Psychometrika (2025), **0**, 1–17
doi:10.1017/psy.2025.9

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Consistency Theory of General Nonparametric Classification Methods in Cognitive Diagnosis

Chengyu Cui[1,†], Yanlong Liu[2,†] and Gongjun Xu[1]

[1]Department of Statistics, University of Michigan, Ann Arbor, MI, USA; [2]Booth School of Business, University of Chicago, Chicago, IL, USA

Corresponding author: Gongjun Xu; Email: gongjun@umich.edu

(Received 30 October 2024; revised 13 January 2025; accepted 10 March 2025)

[†]Chengyu Cui and Yanlong Liu are co-first authors.

Abstract

Cognitive diagnosis models (CDMs) have been popularly used in fields such as education, psychology, and social sciences. While parametric likelihood estimation is a prevailing method for fitting CDMs, nonparametric methodologies are attracting increasing attention due to their ease of implementation and robustness, particularly when sample sizes are relatively small. However, existing consistency results of the nonparametric estimation methods often rely on certain restrictive conditions, which may not be easily satisfied in practice. In this article, the consistency theory for the general nonparametric classification method is reestablished under weaker and more practical conditions.

Keywords: cognitive diagnosis; consistency theory; general nonparametric classification method; Q-matrix

1. Introduction

Cognitive diagnosis models (CDMs), also known as diagnostic classification models (DCMs), are a popular family of discrete latent variable models employed in diagnostic assessments to provide detailed information about subjects' latent attributes based on their responses to designed diagnostic items. For instance, in educational testing, these latent attributes might indicate if a subject has mastered certain skills or not (de la Torre, 2011; Henson et al., 2009; Junker & Sijtsma, 2001); in psychiatric diagnosis, the latent attributes might signal the presence or absence of certain mental disorders (de la Torre et al., 2018; Templin & Henson, 2006).

Parametric models for cognitive diagnosis have been developed and widely applied in practice. Popular examples include the deterministic input, noisy "and" gate (DINA) model (Junker & Sijtsma, 2001), the deterministic input, noisy "or" gate (DINO) model (Templin & Henson, 2006), the general diagnostic model (GDM; von Davier, 2008), the reduced reparameterized unified model (reduced RUM; Hartz, 2002), the log-linear CDM (LCDM; Henson et al., 2009), and the generalized DINA model (GDINA; de la Torre, 2011). In conventional settings with a fixed number of items (J) and a large number of subjects (N), the latent attributes are often viewed as random variables. The corresponding CDMs can thus be viewed as a family of finite mixture models, where each subject's latent attribute profile $\boldsymbol{\alpha}_i$ behaves as a discrete random variable following a categorical distribution. From this perspective, the estimation often takes place through the maximization of the marginal likelihood, relying on methods such as

the expectation-maximization algorithm (de la Torre, 2011; DiBello et al., 2007; von Davier, 2008). However, the maximum likelihood-based approach often necessitates sufficiently large assessments to guarantee the reliability of the item parameter estimation, and it may either produce inaccurate estimates with small sample sizes or suffer from high computational costs (Chiu & Köhn, 2019a; Chiu et al., 2018). Moreover, the parametric CDMs involve certain parametric assumptions about the item response functions, which may raise concerns about the validity of the assumed model and the underlying process (Chiu & Douglas, 2013).

As an alternative, researchers have explored nonparametric cognitive diagnosis methods (Chiu & Köhn, 2019b; Chiu et al., 2009). Instead of modeling the item response functions parametrically, the nonparametric methods aim to directly categorize subjects into latent groups by minimizing certain distance measure between a subject's observed item responses and some expected "centers" of the latent groups. Two popular examples of nonparametric cognitive diagnosis methods include the nonparametric classification (NPC) method (Chiu & Douglas, 2013) and its generalization, the general NPC (GNPC) method (Chiu et al., 2018). The GNPC method, in particular, has received increasing attention in recent years due to its effectiveness in handling complex CDMs and its good performance for sample sizes (Chandía et al., 2023; Chiu & Chang, 2021; Ma, de la Torre, et al., 2023; Wang et al., 2023). The algorithms of the NPC and GNPC methods are straightforward to implement and require minimal computational resources, making them highly appealing for practical applications.

Theoretical properties of the nonparametric methods have also been explored in the literature. Under some regularity conditions, the NPC estimators of the subjects' latent attribute profiles have been shown to be statistically consistent for certain CDMs, including DINA and reduced RUM (Wang & Douglas, 2015), and a similar consistency theory for the GNPC estimator has also been established (Chiu & Köhn, 2019a). However, the current theoretical guarantees for these nonparametric methods depend on relatively stringent assumptions. In the case of the NPC method, the assumptions associated with the ideal binary responses might oversimplify the underlying diagnostic process and thus be challenging to fulfill when dealing with complex underlying CDMs, such as the GDINA model and other general CDMs (Chiu et al., 2018). Although the GNPC method addresses the oversimplification issue of the NPC method, its consistency depends on a key assumption that consistent initial estimators of the latent attribute profiles are available. For instance, Theorem 1 in Chiu and Köhn (2019a) provides theoretical guarantees for the GNPC estimators, assuming an initialization that consistently estimates the ground truth latent memberships. Similarly, Theorems 1–3 in Ma, de la Torre, et al. (2023) require consistent estimation of latent memberships from a calibration dataset to establish their consistency results. The assumption that consistent initial estimators of latent attribute profiles can be obtained or that a calibration dataset is available may be overly restrictive in practice, and the consistency of the GNPC method in more realistic settings remains an open problem.

In this article, we establish the consistency for the GNPC method using different theoretical techniques, without relying on the previous assumption on initial consistent estimators or calibration datasets. Our analysis covers both the original GNPC method in Chiu and Köhn (2019a) and a modified version of the GNPC method in Ma, de la Torre, et al. (2023).

We establish finite-sample error bounds for latent attributes of general nonparametric methods as well as uniform consistency of the item parameters. We would like to clarify that the main contribution of this work lies in the theoretical analysis of the GNPC and modified GNPC methods. For the implementation of these methods, we recommend utilizing the algorithms proposed in the literature (Chiu & Köhn, 2019a; Chiu et al., 2018; Ma, de la Torre, et al., 2023), which have demonstrated the effectiveness of GNPC methods via extensive simulation studies and real data examples.

The rest of the paper is organized as follows: Section 2 provides a brief review of cognitive diagnostic models and discusses the limitations in the existing consistency results. Section 3 establishes consistency results of the GNPC methods. In Section 4, we provide a simulation study to illustrate our theoretical results. Section 5 gives some further discussions, and the Supplementary Material provides the proofs for the main results.

