Variable-Length Cognitive Diagnostic Computerized Adaptive Testing in Small-Scale Assessments

Pablo Nájera[1], Miguel A. Sorrel[2], Chia-Yi Chiu[3], and Francisco J. Abad[2]

[1] Department of Psychology, Universidad Pontificia Comillas

[2] Department of Social Psychology and Methodology, Universidad Autónoma de Madrid

[3] Teachers College, Columbia University

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

Nájera, P., Sorrel, M. A., Chiu, C.-Y., & Abad, F. J. (2025). Variable-length cognitive diagnostic computerized adaptive testing in small-scale assessments. *Journal of Educational and Behavioral Statistics*. Advance online publication. https://doi.org/10.3102/10769986251366581

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/10.3102/10769986251366581.

Author Note

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

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

Chia-Yi Chiu https://orcid.org/0000-0002-3919-2025

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

This work is part of the research project "Avances en los modelos de diagnóstico cognitivo para la evaluación formativa" ["Advances in cognitive diagnosis modeling for formative assessments"; PP2024_20] from the 2024 Call for Funding of Internal Research Projects, financed by Universidad Pontificia Comillas. This research was also funded by MICIU/AEI/10.13039/501100011033, ERDF/EU under the project "Computerized adaptive tests based on new assessment formats" (reference: PID2022-137258NB-I00), the Community of Madrid through the Pluriannual Agreement with the Universidad de Universidad Autónoma de Madrid in its Programa de Estímulo a la Investigación de Jóvenes Doctores (Reference SI3/PJI/2021-00258), and the UAM IIC Chair on Psychometric Models and Applications. Portions of these findings were presented at the 8th Conference of the International Association for Computerized Adaptive Testing (September 2022; Frankfurt, Germany) and the 10th European Congress of Methodology (July 2023; Ghent, Belgium). The simulation codes of this research are publicly available at https://osf.io/mn86s. This study was not preregistered. We have no conflict of interest to disclose.

Corresponding concerning this article should be addressed to Miguel A. Sorrel, Department of Social Psychology and Methodology, Faculty of Psychology, 6 Iván Pavlov St, Madrid, Spain, 28049. Email: miguel.sorrel@uam.es

Abstract

This study examines innovative procedures for cognitive diagnostic computerized adaptive testing (CD-CAT) in small-scale assessments. Traditional CD-CAT methods, based on parametric cognitive diagnostic models (CDMs), often struggle with small calibration samples, leading to overfitting and overestimated reliability. Nonparametric alternatives, while more robust in small-scale settings, lack reliability information, limiting classification certainty and variable-length adaptive testing. To address these challenges, we propose four CD-CAT procedures using the parsimonious *restricted deterministic input, noisy "and" gate* (R-DINA) model, a parametric CDM tailored for small samples. Two of these procedures use a calibration sample (R-GDI and R-NPS), while the other two are calibration-free methods (R-NPS_{ML} and R-NPS_{BM}). Through a simulation study, where calibration sample size, number of attributes, and item quality were manipulated, we compare these methods to the conventional CD-CAT based on the DINA model. Results indicate that R-GDI and R-NPS consistently outperform the conventional CD-CAT in terms of more accurate posterior probability recovery, classification accuracy, and balanced item usage, although they administer a larger number of items. The calibration-free methods also perform satisfactorily but exhibit reliability overestimation with low-quality items. Overall, the proposed procedures offer practical solutions for formative assessments in educational contexts characterized by small sample sizes and time constraints. We provide recommendations for the use and scalability of these methods in real educational settings.

Keywords: cognitive diagnosis, computerized adaptive testing, nonparametric classification, classification accuracy, reliability estimation

Variable-Length Cognitive Diagnostic Computerized Adaptive Testing in Small-Scale Assessments

Cognitive diagnosis modeling (CDM) is a family of statistical models that has gained popularity in educational research as a tool for evaluating whether examinees have mastered a series of *attributes*. This detailed information helps identify students' strengths and weaknesses, which can later guide targeted remedial instruction. CDM, therefore, aligns with the growing interest in alternative formats of educational evaluation, such as formative assessments (de la Torre & Minchen, 2014; Paulsen & Valdivia, 2021). Beyond education, CDM has been also applied to other areas such as clinical psychology (e.g., Templin & Henson, 2006) and organizational psychology (e.g., Sorrel et al., 2016). Attributes are defined as discrete (usually dichotomous) latent variables that represent skills, competences, or psychological processes required to endorse a series of test items. The primary output provided by CDM is the *attribute profile classifications*. For example, consider a test measuring $K = 3$ attributes. Here, $\widehat{\boldsymbol{\alpha}}_i = \{1,0,0\}$ denotes that examinee i has been classified in the attribute profile $\{1, 0, 0\}$, meaning that they have mastered the first attribute but not the second or third. Beyond this categorical classification, CDM has undergone numerous methodological developments in recent decades, making it a comprehensive psychometric framework capable of extracting rich information, including reliability estimates, relative and absolute model fit evaluation, and differential item functioning, among other features (see von Davier & Lee, 2019, for a comprehensive review).

The area of CDMs is a growing field, driven by novel theoretical proposals that allow for the modeling of different types of data (e.g., dichotomous, polytomous, continuous) while considering various aspects, such as multiple strategies for responding to the same item or the hierarchical structure of attributes, among others (for an introduction to recent developments in

CDMs, see, for example, de la Torre & Sorrel, 2023). These modeling capabilities are often associated with high parameterization, which entails a cost in terms of the sample size required to obtain reliable classifications. This is compounded by the fact that these theoretical advances have not yet been widely translated into practice, as highlighted by Sessoms and Henson (2018) in their critical review of existing empirical applications. Considering the specific area of formative assessment in education, it is common for practitioners to face situations with small sample sizes and limited time availability (Paulsen & Valdivia, 2021; Ren et al., 2021). This has led, on the one hand, to the most frequently applied model being a relatively simple one (Sessoms & Henson, 2018), such as the *deterministic, input, noisy "and" gate* (DINA; Junker & Sijtsma, 2001), which only requires two parameters per item. Nonparametric approaches offer an additional solution, as they can potentially generate classifications without requiring a calibration sample or with a very small sample size (Chiu et al., 2018). Regarding time constraints, proposals have been developed to improve assessment efficiency, with the implementation of computerized adaptive testing standing out as a key solution. These tests adjust to the respondent's level during the test, allowing for comparable classification reliability while reducing the number of items (Chang et al., 2019; Sorrel et al., 2021). This area of applying CDMs in computerized adaptive testing has come to be known as cognitive diagnostic computerized adaptive testing (CD-CAT; Cheng, 2009).

As detailed below, while simulation studies have shown that it is possible to obtain reliable classifications using these nonparametric solutions and their application in CD-CAT, they have generally focused on stopping criteria based on the number of items rather than stopping once a desired reliability threshold is achieved, which could enhance efficiency. For example, although 30 items might be administered, if after 20 items the posterior probability that the individual belongs to the assigned latent class is already very high (e.g., greater than .80), the test can stop.

This reduction in the number of items frees up time for instruction, addressing the demand for CDM applications to be implemented in time-constrained environments. The present article aims to develop a solution for this variable-length application of CD-CAT, using the *restricted DINA* (R-DINA) model (Nájera, Abad, et al., 2023) as a starting point. The reason for selecting this model is that, as discussed in this article, it allows operation with small calibration samples, and even without a calibration sample, providing classifications equivalent to those of nonparametric procedures while incrementally incorporating the available information.

The remainder of the paper is organized as follows. First, a brief overview of parametric CDM is provided. Second, diagnostic procedures for small-scale assessments are described, including nonparametric CDM and the R-DINA model. Third, an introduction to CD-CAT is presented. Fourth, we elaborate on our proposal to integrate the R-DINA model into CD-CAT, detailing the four different procedures developed for this purpose. Fifth, the performance of the proposed procedures is tested and compared to that of the traditional CD-CAT by means of a Monte Carlo simulation study. Finally, a discussion section is included that summarizes the main conclusions, limitations, and future research lines, as well as practical recommendations.

