastélum-Cuadras, G., Zueck Enríquez, M. C., Vargas Vitoria, R., Cid, L., Monteiro, D., & Teixeira, D. (2024). [Motivation to learn: an international multilevel study on student autonomy and teacher emphasis on content usefulness] Motivación para aprender: un estudio internacional multinivel sobre la autonomía de los estudiantes y el énfasis de los docentes en la utilidad del contenido. *Annals of Psychology*, *40*(2), 265–271. https://doi.org/10.6018/analesps.571161

Nájera, P., Sorrel, M. A., & Abad, F. J. (2024). *cdmTools: Useful Tools for Cognitive Diagnosis Modeling*. R package version 1.0.5. https://cran.r-project.org/web/packages/cdmTools/

Nájera, P., Sorrel, M. A., de la Torre, J., & Abad, F. J. (2021). Balancing fit and parsimony to improve Q-matrix validation. *British Journal of Mathematical and Statistical Psychology*, *74*, 110–130. https://doi.org/10.1111/bmsp.12228

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ravand, H., & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: a study of reading comprehension. *Educational Psychology, 38*(10), 1255–1277. https://doi.org/10.1080/01443410.2018.1489524

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*. https://doi.org/10.3389/fpsyg.2021.628607

Rupp, A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*, 78–96. https://doi.org/10.1177/0013164407301545

Sanz, S., Kreitchmann, R. S., Nájera, P., Moreno, J. D., Martínez-Huertas, J. Á., & Sorrel, M. A. (2023). FoCo: A Shiny app for formative assessment using cognitive diagnosis modeling. *Psicología Educativa. Revista de los Psicólogos de la Educación*, *29*(2), 149–158. https://doi.org/10.5093/psed2022a14

Secretaría de Educación Pública & Unidad del Sistema para la Carrera de las Maestras y los Maestros. (December 2019). [*Framework for Excellence in Teaching and School Management in Upper Secondary Education: Professional Profiles, Criteria, and Indicators for Teachers, Technical Teaching Staff, and Personnel with Leadership and Supervisory Functions. School Year 2020–2021*] *Marco para la excelencia en la enseñanza y la gestión escolar en la Educación Media Superior: Perfiles profesionales, criterios e indicadores para docentes, técnicos docentes y personal con funciones de dirección y de supervisión. Ciclo Escolar 2020-2021.*

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

Shi, Q., Ma, W., Robitzsch, A., Sorrel, M. A., & Man, K. (2021). Cognitively diagnostic analysis using the G-DINA model in R. *Psych*, *3*(4), 812–835. https://doi.org/10.3390/psych3040052

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. (2017a). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement*, *41*, 614–631. https://doi.org/10.1177/0146621617707510

Sorrel, M. A., de la Torre, J., Abad, F. J., & Olea, J. (2017b). Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 13*(Suppl 1), 39–47. https://doi.org/10.1027/1614-2241/a000131

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and Reliability of Situational Judgement Test Scores: A New Approach Based on Cognitive Diagnosis Models. *Organizational*

Research Methods, *19*(3), 506–532. https://doi.org/10.1177/1094428116630065

Tan, Z., de la Torre, J., Ma, W., Huh, D., Larimer, M. E., & Mun, E. Y. (2022). A tutorial on cognitive diagnosis modeling for characterizing mental health symptom profiles using existing item responses. *Prevention Science*, *24*, 480–492. https://doi.org/10.1007/s11121-022-01346-8

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354. https://doi.org/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. https://doi.org/10.1037/1082-989X.11.3.287