2. Model setup and nonparametric methods

This work focuses on CDMs for multivariate binary data, which are commonly encountered in educational assessments (correct/wrong answers) and social science survey responses (yes/no responses) (von Davier & Lee, 2019). For N subjects and J items, the observed data is an $N \times J$ binary matrix $\mathbf{R} = (R_{i,j})$, where $R_{i,j} = 1$ or 0 denotes whether the ith subject gives a positive response to the jth item. Consider K binary latent attributes. Let the row vector $\boldsymbol{\alpha}_i = (a_{i,1}, \ldots, a_{i,K})$ represent the latent attribute profile for the ith subject, where $a_{i,k} = 1$ or 0 indicates the presence or absence, respectively, of the kth attribute for the ith individual. We further use an $N \times K$ binary matrix, $\mathbf{A} = (a_{i,k}) \in \{0,1\}^{N \times K}$, to represent the latent attribute profiles for all N subjects.

To capture the dependence relationship between items and the latent attributes of subjects, a design matrix called the Q-matrix (Tatsuoka, 1985) is employed. The Q-matrix encodes how the J items depend on the K latent attributes. Specifically, $\mathbf{Q} = (q_{j,k}) \in \{0,1\}^{J \times K}$, where $q_{j,k} = 1$ or 0 indicates whether the jth test item depends on the kth latent attribute, and we denote the jth item's Q-matrix vector as $\mathbf{q}_j = (q_{j,1}, \ldots, q_{j,K})$.

For an integer m, we denote $[m] = \{1, \ldots, m\}$ and for a set $\mathcal{A}$, we denote its cardinality by $|\mathcal{A}|$. We denote $\theta_{j,\boldsymbol{\alpha}} = \mathbb{P}(R_{i,j} = 1 | \boldsymbol{\alpha}_i = \boldsymbol{\alpha})$ for any $i \in [N], j \in [J]$ and $\boldsymbol{\alpha} \in \{0,1\}^K$, and let $\boldsymbol{\Theta} = \{\theta_{j,\boldsymbol{\alpha}}; j \in [J], \boldsymbol{\alpha} \in \{0,1\}^K\}$. We assume each response $R_{i,j}$ follows a Bernoulli distribution with parameter $\theta_{j,\boldsymbol{\alpha}_i}$ and the responses are independent with each other conditional on the latent attribute profiles $\mathbf{A}$ and the structure loading matrix $\mathbf{Q}$. In summary, the data generative process aligns with the following latent class model:

$$\mathbb{P}(\mathbf{R} \mid \mathbf{A}, \boldsymbol{\Theta}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \mathbb{P}(R_{i,j} \mid \boldsymbol{\alpha}_i, \theta_{j,\boldsymbol{\alpha}_i}) = \prod_{i=1}^{N} \prod_{j=1}^{J} (\theta_{j,\boldsymbol{\alpha}_i})^{R_{i,j}} (1 - \theta_{j,\boldsymbol{\alpha}_i})^{1 - R_{i,j}}.$$

To further illustrate the adaptability of the general nonparametric method to the model structures embedded in CDMs imposed by the structural matrix $\mathbf{Q}$, we follow the general assumption for the restricted latent class models (Chiu & Köhn, 2015; Ma, de la Torre, et al., 2023; Xu, 2017) that for different attribute profiles $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}$,

$$\left(\boldsymbol{\alpha} \circ \mathbf{q}_j = \tilde{\boldsymbol{\alpha}} \circ \mathbf{q}_j\right) \Longrightarrow (\theta_{j,\boldsymbol{\alpha}} = \theta_{j,\tilde{\boldsymbol{\alpha}}}), \tag{1}$$

where $\boldsymbol{\alpha} \circ \mathbf{q}_j = (a_1 q_{j,1}, \ldots, a_K q_{j,K})$ denotes the element-wise product of binary vectors $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ and $\mathbf{q}_j$. This implies that the item response parameter $\theta_{j,\boldsymbol{\alpha}}$ only depends on whether the latent attribute profile $\boldsymbol{\alpha}$ contains the required attributes $\mathcal{K}_j := \{k \in [K]; q_{j,k} = 1\}$ for item j. In cognitive diagnostic assessments, the matrix $\mathbf{Q}$ is typically predetermined by domain experts (George & Robitzsch, 2015; Junker & Sijtsma, 2001; von Davier, 2008). In this work, we assume the Q-matrix $\mathbf{Q}$ is specified, and $(\mathbf{A}, \boldsymbol{\Theta})$ are to be estimated from the responses $\mathbf{R}$.

2.1. Parametric CDMs: DINA and DINO models

For parametric CDMs, the structural matrix $\mathbf{Q}$ imposes various constraints on the item parameters based on different cognitive assumptions. For instance, in the DINA (Junker & Sijtsma, 2001) model, a conjunctive relationship among the attributes is assumed. According to this assumption, for a subject to provide a positive (correct) response to an item, mastery of all the required attributes of the item is necessary. In the DINA model, the ideal response for each item $j \in [J]$ and each latent attribute profile $\boldsymbol{\alpha} = (a_1, \ldots, a_K)$ is defined as

$$\eta_{j,\boldsymbol{\alpha}}^{\mathrm{DINA}} = \prod_{k=1}^{K} a_k^{q_{j,k}}.$$

The DINO (Templin & Henson, 2006) model assumes a disjunctive relationship among attributes, where mastery of at least one of the required attributes for an item is necessary for a subject to be considered capable of providing a positive response. In the DINO model, the ideal response is defined as

$$\eta_{j,\boldsymbol{\alpha}}^{\text{DINO}} = 1 - \prod_{k=1}^{K} (1 - a_k)^{q_{j,k}}.$$

The DINA and DINO models further encompass uncertainty by incorporating the slipping and guessing parameters, denoted as s_j and g_j for $j \in [J]$. For each item j, the slipping parameter represents the probability of a capable subject giving a negative response, whereas the guessing parameter signifies the probability of an incapable subject giving a positive response. Specifically, $s_j = \mathbb{P}(R_{i,j} = 0 | \eta_{j,\boldsymbol{\alpha}_i} = 1)$ and $g_j = \mathbb{P}(R_{i,j} = 1 | \eta_{j,\boldsymbol{\alpha}_i} = 0)$ for the ith subject. Therefore, in these two restricted latent class models, the parameter $\theta_{j,\boldsymbol{\alpha}}$ can be expressed as

$$\theta_{j,\boldsymbol{\alpha}} = (1 - s_j)^{\eta_{j,\boldsymbol{\alpha}}} g_j^{1-\eta_{j,\boldsymbol{\alpha}}}.$$