A Review of Parametric CDM

For CDM to classify examinees into attribute profiles, three inputs are required. First, the responses of the N individuals to J items. These responses are typically dichotomous, indicating correct or incorrect answers, although various response formats have been explored in the literature (e.g., W. Ma & de la Torre, 2016; Gao et al., 2020). Second, a *Q-matrix*, which acts as a bridge between the J items and the K attributes. The Q-matrix is usually constructed by domain experts (see Sorrel et al., 2016), who determine which items measure which attributes. For example, with $K = 3$ attributes, $\boldsymbol{q}_j = \{1,0,1\}$ represents the q-vector of item j, indicating that it measures the first

and third attributes, but not the second. In addition to the expert judgment, several Q-matrix estimation and validation (e.g., de la Torre & Chiu, 2016; Nájera, Sorrel, et al., 2021) methods may assist in the Q-matrix specification process from an empirical perspective. Third, as a family of statistical models, CDM can adopt several *item response functions*, which reflect how attributes interact to produce a correct or incorrect response to an item. Two well-known reduced models include the *deterministic input, noisy "and" gate* (DINA) model (Junker & Sijtsma, 2001) and the *deterministic input, noisy "or" gate* (DINO) model (Templin & Henson, 2006). The DINA model assumes a non-compensatory response function, also referred to as conjunctive, which implies that an examinee must master all the attributes involved in an item to endorse it. Conversely, the DINO model assumes a compensatory (or disjunctive) response function, meaning that mastering only one of the attributes measured by an item is sufficient to answer it correctly. The conjunctive/disjunctive nature of the DINA and DINO models is reflected in the *ideal response*:

$$\eta_{lj}^{(c)} = \prod_{k=1}^{K} \alpha_{lk}^{q_{kj}}, \tag{1}$$

and

$$\eta_{lj}^{(d)} = 1 - \prod_{k=1}^{K} (1 - \alpha_{lk})^{q_{jk}}, \tag{2}$$

where $\eta_{lj}^{(c)}$ and $\eta_{lj}^{(d)}$ denote the conjunctive (i.e., DINA) and disjunctive (i.e., DINO) ideal response of examinees in latent class l to item j, respectively, α_{lk} is the attribute k mastery level of examinees in latent class l, and q_{jk} indicates whether item j measures attribute k. These ideal responses are binary and deterministic, but the DINA and DINO models are probabilistic. This means that there is a probability of correctly answering an item for those examinees who are expected to fail (i.e., $\eta_{lj} = 0$), and a probability of failing the item for those who are expected to succeed (i.e., $\eta_{lj} = 1$).

These probabilities are captured by the *guessing* (g_j) and *slip* (s_j) parameters, respectively, resulting in the following item response function:

$$P(y_j = 1|\boldsymbol{\alpha}_l) = {g_j}^{1-\eta_{lj}}(1 - s_j)^{\eta_{lj}}, \quad (3)$$

where $\eta_{lj} = \eta_{lj}^{(c)}$ and $\eta_{lj} = \eta_{lj}^{(d)}$ for the DINA and DINO models, respectively. Equation 3 implies that, in the DINA and DINO models, there are only two parameters (s_j and g_j) per item. These parameters differentiate between examinees expected to provide either a correct ($\eta_{lj} = 1$) or incorrect ($\eta_{lj} = 0$) response to the item. The similarities between the DINA and DINO models are such that they are equivalent under certain transformations (Köhn & Chiu, 2016). Moreover, the DINA model is the most widely used CDM in applied settings (Sessoms & Henson, 2018).

The DINA and DINO models are special cases of the *generalized DINA* (G-DINA) model (de la Torre, 2011), which is a saturated model in that it estimates a different item probability of success for every possible latent group. This makes the G-DINA a more flexible model than DINA and DINO. However, this increased flexibility comes with the trade-off of requiring larger sample sizes to ensure stable and accurate estimation of item and person parameters (Sorrel et al., 2021). This is a significant limitation, especially considering that even the reduced DINA and DINO models require sample sizes of at least 500 individuals to yield accurate parameter estimates (Sen & Cohen, 2021).

CDM for Small-Scale Assessments

The dependency on large sample sizes may be a practical issue, given that one of the most promising applications of CDM is in small-scale assessments. Specifically, the detailed diagnostic feedback provided by these models, delivered in a timely manner, can directly inform remedial instruction or learning efforts at a classroom level (de la Torre & Minchen, 2014; Paulsen & Valdivia, 2021). In this vein, some real CDM applications have been conducted with sample sizes

as small as $N = 105$ (Ren et al., 2021) or even $N = 44$ (Jang et al., 2015). Despite these few examples, the limited number of CDM applications (Sessoms & Henson, 2018) may partially be due to the fact that most methodological developments are not well-suited to practical settings that deal with small sample sizes.

To address this issue, Chiu and Douglas (2013) proposed the *nonparametric classification* (NPC) method, a deterministic procedure that classifies examinees into latent classes without relying on parameter estimation. Namely, the NPC method compares the ideal responses (see Equation 1 and Equation 2) of all possible latent classes with the examinee's observed responses using the Hamming distance, as follows

$$d_h(\boldsymbol{y}_i, \boldsymbol{\eta}_l) = \sum_{j=1}^{J} |y_{ij} - \eta_{lj}|, \tag{4}$$

where η_{lj} can be $\eta_{lj}^{(c)}$ or $\eta_{lj}^{(d)}$ for a conjunctive (i.e., DINA) or disjunctive (i.e., DINO) rules, respectively. Examinees are then classified into the most similar latent class: $\widehat{\boldsymbol{\alpha}}_i = \arg\min_l d_h(\boldsymbol{y}_i, \boldsymbol{\eta}_l)$. The main benefit of the NPC method is that, by not relying on parameter estimation, it provides more accurate attribute profile classifications in settings where the available information for estimation is scarce or poor, such as with small sample sizes or low-quality items (Chiu & Douglas, 2013; Chiu et al., 2018). However, this practical advantage comes with a significant limitation: the inability to assess crucial psychometric properties such as reliability or model fit. Consequently, a practitioner conducting a small-scale assessment with CDM would face a dilemma: either use the NPC method and accept its classifications without additional information on the adequacy of the results, or use a parametric CDM (e.g., DINA model), knowing that it may provide less accurate classifications in suboptimal sample conditions.

To address this, Nájera, Abad, et al. (2023) proposed the *restricted DINA* (R-DINA) model, which is a parametrization of the NPC method. This means that the R-DINA model can provide the same exact attribute profile classifications as the NPC method but, more importantly, allows for the computation of reliability and fit indices to assess its psychometric properties. Specifically, the NPC method can be parametrized as a restricted version of the DINA or DINO model, where the guessing and slip parameters for all items are constrained to have the same value:

$$P\left(y_j = 1 \middle| \boldsymbol{\alpha}_l\right) = \varphi^{1-\eta_{lj}}(1-\varphi)^{\eta_{lj}}, \tag{5}$$

where $\eta_{lj} = \eta_{lj}^{(c)}$ and $\eta_{lj} = \eta_{lj}^{(d)}$ for the R-DINA and R-DINO models, respectively, and φ represents the overall proportion of observed responses that differ from their corresponding ideal responses. Note that the R-DINA model has only one parameter for the entire model: $\varphi = g_j = s_j \forall j$. Despite its over-restrictive nature, the R-DINA model has shown robust performance when its assumptions are violated, outperforming the DINA model in terms of classification accuracy, item parameter recovery, and reliability estimation accuracy under very small sample sizes (N = 25 to 100), even when the generating model was DINA (Nájera, Abad, et al., 2023).

Cognitive Diagnostic Computerized Adaptive Testing

Despite these advancements, small-scale assessments often require longer tests to mitigate the lack of information from the limited number of examinees. However, longer tests require more time to complete, which can limit the feasibility of using these models for continuous formative assessments throughout an academic year (Chang et al., 2018; Paulsen & Valdivia, 2021). A well-established psychometric development that enhances the efficiency, accuracy, and security of assessments is *computerized adaptive testing* (CAT). In CAT, a large item bank is initially calibrated using a large sample size. Once the item parameters are calibrated, each examinee receives a tailored test, with items presented based on their previous responses. When integrated

within the CDM framework, this approach results in *cognitive diagnostic CAT* (CD-CAT; Cheng, 2009).

Since its introduction, CD-CAT has undergone significant advancements, including the adaptation and formulation of various *item selection rules*, as well as the consideration of different *test stopping criteria* and *content restrictions*. The item selection rule refers to the algorithm used to determine the most optimal item to administer to an examinee at a given time, based on the calibrated item parameters and the examinee's responses to previous items. Some widely-explored item selection rules are the *general discrimination index* (GDI; Kaplan et al., 2015), the *Jensen-Shannon divergence index* (Kang et al., 2017), and the *posterior-weighted Kullback-Leibler index* (Cheng, 2009) and its modified version (Kaplan et al., 2015). These rules rely on previously calibrated item parameter estimates and the examinee's responses to select the most discriminative item at a given time. In this study, GDI will be used as a representative of parametric item selection rules, as it is expected to perform similarly to other parametric rules and demonstrates computational efficiency (Kaplan et al., 2015; W. Wang et al., 2020). It is a popular rule implemented in open-access software, such as the R package 'cdcatR' (Sorrel et al., 2022). The GDI is defined as

$$GDI_j = \sum_{l=1}^{L} \pi(\boldsymbol{\alpha}_l)^{(t)} \left[P\left(y_j = 1 \middle| \boldsymbol{\alpha}_l\right) - \bar{P}_j\right]^2, \tag{6}$$

where $\pi(\boldsymbol{\alpha}_l)^{(t)}$ is the posterior probability of latent class l at time t, $P\left(y_j = 1 \middle| \boldsymbol{\alpha}_l\right)$ is the probability of a correct response for latent class l on item j, and $\bar{P}_j$ is the weighted average success probability for all latent classes on item j. The item to be administered at time $t + 1$ is the one with the maximum GDI.