2.2. Nonparametric CDMs: NPC and GNPC

For nonparametric CDMs, the ideal responses described under the DINA and DINO models serve as foundational elements for the NPC analysis. Given a set of 0–1 binary ideal responses, denoted as $\{\eta_{j,\boldsymbol{\alpha}}\}$, the NPC method, as introduced by Chiu and Douglas (2013), estimates the subjects' latent attribute profiles as follows. This method utilizes a distance-based algorithm, leveraging observed item responses to categorize subjects into latent groups. The NPC estimator, $\widehat{\boldsymbol{\alpha}}_i$, for the ith individual's attribute profile, $\boldsymbol{\alpha}_i$, is expressed as

$$\widehat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha} \in \{0,1\}^K}{\arg\min} \sum_{j=1}^{J} (R_{i,j} - \eta_{j,\boldsymbol{\alpha}})^2.$$

In the NPC method, the ideal responses $\eta_{j,\boldsymbol{\alpha}}$ can be based on either the DINA model or the DINO model. However, due to the dependence on these specific model assumptions, which define two extreme relations between $\boldsymbol{q}_j$ and latent attribute profile $\boldsymbol{\alpha}$, the NPC method may fail to handle complex CDMs, such as the GDINA model, and such limitation may lead to misclassifications of the subjects (Chiu & Köhn, 2019a).

To address this issue, the GNPC method (Chiu et al., 2018) offers a solution by considering a more general ideal response that represents a weighted average of the ideal responses from the DINA and DINO models, as in

$$\eta_{j,\boldsymbol{\alpha}}^{(w)} = w_{j,\boldsymbol{\alpha}} \eta_{j,\boldsymbol{\alpha}}^{\text{DINA}} + (1 - w_{j,\boldsymbol{\alpha}}) \eta_{j,\boldsymbol{\alpha}}^{\text{DINO}}. \tag{2}$$

The weights are determined by the data; therefore, the proportional influence of $\eta_{j,\boldsymbol{\alpha}}$ and $w_{j,\boldsymbol{\alpha}}$ on the weighted ideal item response is adapted to the complexity of the underlying CDM data generating process. The GNPC method can be utilized with any CDM that can be represented as a general CDM, without requiring prior knowledge of the underlying model. To obtain estimates of the weights, Chiu et al. (2018) proposed minimizing the L^2 distance between the responses to item j and the weighted ideal responses $\eta_{j,\boldsymbol{\alpha}}^{(w)}$:

$$d_{j,\boldsymbol{\alpha}} = \sum_{i:\boldsymbol{\alpha}_i = \boldsymbol{\alpha}} (R_{i,j} - \eta_{j,\boldsymbol{\alpha}}^{(w)})^2.$$

When $\eta_{j,\boldsymbol{\alpha}}^{\text{DINO}} = \eta_{j,\boldsymbol{\alpha}}^{\text{DINA}}$, this results in $\eta_{j,\boldsymbol{\alpha}}^{(w)} = \eta_{j,\boldsymbol{\alpha}}^{\text{DINA}} = \eta_{j,\boldsymbol{\alpha}}^{\text{DINO}}$, which happens either when $\boldsymbol{\alpha}$ includes all the required latent attributes in $\mathcal{K}_j$, leading to $\eta_{j,\boldsymbol{\alpha}}^{(w)} = 1$, or when $\boldsymbol{\alpha}$ does not contain any required attributes, resulting in $\eta_{j,\boldsymbol{\alpha}}^{(w)} = 0$. Equivalently, these two extreme situations can be summarized as the following constraints:

$$\left(\boldsymbol{\alpha} \cdot \mathbf{q}_j = 0 \Longrightarrow \eta_{j,\boldsymbol{\alpha}}^{(w)} = 0\right) \text{ and } \left(\boldsymbol{\alpha} \cdot \mathbf{q}_j = K_j \Longrightarrow \eta_{j,\boldsymbol{\alpha}}^{(w)} = 1\right), \tag{3}$$

where $\boldsymbol{\alpha}\cdot\mathbf{q}_j = \sum_{k=1}^{K}\alpha_k q_{j,k}$ denotes the inner product of the two vectors and K_j is defined as $\sum_{k=1}^{K} q_{j,k}$, representing the number of latent attributes that the jth item depends on. Thus, in these two extreme situations, the parameters $\eta_{j,\boldsymbol{\alpha}}^{(w)}$ are known and do not need estimation. In scenarios where $\boldsymbol{\alpha}$ includes only some of the required attributes, $\eta_{j,\boldsymbol{\alpha}}^{(w)}$ need to be estimated, and in such cases, minimizing $d_{j,\boldsymbol{\alpha}}$ would lead to

$$\widehat{w}_{j,\boldsymbol{\alpha}} = 1 - \overline{R}_{j,\boldsymbol{\alpha}}, \quad \widehat{\eta}_{j,\boldsymbol{\alpha}}^{(w)} = \overline{R}_{j,\boldsymbol{\alpha}}, \tag{4}$$

where $\overline{R}_{j,\boldsymbol{\alpha}} = \sum_{i:\boldsymbol{\alpha}_i=\boldsymbol{\alpha}} R_{i,j}/|\{i\in[N];\boldsymbol{\alpha}_i=\boldsymbol{\alpha}\}|$, which represents the sample mean of the responses to the jth item for subjects with given latent attribute profile $\boldsymbol{\alpha}$. Since the true latent attribute profiles are unknown, the memberships and the ideal responses will be jointly estimated. Specifically, the optimization problem associated with the GNPC method in Chiu et al. (2018) aims to minimize the following loss function over the membership $\boldsymbol{\alpha}_i$ and the weights $w_{j,\boldsymbol{\alpha}}$ under the constraints imposed by the given Q-matrix:

$$\sum_{\boldsymbol{\alpha}\in\{0,1\}^K}\sum_{i:\boldsymbol{\alpha}_i=\boldsymbol{\alpha}}\sum_{j=1}^{J}(R_{i,j}-\eta_{j,\boldsymbol{\alpha}}^{(w)})^2, \tag{5}$$

under constraint (1), where $\eta_{j,\boldsymbol{\alpha}}^{(w)}$ is given in (2).

A modified GNPC method was studied by Ma, de la Torre, et al. (2023) under a general framework where the item parameters $\theta_{j,\boldsymbol{\alpha}}$ are treated as a certain "centroid." In their framework, the item parameters $\theta_{j,\boldsymbol{\alpha}}$ and latent attributes $\boldsymbol{\alpha}_i$ are obtained by minimizing $L(\mathbf{A},\boldsymbol{\Theta}) = \sum_{\boldsymbol{\alpha}\in\{0,1\}^K}\sum_{i:\boldsymbol{\alpha}_i=\boldsymbol{\alpha}} l(\boldsymbol{R}_i,\boldsymbol{\theta}_{\boldsymbol{\alpha}})$, where $l(\boldsymbol{R}_i,\boldsymbol{\theta}_{\boldsymbol{\alpha}})$ is a loss function that measures the distance between the ith subject's response vector, $\boldsymbol{R}_i = (R_{i,j}, j = 1,\ldots,J)$, and the item parameter vector $\boldsymbol{\theta}_{\boldsymbol{\alpha}} = (\theta_{j,\boldsymbol{\alpha}}, j = 1,\ldots,J)$, given a membership $\boldsymbol{\alpha}$. Under their framework, GNPC method can be derived by taking $l(\boldsymbol{R}_i,\boldsymbol{\theta}_{\boldsymbol{\alpha}}) = \sum_{j=1}^{J}(R_{i,j}-\theta_{j,\boldsymbol{\alpha}})^2$, which leads to minimizing the following loss function:

$$\sum_{\boldsymbol{\alpha}\in\{0,1\}^K}\sum_{i:\boldsymbol{\alpha}_i=\boldsymbol{\alpha}}\sum_{j=1}^{J}(R_{i,j}-\theta_{j,\boldsymbol{\alpha}})^2, \tag{6}$$

with respect to $\theta_{j,\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}$ under constraint (1). To ensure identifiability, we impose the natural constraint $\theta_{j,\boldsymbol{\alpha}} \geq \theta_{j,\tilde{\boldsymbol{\alpha}}}$ if $\boldsymbol{\alpha}\succeq\tilde{\boldsymbol{\alpha}}$. Here $\boldsymbol{\alpha}\succeq\tilde{\boldsymbol{\alpha}}$ if $\alpha_k \geq \tilde{\alpha}_k$ for all $k\in[K]$.

Note that given the membership $\boldsymbol{\alpha}$, the item parameter $\theta_{j,\boldsymbol{\alpha}}$ that minimizes the loss function (6) takes exactly the form of $\overline{R}_{j,\boldsymbol{\alpha}}$ in (4) for all items and $\boldsymbol{\alpha}$'s. Inspired by this, as shown in Ma, de la Torre, et al. (2023), we can see that the solution $(\hat{\boldsymbol{\alpha}}_i,\hat{\eta}_{j,\boldsymbol{\alpha}})$ to the original GNPC estimation method in (5) is the same as the solution $(\hat{\boldsymbol{\alpha}}_i,\hat{\theta}_{j,\boldsymbol{\alpha}})$ to (6) under constraint (1) and the following additional constraint:

$$\left(\boldsymbol{\alpha}\cdot\mathbf{q}_j = 0 \Longrightarrow \theta_{j,\boldsymbol{\alpha}} = 0\right) \text{ and } \left(\boldsymbol{\alpha}\cdot\mathbf{q}_j = K_j \Longrightarrow \theta_{j,\boldsymbol{\alpha}} = 1\right), \tag{7}$$

where the additional constraint (7) corresponds to the constraint (3) under the GNPC setting.

Following the above discussion, both the original GNPC method and the modified GNPC method can be formulated in a unified estimation framework (Ma, de la Torre, et al., 2023) of minimizing (6) under different constraints. In particular, since $\sum_{i=1}^{N}\sum_{j=1}^{J}(R_{i,j}-\theta_{j,\boldsymbol{\alpha}_i})^2 = \sum_{\boldsymbol{\alpha}\in\{0,1\}^K}\sum_{i:\boldsymbol{\alpha}_i=\boldsymbol{\alpha}}\sum_{j=1}^{J}(R_{i,j}-\theta_{j,\boldsymbol{\alpha}})^2$, we can rewrite (6) equivalently as the following loss function:

$$\ell(\mathbf{A},\boldsymbol{\Theta}|\mathbf{R}) = \sum_{i=1}^{N}\sum_{j=1}^{J}(R_{i,j}-\theta_{j,\boldsymbol{\alpha}_i})^2, \tag{8}$$

where minimizing the loss function (8) with respect to $(\mathbf{A},\boldsymbol{\Theta})$ under the constraints (1) and (7) obtains the original GNPC estimators in Chiu and Köhn (2019a) and the modified GNPC estimators in Ma, de la Torre, et al. (2023) can be obtained by minimizing (8) under the constraint (1) only.

2.3. Limitations of existing consistency results for nonparametric CDMs

Existing theoretical research has offered valuable insights into the practical utility of nonparametric methods. It has been shown that the NPC estimators are statistically consistent for estimating subjects' latent attributes under certain CDMs (Wang & Douglas, 2015). Similarly, the GNPC estimator's ability to consistently classify subjects has been established (Chiu & Köhn, 2019a). However, current theoretical assurances for these nonparametric methods come with their own set of limitations.

A fundamental assumption for the NPC method to yield a statistically consistent estimator of $\mathbf{A}$ is that $\mathbb{P}(R_{i,j} = 1|\eta_{j,\boldsymbol{\alpha}} = 0) < 0.5$ and $\mathbb{P}(R_{i,j} = 1|\eta_{j,\boldsymbol{\alpha}} = 1) > 0.5$, where $\eta_{j,\boldsymbol{\alpha}}$ represents the binary ideal responses (either 0 or 1) under the considered model (Wang & Douglas, 2015). However, as previously pointed out, this binary ideal response becomes restrictive when working with more complex CDMs. The binary ideal response, limited to representing the complex latent attribute patterns of examinees through two states, could potentially oversimplify the actual complexity of the scenario. This limitation, in turn, constrains the practical application of the NPC method in instances where the underlying true model is more sophisticated. For instance, Chiu et al. (2018) provided an illustrative example highlighting this restriction, showing the possibility of misclassifications when the underlying true model is the saturated GDINA model.

Although the GNPC method addresses the oversimplification problem of the NPC method, a new restrictive assumption emerges in the existing theory for the GNPC method. Specifically, Theorem 1 in Chiu and Köhn (2019a) assumes initialization of the memberships $\hat{\boldsymbol{\alpha}}_i^{(0)}$s that consistently estimates the ground truth in order to establish the consistency theory for GNPC. Similarly, Ma, de la Torre, et al. (2023) assumes the existence of a calibration dataset that provides consistent estimations $\widehat{\mathbf{A}}_c$ for the true latent class membership $\mathbf{A}_c^0$ of the calibration subjects. Under these assumptions, $\widehat{\eta}_{j,\boldsymbol{\alpha}}^{(w)}$ can be estimated using consistent membership estimations, which further support the consistency theory. The assumption concerning the existence of an initial set of consistent estimates or a calibration dataset may be restrictive and hard to satisfy in practice. To address this issue, we present new theoretical results demonstrating that the consistency of the GNPC method can be established without the need for a consistent initialization or a calibration dataset. These findings are detailed in the subsequent section.