The main issue with all these parametric item selection rules is that, as mentioned earlier, a large calibration sample is usually required to accurately estimate item parameters. To address this practical concern, two nonparametric item selection rules based on Hamming distances have been recently proposed. Chang et al. (2018) developed the *nonparametric item selection* (NPS) rule, which is directly based on the NPC method, while Li and Zheng (2024) proposed the *nonparametric dynamic binary searching item* (NDBS) rule building upon the work in binary searching algorithms (Tatsuoka & Ferguson, 2003; Zheng & C. Wang, 2017). For the sake of simplicity, in this study we will primarily focus on the NPS as a representative nonparametric item selection rule, given its availability in open-access software ('cdcatR' package; Sorrel et al., 2022) and the similar performance of both rules found in an auxiliary analysis presented in the Online Appendix. The NPS rule begins by administering K items in a Q-optimal manner (Xu et al., 2016), which ensures the distinguishability between latent classes. Once the first K items have been administered, at time t, the NPC method is used to calculate the Hamming distance between the examinee's observed responses and the ideal response patterns of the latent classes. This process defines $\widehat{\boldsymbol{\alpha}}_i$ as the most likely latent class (i.e., the one with the lowest Hamming distance) and $\widetilde{\boldsymbol{\alpha}}_i$ as the second most likely latent class (i.e., the one with the second lowest Hamming distance) for examinee i. The NPS rule then randomly selects the next item to be administered from those that elicit a different ideal response for $\widehat{\boldsymbol{\alpha}}_i$ and $\widetilde{\boldsymbol{\alpha}}_i$. By discriminating between the two most likely attribute profiles at a given time, the NPS rule aims to increase the gap between the most likely attribute profile and the others (Chang et al., 2018). Compared to parametric item selection rules, a key practical advantage of the NPS rule is that it is calibration-free, as it does not rely on a calibration sample to estimate item parameters. In other words, items can be used directly in an

adaptive assessment with the NPS rule without needing to be calibrated with a different sample beforehand.

Test stopping criteria refer to the rule used to conclude the CAT for each examinee. They can be broadly divided into *fixed-length* and *variable-length* criteria. In fixed-length CD-CAT, a prespecified number of items is administered to all examinees. While this approach offers notable advantages in terms of efficiency compared to traditional paper-and-pencil assessments, using the same test length for all examinees might be suboptimal. Namely, it can lead to either *inaccurate* assessments (i.e., the test stops before an accurate classification has been made for a particular examinee) or *inefficient* assessments (i.e., the test continues even though an accurate classification has already been made for a particular examinee). In contrast, variable-length criteria allow for administering a different number of items to each examinee, with the test stopping when the desired level of classification certainty is achieved for a particular test taker. Typically, the examinees' latent class posterior probabilities are considered for these criteria. For example, a common variable-length criterion is to stop the assessment when the examinee's maximum latent class posterior probability exceeds a cutoff of .80. This variable-length criterion, here referred to as $c = .80$, implies that 80% of the examinees are expected to be classified into the correct latent class. Thus, this stopping rule, originally introduced by Tatsuoka (2002), has the additional advantage of serving as an estimate of reliability. Another possibility examined in Hsu et al. (2013) is that not only must the largest latent class posterior probability meet or exceed a prespecified value (e.g., .80), but also the second largest latent class posterior probability must not exceed a prespecified value (e.g., .10). Naturally, as the threshold c increases, the differences between these two approaches disappear, which has led open-access software like 'cdcatR' (Sorrel et al., 2022) to default to the simpler rule that the largest latent class posterior probability exceeds .80. There

have also been proposals based on information theory (Guo & Zheng, 2019). In general, these rules tend to require an estimation of the latent class posterior probability. Recently, Li and Zheng (2024) proposed the *nonparametric dynamic binary searching index* (NDBI), which, aligned with the rationale of the NDBS item selection rule and based on binary searching algorithms and Hamming distances, enables nonparametric variable-length CD-CAT (details about the NDBS and NDBI are provided in the Online Appendix). Note that, unlike traditional stopping criteria based on posterior probabilities (e.g., $c = .80$ corresponds to an expected classification accuracy of .80), rules like NDBI do not have this direct translation into reliability. Available studies show that, in fact, this rule can lead to high attribute classification accuracy (Li & Zheng, 2024), but it becomes challenging to associate the score obtained in a specific case with a concrete accuracy estimate. This is a drawback, as score interpretation should be guided by reliability (AERA, APA, & NCME, 2014).

For these reasons, despite the significant advances that have been made, the application of variable-length CD-CAT still poses some practical challenges. On the one hand, estimating latent class posterior probabilities relies on item parameter estimates, which require large sample sizes to be accurately calibrated (Sun et al., 2020). On the other hand, nonparametric approaches available to date (i.e., NPS and NDBS) do not provide posterior probability estimates. This highlights the need to develop an alternative that addresses these two issues.

Lastly, beyond the strictly mathematical aspects, some authors have emphasized the importance of the validity argument in adaptative testing. Specifically, in the context of CDM, content (i.e., attribute) balance is considered an important aspect of test construction (Henson & Douglas, 2005). This relates to the model identifiability problem, where the number of items measuring each attribute is significant (Gu & Xu, 2021). In response, content restrictions have

been incorporated into CD-CAT procedures, either by including content balance as a feature in the item selection rule (e.g., Cheng, 2010; Sun et al., 2021) or by directly imposing the restriction that each attribute must be measured by a minimum number of items (see Cheng et al., 2007).

Integrating the R-DINA Model into CD-CAT

In this section, we explain three different approaches to integrate the R-DINA model within the CD-CAT framework. Two of these procedures resemble traditional implementations because they follow the established process of calibrating model parameters using a calibration sample before conducting the adaptive assessment. The other method is a novel implementation that, consistent with the calibration-free nature of the NPS, does not require a calibration sample. Instead, it estimates the R-DINA parameter on-the-fly for each examinee.

When a Calibration Sample is Accessible

As with any parametric CDM, the R-DINA model can be directly implemented within the traditional, parametric CD-CAT flowchart. This involves first estimating the φ model parameter using a calibration sample, and then using this information to conduct the adaptive testing with the desired parametric item selection rule, test stopping criterion, and content restriction. Compared to other models (e.g., DINA, G-DINA), the R-DINA model allows for the use of parametric CD-CAT even with a small calibration sample ($N < 200$; Nájera, Abad, et al., 2023). In the remainder of the paper, we will use the GDI item selection rule for this first CD-CAT implementation with a calibration sample, which will be referred to as R-GDI.

Despite being a parametric model, the R-DINA has a strong connection with the NPC method, as both procedures are equivalent in terms of attribute classifications due to the parallelism between parametric likelihoods and nonparametric Hamming distances (C. Ma et al., 2023; Nájera, Abad, et al., 2023). Consequently, the R-DINA model can be easily integrated into nonparametric

CD-CAT. Namely, after calibrating the φ model parameter using a (small) calibration sample, a CD-CAT using a nonparametric item selection rule (NPS or NDBS) can be normally conducted. The primary purpose of estimating φ is to compute examinees' posterior probabilities, thereby enabling variable-length assessments in nonparametric CD-CAT, which was not feasible with traditional methods. The NPS rule will be used in the remainder of the study, and thus this second implementation will be referred to as R-NPS.

Note that the R-GDI and R-NPS variants are expected to perform very similarly. If both procedures select the same items for a given examinee, the R-GDI and R-NPS will be equivalent in terms of attribute profile classification and posterior probability estimates, since the underlying model for these calculations is the same (i.e., the R-DINA model). However, the GDI and NPS rules might not always select the same items for a given response pattern due to two reasons. First, if more than one item meets the selection criteria for the GDI (i.e., maximum GDI) or the NPS (i.e., discriminates between the two most likely attribute profiles), then the next item to be administered is randomly selected among the eligible items. Second, while the NPS focuses solely on the point estimates of the two most likely attribute profiles, the GDI considers the posterior probability of all attribute profiles (Sorrel et al., 2020). Moreover, the NPS assumes that all attribute profiles are equally likely in the population, whereas the GDI uses the estimated attribute distribution to compute item discrimination (see Equation 6). These technical differences are not expected to significantly affect CD-CAT performance, as both item selection rules are anticipated to select appropriate items for each examinee throughout the assessment, leading to sound and efficient classifications. This, however, will be one of the questions explored in the simulation study.

When a Calibration Sample is Not Accessible

One of the main practical advantages of nonparametric item selection rules is their direct applicability without requiring a calibration sample (Chang et al., 2018). However, this comes at the significant cost of not providing information on reliability or enabling variable-length assessments. Leveraging the simplicity of the R-DINA model, we propose a calibration-free CD-CAT implementation that supports variable-length tests, thus combining the benefits of both parametric and nonparametric CD-CAT approaches.

This approach utilizes the NPS rule to select items for administration at each stage of the adaptive assessment, although note that the NDBS could be also used in the same fashion. The proposal involves using the R-DINA model to estimate the φ parameter on-the-fly at the examinee-level, meaning with $N = 1$. This parameter is then used to calculate the posterior probabilities of the latent classes, which in turn are used to determine when to stop the CD-CAT based on a variable-length stopping criterion. The pseudo-algorithm for this on-the-fly approach is as follows:

1) Administer K items according to the Q-optimal criterion (Xu et al., 2016) to ensure distinguishability among all latent classes. This starting rule is identical to the one used in the NPS method.
2) Estimate the R-DINA model using the examinee's responses to these items.
3) Calculate the posterior probabilities based on the estimated φ parameter.
4) If the variable-length stopping criterion (e.g., $c = .80$) is met, terminate the CD-CAT. Otherwise, use the NPS rule to select the next item to administer.
5) Repeat steps 2 to 4 until the stopping criterion is satisfied.