3. Main results

Based on the unified framework of two GNPC methods outlined in Section 2, we will establish the theoretical properties of both the original GNPC method (Chiu & Köhn, 2019a) and the modified GNPC method (Ma, de la Torre, et al., 2023) under less stringent conditions. Regarding implementation, estimation algorithms for both the original and modified GNPC methods have been detailed in Chiu et al. (2018) and Ma, de la Torre, et al. (2023), respectively. We recommend using these well-established methods for estimation.

Before delving into the statistical behaviors of the aforementioned general nonparametric estimators, we outline the needed regularity conditions. Consider a model sequence indexed by (N,J), where both N and J tend to infinity, while K is held constant. For clarity, let the true parameters generating the data be represented as $(\boldsymbol{\Theta}^0,\mathbf{Q}^0,\mathbf{A}^0)$, and other true parameters are also denoted with superscript 0. Assumptions are made on these true parameters as follows.

Assumption 1. There exists $\delta > 0$ such that

$$\min_{1\le j\le J}\left\{\min_{\boldsymbol{\alpha}\circ\mathbf{q}_j\neq\tilde{\boldsymbol{\alpha}}\circ\mathbf{q}_j}(\theta^0_{j,\boldsymbol{\alpha}}-\theta^0_{j,\tilde{\boldsymbol{\alpha}}})^2\right\}\ge\delta.$$

Assumption 2. There exist $\{\delta_J : \delta_J > 0\}_{J=1}^{\infty}$ and a constant $\varepsilon > 0$ such that

$$\min_{1\le k\le K}\frac{1}{J}\sum_{j=1}^{J}\mathbb{1}\{\mathbf{q}_j^0=\mathbf{e}_k\}\ge\delta_J; \tag{9}$$

$$\min_{\boldsymbol{\alpha}\in\{0,1\}^K}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{\boldsymbol{\alpha}_i^0=\boldsymbol{\alpha}\}\ge\varepsilon. \tag{10}$$

Assumption 1 serves as an identification condition for local latent classes at each item level, ensuring that the item parameters of different local latent classes, influenced by $\mathcal{K}_j$, are sufficiently distinct. The gap, denoted as δ, measure the separation between latent classes, thereby quantifying the strength of the signals. In the finite-J regime, a $\delta > 0$ is required for studying identifiability (Gu & Xu, 2019, 2023; Köhn & Chiu, 2017; Xu & Shang, 2018). Assumption 2 pertains to the discrete structures of $\mathbf{Q}$ and $\mathbf{A}$. Here, (10) implies that the 2^K latent patterns are not too unevenly distributed in the sample. An equivalent requirement in random-effect latent class models is $p_{\boldsymbol{\alpha}} > 0$ for all $\boldsymbol{\alpha} \in \{0,1\}^K$, where $p_{\boldsymbol{\alpha}}$ represents the population proportion of latent pattern $\boldsymbol{\alpha}$. For Assumption 2, within the finite-J regime, (9) is similar to the requirement that "$\mathbf{Q}$ should contain an identity submatrix $\mathbf{I}_K$" (Chen et al., 2015; Xu & Shang, 2018). However, as J approaches infinity, a finite number of submatrices $\mathbf{I}_K$ in $\mathbf{Q}$ may not be sufficient to ensure estimability and consistency. Therefore, (9) necessitates that $\mathbf{Q}$ includes an increasing number of identity submatrices, $\mathbf{I}_K$, as J grows. A similar assumption on $\mathbf{Q}$ was made by Wang and Douglas (2015) when they were establishing the consistency of the NPC method. It is worth mentioning that the lower bound δ_J in (9) in Assumption 2 is allowed to decrease to zero as J goes to infinity.

In the following subsections, we study the consistency properties of the modified GNPC method with the constraint (1) and the original GNPC method with both constraints (1) and (7). As the modified GNPC method involves less constraints compared to the original GNPC method, for convenience, we first present results for the modified GNPC method in Section 3.1 and then for the original GNPC method in Section 3.2.

3.1. Consistency results for modified GNPC

In this section, we discuss the consistency results for the modified GNPC method. The following main theorem first validates the consistency of the modified GNPC method under the constraint (1) and provides a bound for its rate of convergence in recovering the latent attribute profiles. We use $O(\cdot)$ and $o(\cdot)$ to denote the big-O and small-o notations, respectively, and $O_p(\cdot)$ and $o_p(\cdot)$ as their probability versions for convergence in probability.

Theorem 1 (Consistency of modified GNPC method). *Consider the* $(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{\Theta}}) = \arg\min_{(\boldsymbol{A},\boldsymbol{\Theta})} \ell(\boldsymbol{A}, \boldsymbol{\Theta} | \boldsymbol{R})$ *under the constraint (1). When* $N, J \to \infty$ *jointly, suppose* $\sqrt{J} = O\left(N^{1-c}\right)$ *for some small constant* $c \in (0,1)$. *Under Assumptions 1 and 2, the classification error rate is*

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{\widehat{\boldsymbol{\alpha}}_i \neq \boldsymbol{\alpha}_i^0\} = o_p\left(\frac{(\log J)^{\tilde{\varepsilon}}}{\delta_J \sqrt{J}}\right), \tag{11}$$

where for a small positive constant $\tilde{\varepsilon} > 0$.

Theorem 1 bounds the error of the estimator $\widehat{\mathbf{A}}$, which establishes the consistency of the latent attributes of the nonparametric method, and even allows the rate δ_J to go to zero. Theorem 1 also offers insight into the accuracy of estimating $\mathbf{A}$ with finite samples and finite J. In particular, if δ_J is a constant, then the finite sample error bound in (11) becomes $o_p((\log J)^{\tilde{\varepsilon}}/\sqrt{J})$. Ignoring the log terms, the result shows that the classification error rate can be dominated by the order of $J^{-1/2}$, indicating that a longer item set facilitates more accurate classification for the latent profiles of all subjects. Note the scaling condition that $N\exp(-Jt^2) \to 0$ for any positive fixed $t > 0$ in Chiu and Köhn (2019a) and Wang and Douglas (2015) essentially requires the growth rate of J to be at least the order of $\log N$. In contrast, our scaling condition only assumes that the number of items goes jointly with N at a slower rate, which can be more easily satisfied.