The straightforward implementation of this pseudo-algorithm involves estimating the R-DINA model using marginal maximum likelihood (ML) in a standard fashion (Nájera, Abad, et al., 2023), although with $N = 1$, and will henceforth be referred to as R-NPS$_{ML}$. The R-NPS$_{ML}$

implementation facilitates variable-length CD-CAT without requiring a calibration sample, by using the responses of each examinee as the calibration dataset. However, there is an important limitation to this approach. Although the R-DINA model is simple enough to provide accurate parameter estimates with small sample sizes, relying on only a few responses from a single examinee might provide insufficient data, potentially compromising the reliability of the estimate. It is likely that, after responding to only a few items, the observed response pattern of examinee i will perfectly match the ideal response pattern of latent class l. In such an overfitting scenario, the Hamming distance between the response patterns will be zero, leading to $\hat{\varphi}_i = 0$. This boundary problem, which has also been observed in other more complex CDMs under small sample conditions (Garre & Vermunt, 2006; Kreitchmann et al., 2023; W. Ma & Guo, 2019; W. Ma & Jiang, 2021), will lead to the posterior probability for latent class l being equal to 1. Consequently, the CD-CAT might terminate at a very early stage of the assessment, potentially resulting in a greatly overestimated reliability estimate.

To address this problem, we propose using a Bayes modal (BM) estimation algorithm. BM was first applied to CDM by W. Ma and Jiang (2021), who introduced it to overcome boundary issues in the G-DINA model when using ML estimation in small-scale scenarios. In ML estimation, the probability of success for latent class l on item j is estimated as $\hat{P}_j(\boldsymbol{\alpha}_l) = r_{jl}/n_l$, where n_l is the expected number of individuals in latent class l and r_{jl} is the expected number of correct responses among those individuals (de la Torre, 2011). Boundary problems are likely to occur when latent class l is sparse, causing the proportion of correct responses to skew towards 0 or 1. In contrast, BM estimation uses a Beta prior distribution, $\text{Beta}(\beta_1, \beta_2)$, to mitigate these extreme estimates. The BM estimation focuses on the mode of the posterior distribution, providing a single point estimate for each probability of success. The BM estimate is calculated as follows:

$$\hat{P}_j(\boldsymbol{\alpha}_l) = \frac{r_{jl} + (\beta_1 - 1)}{n_l + (\beta_1 + \beta_2 - 2)}. \quad (7)$$

W. Ma and Jiang (2021) used $\text{Beta}(1.5, 2.5)$ as the prior distribution of the guessing parameter, which improved the estimates under challenging conditions by mitigating boundary issues.

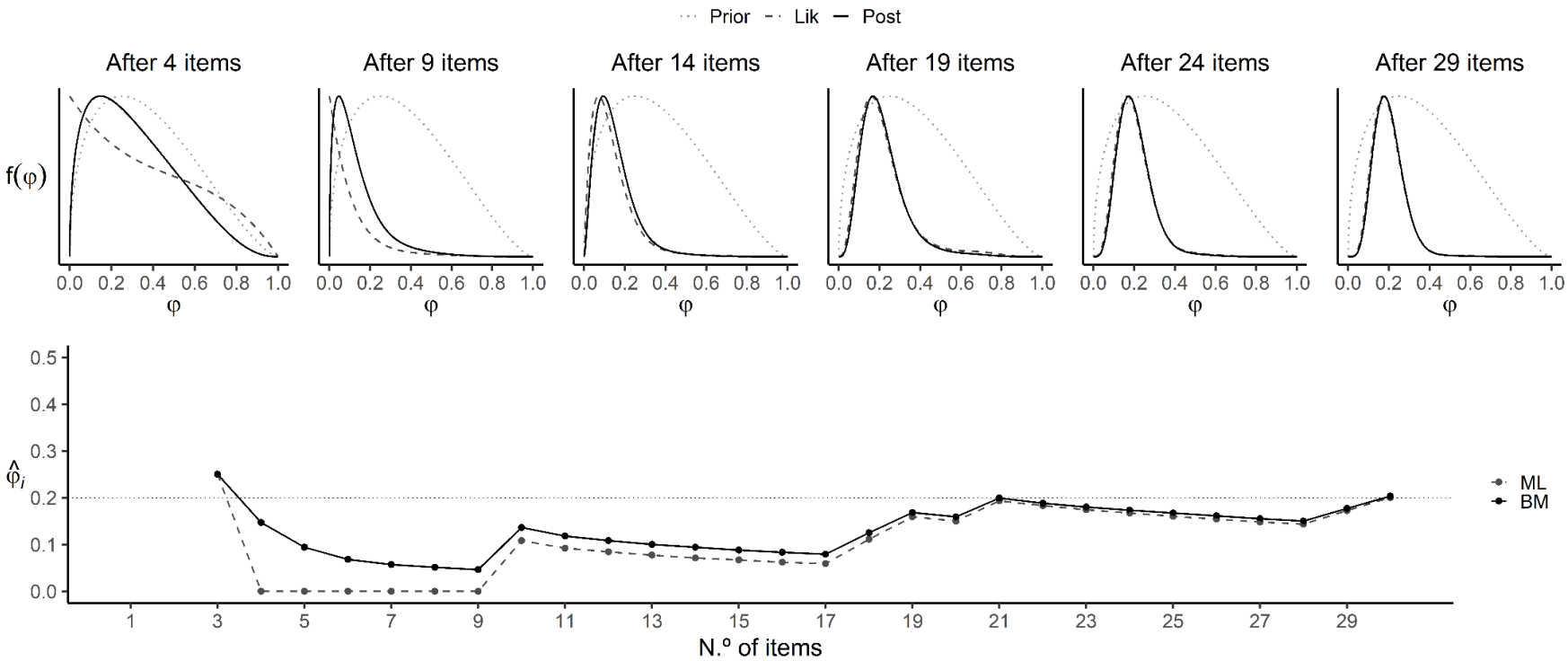
Following their work, we propose using a $\text{Beta}(1.5, 2.5)$ prior distribution for φ_i to prevent the CD-CAT from stopping prematurely before an accurate estimate has been reached. Under this approach, $\hat{\varphi}_i$ is defined as the mode of the posterior probability, obtained by combining the prior distribution with the likelihood function. This CD-CAT implementation of the R-DINA model, referred to as R-NPS$_{\text{BM}}$, is illustrated in Figure 1 alongside the R-NPS$_{\text{ML}}$ method. Note that BM estimation prevents premature stopping by adding stability to the estimate. As more items are administered, the likelihood function becomes more informative, and the influence of the prior diminishes. This approach ensures that the assessment is more robust and less likely to be halted due to early overfitting.

Simulation Study

The main goal of this study is to evaluate the performance of four different implementations of the R-DINA model within the CD-CAT framework. Two of these implementations require a (small) calibration sample (R-GDI and R-NPS), while the other two are directly applicable without a calibration step (R-NPS$_{\text{ML}}$ and R-NPS$_{\text{BM}}$). These procedures will also be compared to the traditional parametric CD-CAT based on a DINA-calibrated model and the GDI item selection rule (referred to here as the D-GDI procedure). The focus of the study will be on variable-length CD-CAT, given its abovementioned advantages over fixed-length CD-CAT. This approach will allow for a more appropriate assessment of the precision of the different procedures in terms of parameter estimates throughout the adaptive implementations, as a poor estimate might lead to CD-CAT stopping either prematurely (inaccuracy) or unnecessarily (inefficiency).

Figure 1

Illustration of the On-the-Fly Estimation of φ_i throughout an Adaptive Implementation of the R-DINA Model



Note. Prior = $\text{Beta}(1.5, 2.5)$ prior distribution (represented as a light gray dotted line); Lik = likelihood distribution (represented as a dark gray dashed line); Post = posterior distribution (represented as a solid black line); ML = maximum likelihood (represented as a dark gray dashed line); BM = Bayes modal (represented as a solid black line). The upper panel displays the prior, likelihood, and posterior distribution of $\hat{\varphi}_i$ after examinee i has taken a different number of items. Based on that information, the lower panel shows the value of $\hat{\varphi}_i$ under each moment in the CD-CAT application for both ML (the maximum of the likelihood distribution) and BM (the maximum of the posterior distribution). The dotted line in the lower panel represents the true, generating φ in this example.

Given the dual nature of the R-DINA model, which bridges the parametric DINA model and the nonparametric NPC method (Nájera, Abad, et al., 2023), we anticipate that the R-GDI and R-NPS will exhibit similar performance. We also expect these procedures to outperform the calibration-free alternatives (R-NPS$_{ML}$ and R-NPS$_{BM}$), as the φ parameter will be more accurately estimated, even with a small calibration sample. Additionally, R-NPS$_{BM}$ is expected to outperform R-NPS$_{ML}$ due to the latter's potential boundary issues during the early stages of adaptive testing.

CD-CAT Implementation

Six CD-CAT procedures are tested in the present simulation study: T-GDI, D-GDI, R-GDI, R-NPS, R-NPS$_{BM}$, and R-NPS$_{ML}$. Table 1 summarizes the characteristics of these implementations. As detailed later, the DINA model, given its popularity (Sessoms & Henson, 2018), was chosen as the data-generating model. The T-GDI procedure serves as an upper baseline, representing a parametric CD-CAT that uses the true, generating item parameters (i.e., there is no calibration error) in the computation of GDI (Equation 6). On the other hand, in D-GDI, the parameters will be estimated using a calibration sample. Smaller sample sizes in the calibration sample will lead to higher item calibration error, which will affect the performance of the CD-CAT procedure. All procedures employ a variable-length stopping criterion with a cutoff of $c = .80$ (i.e., the test stops once the maximum posterior probability for a latent class is equal to or higher than .80). As indicated earlier, we selected this rule because it is the simplest and is included by default in the available open-source software. As stated in the section on performance measures, the focus of the article is on determining whether the procedures operate under an estimated posterior probability close to the true estimated probability. As long as different stopping rules use this estimated posterior probability, the results are generalizable to other stopping rules. Additionally, content restrictions are imposed, requiring that each attribute must be measured by at least three

items before the CD-CAT can terminate. Specifically, after applying either the GDI (T-GDI, D-GDI, and R-GDI) or NPS (R-NPS, R-NPS$_{BM}$, and R-NPS$_{ML}$) item selection rule, if the cutoff of c = .80 is reached, it is verified that all attributes have been measured by at least three items. If not, the most optimal item (according to either the GDI or NPS rule) among those measuring the insufficiently explored attributes is administered. This process continues until both the c = .80 cutoff and the three-items-per-attribute criteria are met. Attribute profiles were estimated using *maximum a posteriori* (MAP) in all CD-CAT procedures.

Table 1

Specification of the CD-CAT Procedures

Procedure	Model	Calibration sample	ISR	Estimation method
T-GDI	DINA (TRUE)	Yes	GDI	ML
D-GDI	DINA	Yes	GDI	ML
R-GDI	R-DINA	Yes	GDI	ML
R-NPS	R-DINA	Yes	NPS	ML
R-NPS$_{ML}$	R-DINA	No	NPS	ML
R-NPS$_{BM}$	R-DINA	No	NPS	BM

Note. ISR = item selection rule; ML = maximum likelihood; BM = Bayes modal.

Design and Data Generation

Data were generated using the DINA model. Three independent variables were systematically manipulated: calibration sample size (N = 25, 50, 100), number of attributes (K = 3, 5), and item quality (IQ = low, medium, high, mixed). The chosen calibration sample size levels reflect those from applied small-scale assessments (e.g., Jang et al., 2015; Ren et al., 2021) as well as simulation studies focused on small sample sizes (e.g., Oka & Okada, 2021). Note that a calibration sample is required for the D-GDI, R-GDI, and R-NPS, but not for R-NPS$_{ML}$ or R-NPS$_{BM}$ (see Table 1). Regarding the number of attributes, Sessoms and Henson (2018) found that four attributes are most common in applied research.

Item quality was varied by manipulating the probability of correctly answering an item for latent class {**0**} (i.e., non-masters of all attributes) and latent class {**1**} (i.e., masters of all

attributes) as follows: $P(\mathbf{0})\sim U(0.2,0.4)$ and $P(\mathbf{1})\sim U(0.6,0.8)$ for low-quality items, $P(\mathbf{0})\sim U(0.1,0.3)$ and $P(\mathbf{1})\sim U(0.7,0.9)$ for medium-quality items, and $P(\mathbf{0})\sim U(0,0.2)$ and $P(\mathbf{1})\sim U(0.8,1)$ for high-quality items. This results in average item quality of $\overline{IQ} = P(\mathbf{1}) - P(\mathbf{0}) \approx 0.4$, 0.6, and 0.8 for low, medium, and high item quality, respectively. Furthermore, the mixed item quality condition combined $P(\mathbf{0})\sim U(0,0.2)$ and $P(\mathbf{1})\sim U(0.6,0.8)$ for half of the items, and $P(\mathbf{0})\sim U(0.2,0.4)$ and $P(\mathbf{1})\sim U(0.8,1)$ for the other half. This mixed condition directly violates the assumptions of the R-DINA model (i.e., same guessing and slip parameters for all items), making it particularly challenging for this model (Nájera, Abad, et al., 2023).

The item bank consisted of 300 items. The Q-matrices were randomly generated with the constraint of containing 150 one-attribute items, 120 two-attribute items, and 30 three-attribute items. This distribution ensures the completeness of the Q-matrices (Köhn & Chiu, 2017) while mimicking the complexity typically found in applied studies (Nájera, Abad, et al., 2021). Additionally, attribute profiles were generated using the multivariate normal threshold model (Chiu et al., 2009). Specifically, K continuous latent variables were drawn from a multivariate normal distribution with a mean of 0 and correlations of 0.5, reflecting the moderately large attribute correlations found in applied studies (Sessoms & Henson, 2018). These continuous latent variables were then dichotomized by assigning α_{ik} = 0 or 1 depending on whether the continuous score was lower or higher than 0, respectively.

Data generation and analyses were performed using R (R Core Team, 2023) with several packages: 'GDINA' version 2.8.7 (W. Ma & de la Torre, 2020), 'cdmTools' version 1.0.2 (Nájera, Sorrel, et al., 2023), 'NPCD' version 1.0-11 (Zheng & Chiu, 2022), and 'cdcatR' version 1.0.6 (Sorrel et al., 2022). The R code for the simulations and analyses is publicly available at https://osf.io/mn86s.

Performance Measures

All CD-CAT methods were evaluated in terms of attribute classification accuracy, item parameter recovery, and efficiency. The primary dependent variable of the study was the *relative measurement precision* (RMP; Huang, 2018), as it serves as an omnibus measure encompassing classification accuracy, parameter recovery, and efficiency. Specifically, the RMP is defined as

$$\text{RMP} = \frac{\sum_{i=1}^{N} \max_{l}\left[P\left(\boldsymbol{\alpha}_l \middle| \boldsymbol{y}_i, \widehat{\boldsymbol{\delta}}\right)\right]}{\sum_{i=1}^{N} \max_{l}\left[P\left(\boldsymbol{\alpha}_l \middle| \boldsymbol{y}_i, \boldsymbol{\delta}\right)\right]}, \tag{8}$$

where $P\left(\boldsymbol{\alpha}_l \middle| \boldsymbol{y}_i, \widehat{\boldsymbol{\delta}}\right)$ denotes the posterior probability of latent class l for examinee i based on the estimated item parameters, and $P(\boldsymbol{\alpha}_l | \boldsymbol{y}_i, \boldsymbol{\delta})$ is based on the generating item parameters. Table 2 summarizes the relationship between RMP, true classification accuracy, and the number of items administered in the CD-CAT application (i.e., test length). RMP reflects the overall accuracy of item parameter estimates. Item parameter estimates directly affect the calculation of posterior probabilities of attribute mastery, which are used to determine reliability (e.g., estimated classification accuracy) and, particularly in CD-CAT, as a stopping criterion for the test. An RMP close to 1 indicates that the posterior probabilities of attribute profiles are accurately recovered, implying that item parameters have been correctly estimated. In contrast, A CD-CAT method that overestimates reliability, as indicated by an RMP greater than 1, will meet the stopping criterion prematurely, potentially resulting in overly short and inaccurate CD-CAT applications. Conversely, methods that underestimate reliability, as indicated by an RMP less than 1, may fail to meet the stopping criterion, leading to longer and less efficient CD-CAT applications.

Although the RMP indicates whether the desired classification accuracy has been reached, classification accuracy was directly assessed using the *proportion of correctly classified vectors* (PCV):

$$\text{PCV} = \frac{\sum_{i=1}^{N} I(\widehat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i)}{N}, \tag{9}$$

where $I(\cdot)$ is the indicator function, and $\widehat{\boldsymbol{\alpha}}_i$ and $\boldsymbol{\alpha}_i$ represent the estimated and generated attribute profile for examinee *i*, respectively.

Table 2

Expected Relation Between Several Performance Measures

RMP	PCV	Test Length	Label
RMP < 1	PCV > *c*	Overly long	Inefficient
RMP ≈ 1	PCV ≈ *c*	Optimal	Optimal
RMP > 1	PCV < *c*	Overly short	Inaccurate

Note. RMP = relative measurement precision; PCV = proportion of correctly classified vectors.

Additionally, CD-CAT efficiency was evaluated based on the average *test length* (TL), which reflects the average number of items administered to examinees, and the *test overlap rate* (TOR; Chen et al., 2003), defined as:

$$\text{TOR} = \frac{J}{J^*} S_r^2 + \frac{J^*}{J}, \tag{10}$$

where J is the number of items administered, J^* is the item bank length (i.e., 300), and S_r^2 is the sample variance of item exposure rates. A high TOR implies that some items have been overexposed while many items have been underused, suggesting potential issues with test security.

Results

Relative Measurement Precision

Table 3 presents the RMP for the six CD-CAT procedures across the different levels of the number of attributes (*K*) and item quality (*IQ*). Note that the calibration sample size (*N*) is not included as an independent variable, given its negligible effect on all CD-CAT procedures except for D-GDI (note the small standard deviations in Table 3). Therefore, given that calibration sample size only significantly impacted D-GDI, the results for this procedure are separated into D-GDI$_{25}$, D-GDI$_{50}$, and D-GDI$_{100}$, corresponding to the different calibration sample sizes.