The following corollary demonstrates that under certain conditions, the item parameters can be consistently estimated via the modified GNPC method as $N, J \to \infty$.

Theorem 2 (Item Parameters Consistency). *Under Assumptions 1 and 2 and the scaling conditions given in Theorem 1, we have the following uniform consistency result for all* $j \in [J]$ *and* $\boldsymbol{\alpha} \in \{0,1\}^K$*:*

$$\max_{j,\boldsymbol{\alpha}} \left|\widehat{\theta}_{j,a} - \theta^0_{j,\boldsymbol{\alpha}}\right| = o_p\left(\frac{1}{\sqrt{N^{1-\tilde{c}}}}\right) + o_p\left(\frac{(\log J)^{\tilde{\varepsilon}}}{\delta_J\sqrt{J}}\right),$$

where $\tilde{c}$ *and* $\tilde{\varepsilon}$ *are small positive constants.*

This theorem builds on the consistency result established in Theorem 1 to establish the uniform consistency in parameter estimation. The condition $\sum_{i=1}^N \mathbb{1}\{\boldsymbol{\alpha}_i^0 = \boldsymbol{\alpha}\} \geq N\varepsilon$ for all $\boldsymbol{\alpha} \in \{0,1\}^K$ ensures that there are enough samples within each class to provide accurate estimates of item parameters. This is reflected in the first error term $o_p(1/\sqrt{N^{1-\tilde{c}}})$, which achieves nearly optimal $\sqrt{N}$-consistency. Notably, the added $\tilde{c}$ term arises due to the number of parameters going to infinity jointly with the sample size N, causing a slight deviation from the optimal error rate of $O_p(1/\sqrt{N})$. The maximum deviation $\max_{j,\boldsymbol{\alpha}} |\widehat{\theta}_{j,\boldsymbol{\alpha}} - \theta^0_{j,\boldsymbol{\alpha}}|$ is also influenced by the classification errors for the unknown latent attributes, which is shown in the second error term $o_p((\log J)^{\tilde{\varepsilon}}/(\delta_J\sqrt{J}))$. In conclusion, the upper bound for the maximal error in estimating item parameters comprises a term that denotes nearly optimal $\sqrt{N}$-consistency, accompanied by an additional term related to the errors in classifying the latent attributes. Our theory suggests that both the sample size and the test length need to be sufficiently large to ensure accurate estimation of the item parameters, given that the latent attributes of the subjects must also be estimated.

3.2. Consistency results for original GNPC

In this section, we discuss the consistency result of the original GNPC method. Since the original method adds an additional constraint (7) compared to the modified method, which causes some of the parameters $\theta_{j,\boldsymbol{\alpha}}$ to be 0 or 1, additional notations are needed to characterize how this potential variation affects the consistency outcome. Denote

$$\lambda^2_{N,J} = \frac{1}{NJ}\sum_{j=1}^J \left(\sum_{i:\boldsymbol{\alpha}_i^0\cdot\mathbf{q}_j=0} (\theta^0_{j,\boldsymbol{\alpha}_i^0} - 0)^2 + \sum_{i:\boldsymbol{\alpha}_i^0\cdot\mathbf{q}_j=K_j} (1 - \theta^0_{j,\boldsymbol{\alpha}_i^0})^2 \right), \tag{12}$$

which represents the average squared distance between the true parameters and the associated zero/one values. To establish the consistency for the original GNPC method, an additional assumption is needed.

Assumption 3. For any $j \in [J]$, we have $\theta^0_{j,\boldsymbol{\alpha}=\mathbf{0}} < 1/2 < \theta^0_{j,\boldsymbol{\alpha}=\mathbf{1}}$.

Assumption 3 plays a similar role to Assumption 1, as both measure the separation between different latent classes. While this appears to be a relatively mild condition and may seem similar to the one presented in Wang and Douglas (2015), as discussed in Section 2, it remains applicable to complex CDMs. The following theorem validate the consistency of the NPC method under the original GNPC setting, provides a similar bound as Theorem 1 for the misclassification rate.

Theorem 3 (GNPC Consistency). *Consider the* $(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{\Theta}}) = \arg\min_{(\boldsymbol{A},\boldsymbol{\Theta})} \ell(\boldsymbol{A}, \boldsymbol{\Theta}|\boldsymbol{R})$ *under the constraints (1) and (7). When* $N,J \to \infty$ *jointly, suppose* $\sqrt{J} = O\left(N^{1-c}\right)$ *for some small constant* $c \in (0,1)$*. Under Assumptions 2 and 3, the classification error rate is*

$$\frac{1}{N}\sum_{i=1}^N \mathbb{1}\{\widehat{\boldsymbol{\alpha}}_i \neq \boldsymbol{\alpha}_i^0\} \leq o_p\left(\frac{(\log J)^{\tilde{\varepsilon}}}{\delta_J\sqrt{J}}\right) + \frac{4\lambda^2_{N,J}}{\delta_J}.$$

The classification error rate for the original GNPC method is slightly different from the result given in Theorem 1. An extra item $4\lambda^2_{N,J}/\delta_J$ is added into the error rate. This additional term reflects the number of items that violate constraint (7), as (12) will be larger when there are more items with $\theta_{j,\boldsymbol{\alpha}}$ that is neither 0 nor 1. The impact of the additional error introduced by $\lambda_{N,J}$ is further illustrated by Example

1 and the simulation studies in Section 4. The details of Theorem 3's proof can be found in Section C of the Supplementary Material.

It is worth mentioning that without further regularity conditions, it might be challenging to avoid the additional error term $4\lambda_{N,J}^2/\delta_J$. In the existing consistency results for both the NPC and the modified GNPC methods (Chiu & Köhn, 2019a; Wang & Douglas, 2015), a crucial step involves ensuring that for each examinee i, the true attribute profile minimizes $\mathbb{E}[d_i(\boldsymbol{\alpha}_m)]$ across all m. Here, $d_i(\boldsymbol{\alpha}_m) = \sum_{j=1}^{J} d(R_{i,j}, \widehat{\eta}_{j,\boldsymbol{\alpha}_m})$ represents the distance functions used in the respective nonparametric methods. A similar approach is required in the proof of Theorem 1 for the modified GNPC method. If we denote $\overline{\ell}(\mathbf{A},\boldsymbol{\Theta}) = \mathbb{E}[\ell(\mathbf{A},\boldsymbol{\Theta}|\mathbf{R})]$ and $\overline{\ell}(\mathbf{A}) = \inf_{\boldsymbol{\Theta}} \overline{\ell}(\mathbf{A},\boldsymbol{\Theta})$, then the true latent class profiles, $\mathbf{A}^0$, are found to minimize $\overline{\ell}(\mathbf{A})$.