Table 3

Means (and Standard Deviations) of the CD-CAT Procedures

K	*IQ*	T-GDI	D-GDI$_{25}$	D-GDI$_{50}$	D-GDI$_{100}$	R-GDI	R-NPS	R-NPS$_{ML}$	R-NPS$_{BM}$
					Relative Measurement Precision (RMP)				
3	Low	1.000	1.694 (.106)	1.339 (.061)	1.166 (.026)	**0.999** (.012)	**0.999** (.012)	1.155 (.012)	1.143 (.011)
	Mixed	1.000	1.244 (.048)	1.132 (.031)	1.066 (.018)	**0.996** (.007)	**0.996** (.007)	**1.010** (.005)	**0.998** (.005)
	Medium	1.000	1.295 (.037)	1.175 (.031)	1.083 (.016)	**0.999** (.007)	**1.000** (.006)	**1.012** (.005)	**0.999** (.004)
	High	1.000	**1.026** (.007)	**1.020** (.010)	**1.016** (.007)	**0.998** (.003)	**0.997** (.004)	**0.979** (.002)	**0.970** (.002)
5	Low	1.000	2.298 (.222)	1.502 (.102)	1.212 (.039)	**0.999** (.015)	**0.999** (.015)	1.232 (.016)	1.208 (.018)
	Mixed	1.000	1.404 (.077)	1.203 (.044)	1.090 (.018)	**0.994** (.009)	**0.994** (.008)	**1.039** (.008)	**1.018** (.007)
	Medium	1.000	1.523 (.068)	1.277 (.058)	1.112 (.020)	**1.000** (.009)	**1.000** (.008)	**1.041** (.007)	**1.020** (.006)
	High	1.000	**1.046** (.012)	**1.044** (.013)	**1.026** (.010)	**0.997** (.004)	**0.997** (.004)	**0.984** (.003)	**0.972** (.003)
					Proportion of Correctly Classified Vector (PCV)				
3	Low	.855 (.016)	.468 (.055)	.634 (.043)	.730 (.026)	**.827** (.020)	**.829** (.019)	.772 (.019)	.775 (.019)
	Mixed	.943 (.013)	.698 (.052)	.798 (.048)	**.873** (.026)	**.881** (.015)	**.881** (.015)	**.910** (.014)	**.914** (.014)
	Medium	.920 (.013)	.657 (.038)	.734 (.039)	**.836** (.027)	**.881** (.015)	**.880** (.016)	**.907** (.014)	**.912** (.013)
	High	.990 (.007)	**.967** (.010)	**.961** (.019)	**.961** (.017)	**.947** (.011)	**.942** (.011)	**.980** (.006)	**.983** (.006)
5	Low	.827 (.017)	.309 (.049)	.531 (.051)	.676 (.033)	.796 (.020)	**.802** (.021)	.696 (.020)	.706 (.021)
	Mixed	.921 (.014)	.565 (.062)	.722 (.050)	**.831** (.027)	**.869** (.017)	**.870** (.016)	**.858** (.015)	**.872** (.015)
	Medium	.898 (.014)	.488 (.043)	.641 (.055)	.790 (.032)	**.867** (.015)	**.871** (.016)	**.860** (.016)	**.873** (.016)
	High	.979 (.009)	**.932** (.024)	**.909** (.027)	**.934** (.025)	**.940** (.011)	**.931** (.012)	**.960** (.009)	**.968** (.008)

Note. *K* = number of attributes; *IQ* = item quality. For the sake of simplicity, T-GDI, R-GDI, and R-NPS are not split by sample size due to the negligible effect of this variable on their performance, as indicated by the small standard deviations reported in this table. RMP values between 1.050 and 0.950 are shown in bold (excluding the T-GDI procedure). PCV values higher than 0.800 are shown in bold (excluding the T-GDI procedure).

Table 3 (Continued)

Means (and Standard Deviations) of the CD-CAT Procedures

K	*IQ*	T-GDI	D-GDI$_{25}$	D-GDI$_{50}$	D-GDI$_{100}$	R-GDI	R-NPS	R-NPS$_{ML}$	R-NPS$_{BM}$
					Test Length (TL)				
3	Low	11.8 (0.35)	**7.8** (0.43)	**9.4** (0.47)	10.5 (0.38)	19.8 (1.03)	20.2 (1.01)	19.2 (0.81)	19.0 (0.80)
	Mixed	7.5 (0.16)	**7.0** (0.16)	**7.3** (0.21)	**7.4** (0.18)	10.1 (0.17)	10.5 (0.17)	12.0 (0.30)	12.1 (0.29)
	Medium	7.8 (0.14)	**6.9** (0.19)	**7.3** (0.18)	**7.5** (0.18)	10.0 (0.18)	10.5 (0.17)	12.0 (0.30)	12.1 (0.29)
	High	6.7 (0.14)	**7.1** (0.09)	**6.9** (0.18)	**6.7** (0.14)	**7.6** (0.12)	**8.0** (0.13)	8.9 (0.13)	9.0 (0.13)
5	Low	23.6 (1.07)	**14.0** (0.96)	18.5 (1.30)	21.6 (1.23)	39.5 (1.82)	41.7 (2.04)	36.3 (1.86)	36.7 (1.89)
	Mixed	12.9 (0.24)	**11.8** (0.33)	**12.5** (0.33)	**12.8** (0.27)	18.2 (0.44)	19.3 (0.43)	20.0 (0.48)	20.5 (0.50)
	Medium	13.6 (0.25)	**11.6** (0.27)	**12.6** (0.32)	**13.2** (0.28)	18.2 (0.38)	19.4 (0.39)	20.1 (0.49)	20.5 (0.47)
	High	11.2 (0.22)	**12.0** (0.21)	**11.4** (0.24)	**11.2** (0.24)	**13.1** (0.16)	13.6 (0.14)	14.4 (0.20)	14.7 (0.20)
					Test Overlap Rate (TOR)				
3	Low	.488 (.031)	.527 (.042)	.502 (.034)	.495 (.029)	**.107** (.014)	**.103** (.013)	**.096** (.014)	**.096** (.015)
	Mixed	.490 (.035)	.488 (.045)	.489 (.041)	.482 (.035)	**.067** (.013)	**.065** (.012)	**.070** (.012)	**.071** (.012)
	Medium	.497 (.032)	.478 (.051)	.498 (.036)	.489 (.030)	**.067** (.013)	**.065** (.013)	**.071** (.013)	**.071** (.013)
	High	.467 (.035)	.258 (.053)	.417 (.047)	.441 (.037)	**.057** (.013)	**.056** (.013)	**.059** (.013)	**.059** (.013)
5	Low	.478 (.035)	.539 (.039)	.503 (.039)	.477 (.028)	**.176** (.018)	**.174** (.017)	**.151** (.019)	**.153** (.019)
	Mixed	.511 (.030)	.513 (.037)	.507 (.033)	.504 (.029)	**.103** (.013)	**.097** (.013)	**.097** (.013)	**.100** (.014)
	Medium	.504 (.028)	.512 (.038)	.516 (.034)	.501 (.026)	**.103** (.013)	**.097** (.014)	**.098** (.014)	**.100** (.014)
	High	.482 (.027)	.381 (.045)	.458 (.030)	.466 (.025)	**.085** (.012)	**.075** (.012)	**.077** (.012)	**.079** (.012)

Note. *K* = number of attributes; *IQ* = item quality. For the sake of simplicity, T-GDI, R-GDI, and R-NPS are not split by sample size due to the negligible effect of this variable on their performance, as indicated by the small standard deviations reported in this table. Lowest TL values (differences lower than 2 are not considered) are shown in bold (excluding the T-GDI procedure). Lowest TOR values (differences lower than 0.050 are not considered) are shown in bold (excluding the T-GDI procedure).

Overall, R-GDI and R-NPS ($0.994 \leq RMP \leq 1.000$) produced the best results in terms of RMP, accurately recovering posterior probabilities across all conditions. The two calibration-free procedures (R-NPS_{ML} and R-NPS_{BM}) also provided accurate posterior probabilities across all conditions ($0.972 \leq RMP \leq 1.041$), except for low-quality items, where they exhibited a tendency to overestimate reliability, resulting in higher RMP values ($1.143 \leq RMP \leq 1.232$). Finally, the D-GDI procedure only yielded accurate posterior probabilities with high-quality items ($1.016 \leq RMP \leq 1.046$). In the remaining conditions, it consistently overestimated reliability ($1.066 \leq RMP \leq 2.298$), with this tendency becoming more pronounced with smaller calibration sample sizes and larger number of attributes.

Classification Accuracy

As shown in Table 3, and consistent with the RMP results, the D-GDI procedure consistently provided the lowest classification accuracies. Given that the variable-length stopping criterion was $c = .80$, D-GDI_{25} and D-GDI_{50} only achieved a $PCV \geq .80$ with high-quality items, exhibiting poor classification accuracy ($.309 \leq PCV \leq .798$) under the remaining conditions. With a larger calibration sample size, D-GDI_{100} managed to reach the desired classification accuracy in more situations: namely, when item quality was not low and 3 attributes were measured ($.873 \leq PCV \leq .961$), and when item quality was mixed or high and 5 attributes were measured ($.831 \leq PCV \leq .934$). However, with low item quality, it produced unsatisfactory results with 3 attributes ($PCV = .730$) and 5 attributes ($PCV = .676$).

In contrast, R-GDI and R-NPS achieved a $PCV \geq .80$ under all conditions ($.802 \leq PCV \leq .947$), with the only exception being R-GDI under 5 attributes and low-quality items, where the PCV was still close to the desired cutoff ($PCV = .796$). Lastly, R-NPS_{ML} and R-NPS_{BM} provided satisfactory classification accuracy across all conditions ($.858 \leq PCV \leq .983$), except for low-

quality items (.696 ≤ PCV ≤ .775). These results align with those of the RMP, where the two calibration-free procedures only performed poorly when item quality was low. Lastly, and expectedly, the T-GDI procedure achieved a satisfactory classification accuracy (PCV ≥ .827) under all conditions.

Test Efficiency

Table 3 also presents the average test length and test overlap rate across all conditions. As expected, given the large RMP values obtained by the D-GDI procedure, its consistent tendency to overestimate reliability led to tests being stopped prematurely after administering only a few items. Consequently, D-GDI resulted in the shortest test lengths, particularly when the calibration sample size was smallest (N = 25). In contrast, the four procedures based on the R-DINA model administered more items. These differences were especially pronounced with low-quality items, where the test lengths of these procedures were up to 1.77 times that of the TRUE procedure. However, with medium or high-quality items, these differences were less pronounced, particularly for R-GDI and R-NPS (up to 1.43 times the test length of the TRUE procedure).

Table 4

Observed Relation Between Several Performance Measures

K	*IQ*	*r*(RMP, PCV)	*r*(RMP, TL)	*r*(PCV, TL)
3	Low	–.967	–.810	.852
	Mixed	–.902	–.739	.754
	Medium	–.944	–.774	.836
	High	–.545	–.931	.552
5	Low	–.971	–.819	.873
	Mixed	–.964	–.778	.756
	Medium	–.977	–.816	.845
	High	–.719	–.907	.652

Note. *K* = number of attributes; *IQ* = item quality; RMP = relative measurement precision; PCV = proportion of correctly classified vectors; TL = test length.

Table 4 displays the correlations between RMP, PCV, and test length, considering all CD-CAT procedures except T-GDI (which does not exhibit variability in RMP). Specifically, RMP was

inversely correlated with PCV ($-.977 \leq r \leq -.545$) and test length ($-.931 \leq r \leq -.739$), while PCV and test length were positively correlated ($.552 \leq r \leq .873$). These results align with the expected relationships among these three variables summarized in Table 2.

Lastly, the parametric procedures (T-GDI and D-GDI) exhibited a large test overlap rate ($.258 \leq \text{TOR} \leq .539$), indicating that some items were overexposed while others were underused. On the other hand, the procedures based on the R-DINA model showed a much lower TOR ($.056 \leq \text{TOR} \leq .176$), suggesting a more balanced use of the item bank.

Discussion

CDM is a family of restricted latent class models that can enhance educational formative assessments by identifying students' strengths and weaknesses (de la Torre & Minchen, 2014; Paulsen & Valdivia, 2021). Despite the significant methodological advancements within the CDM framework, the number of applied studies remains limited (Sessoms & Henson, 2018). One potential reason for this could be a misalignment between the focus of theoretical research and practical needs. Although recent efforts have concentrated on developing and testing CDM procedures for small-scale assessments or suboptimal conditions (e.g., Chiu & Douglas, 2013; Chiu et al., 2018; W. Ma & Jiang, 2021; Nájera, Abad, et al., 2023; Oka & Okada, 2021; Paulsen & Valdivia, 2021), these developments have primarily focused on traditional "paper-and-pencil" assessments. However, the educational goals of CDM require efficient testing to be effectively integrated as an evaluation tool throughout a course to guide formative assessment. To address this, the present paper proposes and evaluates a procedure that combines the R-DINA model with the CD-CAT framework, enabling diagnostic, contrastable, and efficient assessments in small-scale contexts, such as those typical in educational settings. This new approach is flexible enough to accommodate two types of scenarios: a traditional one where a calibration sample is used to

estimate item parameters, which are then administered adaptively to new examinees, and a more innovative approach aligned with the NPS method (Chang et al., 2018), where the CD-CAT can be applied without a calibration sample, as the parameter of the R-DINA model is calibrated on-the-fly. In this paper, two different variants have been proposed for each scenario. In the former, after calibrating the item bank, either the GDI or the NPS item selection rule can be used for the adaptive administration. In the latter, ML or BM estimation can be employed for calibrating the φ parameter on-the-fly.

In the simulation study, the performance of the four different implementations of the proposed method was compared to that of the traditional CD-CAT, using the well-known DINA model for calibration and the GDI item selection rule. The results indicate that when the calibration sample is smaller than 100 examinees, the DINA model struggles to obtain accurate item parameters due to overfitting, leading to extreme posterior probabilities and, in turn, to reliability overestimation. These findings are particularly problematic in an adaptative testing situation, as they cause the CD-CAT to stop before reaching the desired level of precision, resulting in inaccurate assessments. In contrast, the two new procedures based on the R-DINA with a calibration sample, namely the R-GDI and R-NPS, performed satisfactorily across all conditions in terms of posterior probability recovery (relative measurement precision), classification accuracy, and test overlap rate. These results are consistent with previous work comparing models with varying degrees of complexity, which has found that simpler models may be preferred over more complex ones, even when the latter are closer to the data generation process, in conditions where the available information is insufficient to estimate several parameters (Nájera, Abad et al., 2023; Sorrel, Nájera et al., 2021).While GDI and NPS differ in how they define item discrimination and treat prior information, their comparable performance under the R-DINA model likely stems

from relying on the same underlying classification mechanism. When both rules select similar items, the resulting classifications tend to align. This practical convergence, despite theoretical differences, highlights a potential pathway for connecting parametric and nonparametric CDMs, and invites further exploration into the conditions under which such alignment occurs (C. Ma et al., 2023; Nájera, Abad et al., 2023). To achieve these satisfactory results, the new R-GDI and R-NPS methods tended to administer a larger number of items. However, these longer tests are offset by much greater confidence in the estimates and classification accuracy they provided compared to the traditional CD-CAT when the calibration sample size is small.

Regarding the two proposed calibration-free procedures, both methods provided similar results, which were generally satisfactory except for low-quality items, where they tended to overestimate reliability and, like the DINA model, resulted in low classification accuracy. It should be highlighted that these generally satisfactory results were obtained using an on-the-fly estimation approach with responses from single individuals (i.e., $N = 1$). One reason why R-NPS$_{ML}$ performed very similarly to the R-NPS$_{BM}$ is the content restriction imposed in the CD-CAT; without this restriction, R-NPS$_{ML}$ would have terminated the test much earlier than R-NPS$_{BM}$ due to the boundary problem (see Figure 1). The need to administer additional items to comply with the content restriction helped achieve a more accurate estimation of the φ parameter using maximum likelihood.

Limitations, Future Research, and Practical Recommendations

The study is not without limitations, which are listed here to be considered when interpreting these results and in future research. First, it is worth noting that this study employed GDI as a representative of parametric item selection rules, but many other alternatives are available (e.g., Kaplan et al., 2015; C. Wang, 2013; Xu et al., 2016). However, it should also be noted that

many of these selection rules are expected to yield similar results when considering the relationships between them (W. Wang et al., 2020). There is a similar situation with nonparametric item selection rules: an auxiliary analysis revealed that CD-CAT using the R-DINA model performed similarly with either the NPS or NDBS rule (see Online Appendix). For studies focusing more on the comparison between parametric and nonparametric rules, Chiu and Chang (2021) can be consulted. Second, the data were generated using a single model, the DINA model, due to its popularity. The results should be generalizable to other models, such as DINO (Templin & Henson, 2006), considering it is equally complex (Köhn & Chiu, 2016). It would be beneficial to extend the proposal to other types of data (e.g., polytomous data, Gao et al., 2020) and models (e.g., multiple strategies, D. Wang et al., 2024; coded distractors, Y. Wang et al., 2024). Finally, a simple and widely used stopping rule was chosen, aligned with those commonly employed in existing CD-CAT applications (Li et al., 2023) and incorporated into freely available software, such as the R package 'cdcatR' (Sorrel et al., 2022). This rule, based on posterior probabilities, can be directly interpreted in terms of expected reliability; in this regard, the use of the R-DINA as a bridge between parametric and nonparametric CDM enables retrieving this reliability information from nonparametric procedures (e.g., NPC, NPS, NDBS). Moreover, as long as the posterior probability and item parameters are accurately estimated, any other stopping rules based on this information are expected to perform adequately. Future studies may focus on the comparison of different stopping criteria to discuss potential benefits of other alternatives (e.g., Guo & Zheng, 2019; Li and Zheng, 2024).

The proposed procedures can be employed to enhance formative assessments in real-world settings, as they offer solutions to the challenges of small sample sizes and time constraints that are common in many educational contexts. Specifically, the use of a model with low

parameterization, such as the R-DINA developed here for its application in variable test length adaptive testing, should allow for efficient, accurate classifications. This enables teachers to focus less on testing time and implement tailored instruction based on the strengths and weaknesses detected in students. There are a few important considerations. The first concerns the restrictive nature of the R-DINA model, which underpins the four proposed CD-CAT procedures. The R-DINA model assumes a conjunctive relationship between attributes, which may or may not be appropriate. Although the R-DINA model has been shown to outperform the DINA model in small-scale settings, even when the DINA model was the generating model, it is essential to carefully check model fit to ensure that the conjunctive item response function is suitable for the data (Nájera, Abad, et al., 2023). The same caution applies to the disjunctive rule used in the R-DINO and DINO models. It is possible that no reduced model can fully capture the complexity of a given dataset, in which case a more general model should be preferred. However, general models require larger sample sizes to achieve accurate estimates, so there is currently no optimal solution for this inconvenient scenario. Thus, it is particularly important to carefully design the assessment when working with small samples, reflecting on the item response process (e.g., conjunctive, disjunctive, general). In this vein, evaluations of mathematical abilities and language mastery have commonly found that, in these domains, the conjunctive rule (i.e., DINA model) often reflects the relationship between the attributes (e.g., George & Robitzsch, 2021; Groß et al., 2016).