One challenge in establishing the consistency for the original GNPC method lies in the fact that, with the inclusion of the constraint (7), the true latent class profiles $\mathbf{A}^0$ might not necessarily minimize $\overline{\ell}(\mathbf{A})$. Let $\tilde{\mathbf{A}} = \arg\min_{\mathbf{A}} \overline{\ell}(\mathbf{A})$, it can be intuitively understood that $\widehat{\mathbf{A}}$ might approach $\tilde{\mathbf{A}}$ more closely than $\mathbf{A}^0$. Thus the additional error term originates from the discrepancy between $\overline{\ell}(\mathbf{A}^0)$ and $\overline{\ell}(\tilde{\mathbf{A}})$. Indeed, in the proof of Theorem 3, we employ the following upper bound to account for this deviation:

$$\lambda_{N,J}^2 \geq \frac{1}{NJ}\left(\overline{\ell}(\mathbf{A}^0) - \overline{\ell}(\tilde{\mathbf{A}})\right). \tag{13}$$

The above inequality in (13) is sharp up to a constant multiple of $\lambda_{N,J}^2$, below is an illustrative example.

Example 1. In this example, we assume that the number of sample size N is $8M$ for some positive integer M, the number of items J is 4, and the dimension of latent attribute profiles K is also 4. We further assume that the four corresponding row vectors for the items in the Q-matrix are $\boldsymbol{q}_1 = (1,0,0,1)$, $\boldsymbol{q}_2 = (1,1,0,0)$, $\boldsymbol{q}_3 = (0,1,1,0)$, and $\boldsymbol{q}_4 = (0,0,1,1)$, where $\boldsymbol{q}_j$ encodes the required latent attributes for the jth item. For the true latent attribute profiles of the $8M$ samples, it is assumed that $4M$ samples exhibit latent attribute profile $(1,1,1,1)$, while the remaining $4M$ display the profile $(0,0,0,0)$. It is noteworthy that all parameters in the original GNPC method will be treated as exactly zero or one under the true latent attribute profiles $\mathbf{A}^0$ in this example, as stipulated by the constraint (7). The last assumption in this example is that there exists some $\lambda \in (0,1/2)$ such that $\theta^0_{j,\boldsymbol{\alpha}=0} = 1/2 - \lambda$ and $\theta^0_{j,\boldsymbol{\alpha}=1} = 1/2 + \lambda$. Under these assumptions, the expected loss under the true latent attribute profiles satisfies

$$\overline{\ell}(\mathbf{A}^0) - \sum_{i=1}^{N}\sum_{j=1}^{J} P_{i,j}(1-P_{i,j}) = (NJ)\left(\frac{1}{2} - \lambda\right)^2, \tag{14}$$

where $P_{i,j} := \mathbb{P}(R_{i,j} = 1)$ are true item response parameters, independent of the estimation process. The derivation of (14) is detailed in Section D of the Supplementary Material. To demonstrate the sharpness of inequality (13), we construct an alternative set of latent attribute profiles, denoted as $\mathbf{A}^1$. This set contains $2M$ samples of $\mathbf{e}_k \in \{0,1\}^4$ for each $k \in [4]$, where each $\mathbf{e}_k$ only contains the kth latent attribute. For instance, $\mathbf{e}_1 = (1,0,0,0)$, $\mathbf{e}_2 = (0,1,0,0)$, and so on. There is a correspondence between the true latent profiles $\mathbf{A}^0$ and the constructed $\mathbf{A}^1$. Specifically, for the $2M$ samples assigned to $\mathbf{e}_k$ within $\mathbf{A}^1$, the true latent attribute profiles are equally divided, with half being $(0,0,0,0)$ and the other half $(1,1,1,1)$. Hence, the expected loss under the constructed latent attribute profiles fulfills

$$\overline{\ell}(\mathbf{A}^1) - \sum_{i}\sum_{j} P_{i,j}(1-P_{i,j}) = (NJ)\cdot\left(\lambda^2 + \frac{1}{8}\right). \tag{15}$$

The derivation of (15) is detailed in Section D of the Supplementary Material. Thus, we have $(NJ)^{-1}\left(\overline{\ell}(\mathbf{A}^0) - \overline{\ell}(\mathbf{A}^1)\right) = -\lambda + 1/8$. Note that in this example $\lambda_{N,J}^2 = (\lambda - 1/2)^2$. If $\lambda \leq 1/13$, then one can easily verify that $-\lambda + 1/8 > \lambda_{N,J}^2/4$, and therefore, in this case, we deduce that

$$\lambda_{N,J}^2 \geq \frac{1}{NJ}\left(\overline{\ell}(\mathbf{A}^0) - \overline{\ell}(\tilde{\mathbf{A}})\right) \geq \frac{1}{NJ}\left(\overline{\ell}(\mathbf{A}^0) - \overline{\ell}(\mathbf{A}^1)\right) \geq \frac{\lambda_{N,J}^2}{4},$$

which implies the order $\lambda_{N,J}^2$ in the inequality in (13) is sharp. The details of the proof can be found in Section D of the Supplementary Material.

The magnitude of the additional classification error term arising from the aforementioned discrepancy is of the order $O((NJ)^{-1}(\overline{\ell}(\mathbf{A}^0)-\overline{\ell}(\tilde{\mathbf{A}})))$. As demonstrated in Example 1, $O(\lambda_{N,J}^2)$ provides a tight estimation of the order of the discrepancy $(NJ)^{-1}(\overline{\ell}(\mathbf{A}^0)-\overline{\ell}(\tilde{\mathbf{A}}))$. Therefore, the additional error term $4\lambda_{N,J}^2/\delta_J$ in Theorem 3 may not be significantly reducible.

4. Simulation study

In this section, we conduct a comprehensive simulation study to illustrate our theoretical findings of both the original GNPC (Chiu et al., 2018) and the modified GNPC (Ma, de la Torre, et al., 2023). Note that in the existing literature (Chiu et al., 2018; Ma, de la Torre, et al., 2023), various numerical studies have already demonstrated the effectiveness of both the original GNPC method and the modified GNPC method in small sample settings. Therefore, our focus here primarily lies on scenarios where both the sample size and test length are relatively large to illustrate our theoretical results.