Given these considerations, we recommend using the proposed methods primarily in low-stakes contexts, which naturally align with the purpose of formative assessment (Paulsen & Valdivia, 2021). Additionally, when designing an educational CD-CAT project, the R-NPS$_{ML}$ and R-NPS$_{BM}$ methods could be utilized with the first cohort when there is no prior informative available for model calibration. For subsequent cohorts, a small calibration sample from previous

cohorts would be available, making the R-GDI and R-NPS methods more appropriate. Once the calibration sample is sufficiently large, more complex models, such as the DINA or even the G-DINA model, could be compared in terms of fit and reliability to achieve an optimal solution.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

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, S.-Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, *40*(2), 129–145. https://doi.org/10.1111/j.1745-3984.2003.tb01100.x

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*(4), 619–632. http://dx.doi.org/10.1007/S11336-009-9123-2

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement*, *70*(6), 902–913. https://doi.org/10.1177/0013164410366693

Cheng, Y., Chang, H.-H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement*, *31*(6), 467–482. https://doi.org/10.1177/0146621606292933

Chiu, C. Y., & Chang, Y. P. (2021). Advances in CD-CAT: The general nonparametric item selection method. *Psychometrika*, *86*, 1039–1057. https://doi.org/10.1007/s11336-021-09792-z

Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, *30*, 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*, 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*, 355–375. https://doi.org/10.1007/s11336-017-9595-4

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

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273. https://doi.org/10.1007/s11336-015-9467-8

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*, 89–97. http://dx.doi.org/10.1016/j.pse.2014.11.001

de la Torre, J., & Sorrel, M. A. (2023). Cognitive diagnosis models. In F. G. Ashby, H. Colonius, & E. N. Dzhafarov (Eds.), *New handbook of mathematical psychology. Volume 3, Perceptual and cognitive processes* (pp. 385–420). Cambridge University Press.

Gao, X., Wang, D., Cai, Y., & Tu, D. (2020). Cognitive diagnostic computerized adaptive testing for polytomously scored items. *Journal of Classification*, *37*, 709–729. https://doi.org/10.1007/s00357-019-09357-x

Garre, F. G., & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika*, *33*, 43–59. https://doi.org/10.2333/bhmk.33.43

George, A. C., & Robitzsch, A. (2021). Validating theoretical assumptions about reading with cognitive diagnosis models. *International Journal of Testing*, *21*(2), 105–129. https://doi.org/10.1080/15305058.2021.1931238

Groß, J., Robitzsch, A., & George, A. C. (2016). Cognitive diagnosis models for baseline testing of educational standards in math. *Journal of Applied Statistics*, *43*(1), 229–243. https://doi.org/10.1080/02664763.2014.1000841

Gu, Y., & Xu, G. (2021). Sufficient and necessary conditions for the identifiability of the Q-matrix. *Statistica Sinnica*, *31*, 449–472. https://doi.org/10.5705/ss.202018.0410

Guo, L., & Zheng, C. (2019). Termination rules for variable-length CD-CAT from the information theory perspective. *Frontiers in Psychology*, *10*, 1122. https://doi.org/10.3389/fpsyg.2019.01122

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

Hsu, C. L., Wang, W. C., & Chen, S. Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement*, *37*(7), 563–582. https://doi.org/10.1177/0146621613488642

Huang, H.-Y. (2018). Effects of item calibration errors on computerized adaptive testing under cognitive diagnosis models. *Journal of Classification*, *35*, 437–465. https://doi.org/10.1007/s00357-018-9265-y

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

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

Kang, H.-A., Zhang, S., & Chang, H.-H. (2017). Dual-objective item selection criteria in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, *54*(2), 165–183. https://doi.org/10.1111/jedm.12139

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, *39*(3), 167–188. https://doi.org/10.1177/0146621614554650

Köhn, H. F., & Chiu, C.-Y. (2016). A proof of the duality of the DINA model and the DINO model. *Journal of Classification*, *33*, 171–184. https://doi.org/10.1007/s00357-016-9202-x

Köhn, H. F., & Chiu, C.-Y. (2017). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika*, 82, 112–132.. https://doi.org/10.1007/s11336-016-9536-7

Kreitchmann, R. S., de la Torre, J., Sorrel, M. A., Nájera, P., & Abad, F. J. (2023). Improving reliability estimation in cognitive diagnosis modeling. *Behavior Research Methods*, *55*, 3446–3460. https://doi.org/10.3758/s13428-022-01967-5

Li, Y., Huang, C., & Liu, J. (2023). Diagnosing Primary Students' Reading Progression: Is Cognitive Diagnostic Computerized Adaptive Testing the Way Forward? *Journal of*

Educational and Behavioral Statistics, *48*(6), 842–865. https://doi.org/10.3102/10769986231160668

Li, J., & Zheng, H. (2024). Non-parametric CD-CAT item selection strategy and termination rules based on binary search algorithm. *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语期刊*, *5*(1), 1-20. https://doi.org/10.59863/DKUI7768

Ma, C., de la Torre, J. & Xu, G. (2023). Bridging parametric and nonparametric methods in cognitive diagnosis. *Psychometrika*, *88*, 51–75. https://doi.org/10.1007/s11336-022-09878-2

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, *69*, 253–275. https://doi.org/10.1111/bmsp.12070

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., & Guo, W. (2019). Cognitive diagnosis models for multiple strategies. *British Journal of Mathematical and Statistical Psychology*, *72*(2), 370–392. https://doi.org/10.1111/bmsp.12155

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., 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*, *48*(6), 719–749. https://doi.org/10.3102/10769986231158829

Nájera, P., Sorrel, M. A., & Abad, F. J. (2023). *cdmTools: Useful tools for cognitive diagnosis modeling*. R package Version 1.0.3. https://cran.r-project.org/package=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

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., & Valdivia, D. S. (2021). Examining cognitive diagnostic modeling in classroom assessment conditions. *The Journal of Experimental Education*, *90*(4), 916–933. https://doi.org/10.1080/00220973.2021.1891008

R Core Team. (2023) *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

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., Barrada, J. R., de la Torre, J., & Abad, F. J. (2020). Adapting cognitive diagnosis computerized adaptive testing item selection rules to traditional item response theory. *PLoS ONE*, *15*(1):e0227196. https://doi.org/10.1371/journal. pone.0227196

Sorrel, M. A., Nájera, P., & Abad, F. J. (2022). *cdcatR: Cognitive Diagnostic Computerized Adaptive Testing*. R package version 1.0.6. https://CRAN.R-project.org/package=cdcatR

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, *19*(3), 506–532. https://doi.org/10.1177/1094428116630065

Sun, X., Andersson, B., & Xin, T. (2021). A new method to balance measurement accuracy and attribute coverage in cognitive diagnostic computerized adaptive testing. *Applied*

Psychological Measurement, *45*(7-8), 463–476. https://doi.org/10.1177/01466216211040489

Sun, X., Liu, Y., Xin, T., & Song, N. (2020). The impact of item calibration error on variable-length cognitive diagnostic computerized adaptive testing. *Frontiers in Psychology*, *11*, 575141. https://doi.org/10.3389/fpsyg.2020.575141

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *51*(3), 337-350. https://doi.org/10.1111/1467-9876.00272

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *65*(1), 143-157. https://doi.org/10.1111/1467-9868.00377

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

von Davier, M., & Lee, Y.-S. (Eds.). (2019). *Handbook of Diagnostic Classification Models*. Springer. https://doi.org/10.1007/978-3-030-05584-4

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, 73(6), 1017–1035. https://doi.org/10.1177/0013164413498256

Wang, D., Ma, W., Cai, Y., & Tu, D. (2024). A general nonparametric classification method for multiple strategies in cognitive diagnostic assessment. *Behavior Research Methods*, *56*(2), 723–735. https://doi.org/10.3758/s13428-023-02075-8

Wang, W., Song, L., Wang, T., Gao, P., & Xiong, J. (2020). A note on the relationship of the Shannon entropy procedure and the Jensen–Shannon divergence in cognitive diagnostic computerized adaptive testing. *Sage Open*, *10*(1). https://doi.org/10.1177/2158244019899046

Wang, Y., Chiu, C. Y., & Köhn, H. F. (2024). Nonparametric CD-CAT for multiple-choice items: Item selection method and Q-optimality. *British Journal of Mathematical and Statistical Psychology*, *78*, 61–83. https://doi.org/10.1111/bmsp.12350

Xu, G., Wang, C., & Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 291–315. https://doi.org/10.1111/bmsp.12072

Zheng, C., & Wang, C. (2017). Application of binary searching for item exposure control in cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *41*(7), 561-576. https://doi.org/10.1177/0146621617707509

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