For the data-generating process, followed by the simulation design of Chiu et al. (2018) and Ma, de la Torre, et al. (2023), we consider two settings: (1) items are simulated using the DINA model, and (2) items are simulated from GDINA model, as detailed in Section 2. The manipulated conditions include: the sample size $N \in \{300,600,1,000\}$; the test length $J \in \{50,100,200,300,400,500\}$; the number of latent attributes $K \in \{3,5\}$. For $K = 3$, the Q-matrix is constructed with two identity $K \times K$ submatrices, and the remaining items are generated uniformly from all possible non-zero patterns. It is worth mentioning that this generating process adheres to (9) in Assumption 2. For the case of $K = 5$, the Q-matrix is restricted to contain items that measure up to three attributes and constructed the same way as that for $K = 3$. For the data conforming to the DINA model, we simulate s_j and g_j independently from a uniform distribution $Unif[0,r]$ with $r \in \{0.2,0.4\}$. For data generated under the GDINA model, the item parameters are simulated following the framework outlined in Chiu et al. (2018) as follows. For any item j, let $K_j^* = \sum_{k=1}^{K} q_{jk}$ be the number of required attributes of item j, where q_{jk} is the (j,k)th entry in the Q-matrix. Without loss of generality, we assume that these attributes with $q_{jk} = 1$ are the first K_j^* attributes. For instance, if $K_j^* = 3$, that is, item j requires three attributes, we then denote the possible proficiency classes as $\boldsymbol{\alpha}_1^* = (000)$, $\boldsymbol{\alpha}_2^* = (100)$, $\boldsymbol{\alpha}_3^* = (010)$, $\boldsymbol{\alpha}_4^* = (110)$, $\boldsymbol{\alpha}_5^* = (001)$, $\boldsymbol{\alpha}_6^* = (101)$, $\boldsymbol{\alpha}_7^* = (011)$, and $\boldsymbol{\alpha}_8^* = (111)$. The item parameters for item j are specified by the probabilities of making the correct responses for all $\boldsymbol{\alpha}_i^*$ with $1 \le i \le 8$. If $K_j^* = 2$, we only need to specify the probabilities for $\boldsymbol{\alpha}_i^*$ with $1 \le i \le 4$ since the remaining attributes are irrelevant for distinguishing among the proficiency classes, and if $K_j^* = 1$, we only need to specify the probabilities for $\boldsymbol{\alpha}_1^*$ and $\boldsymbol{\alpha}_2^*$ (Chiu et al., 2018). Analogous to the data generation process under the DINA model, we simulate two noise levels under the GDINA model as in Ma, de la Torre, et al. (2023), with item parameters provided in Table 1 for small noises and Table 2 for large noises, respectively. Note that Table 1 contains seven rows, while Table 2 contains six, each row representing a distinct set of item parameters. For each noise level, the set of item parameters for each item j is sampled randomly from those rows with $K^* = K_j^*$ in each table.

For the latent attribute patterns, they are generated using either a uniform setting, where each proficiency class is drawn with a uniform probability of 2^{-K}, or a multivariate normal threshold model as described by Chiu et al. (2018). In this model, each subject's attribute profile is linked to a latent continuous ability vector $\mathbf{z} \sim \mathcal{N}_K(0,\boldsymbol{\Sigma})$. The diagonal elements of $\boldsymbol{\Sigma}$ are fixed at 1, while the off-diagonal elements are set to 0.3 for a low-correlation scenario and 0.7 for a high correlation scenario. The attribute profile is then derived from $\mathbf{z}$ by applying a truncation process as follows:

$$\alpha_{ik} = \begin{cases} 1, & z_{ik} \ge \Phi^{-1}\left(\dfrac{k}{K+1}\right), \\ 0, & \text{otherwise}, \end{cases}$$

where Φ is the cumulative distribution function of the standard normal distribution.

Table 1. Item response parameters for GDINA with small noises, where K^* denotes the number of required attributes of a considered item

	$P(\boldsymbol{\alpha}_1^*)$	$P(\boldsymbol{\alpha}_2^*)$	$P(\boldsymbol{\alpha}_3^*)$	$P(\boldsymbol{\alpha}_4^*)$	$P(\boldsymbol{\alpha}_5^*)$	$P(\boldsymbol{\alpha}_6^*)$	$P(\boldsymbol{\alpha}_7^*)$	$P(\boldsymbol{\alpha}_8^*)$
	0.2	0.9						
$K^* = 1$	0.1	0.8						
	0.1	0.9						
	0.2	0.5	0.4	0.9				
$K^* = 2$	0.1	0.3	0.5	0.9				
	0.1	0.2	0.6	0.8				
$K^* = 3$	0.1	0.2	0.3	0.4	0.4	0.5	0.7	0.9

Table 2. Item response parameters for GDINA with large noises, where K^* denotes the number of required attributes of a considered item

	$P(\boldsymbol{\alpha}_1^*)$	$P(\boldsymbol{\alpha}_2^*)$	$P(\boldsymbol{\alpha}_3^*)$	$P(\boldsymbol{\alpha}_4^*)$	$P(\boldsymbol{\alpha}_5^*)$	$P(\boldsymbol{\alpha}_6^*)$	$P(\boldsymbol{\alpha}_7^*)$	$P(\boldsymbol{\alpha}_8^*)$
$K^* = 1$	0.3	0.7						
	0.3	0.8						
	0.3	0.4	0.7	0.8				
$K^* = 2$	0.3	0.4	0.6	0.7				
	0.2	0.3	0.6	0.7				
$K^* = 3$	0.2	0.3	0.3	0.4	0.4	0.5	0.6	0.7

To illustrate our theoretical results, we compute the pattern-wise agreement rate (PAR):

$$\text{PAR} = \frac{1}{N}\sum_{i=1}^{N} I\{\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i\}.$$

We apply the original GNPC and modified GNPC estimation methods under all manipulated scenarios. Following the algorithms proposed by Chiu et al. (2018) and Ma, de la Torre, et al. (2023), both methods are initialized using latent attributes estimated with the NPC method for computational efficiency. In each scenario, we conduct 100 replications and calculate the mean value of the PARs. The iteration process is terminated when $\frac{1}{N}\sum_{i=1}^{N} I\{\hat{\boldsymbol{\alpha}}_i^{(t-1)} \neq \hat{\boldsymbol{\alpha}}_i^{(t)}\} < 0.001$ or exceeding the maximal number of iterations set as 500. Additionally, we conducted simulations in these scenarios where both methods are initialized with random latent attributes, resulting in estimation errors similar to those obtained with NPC initialization. Details and results are provided in Section E of the Supplementary Material.

Figures 1–4 present the PAR results when the data are generated under the DINA model, and the PAR results under the GDINA model are shown in Figures 5–8. In each figure, the upper panel presents the estimation results using the original GNPC method, while the lower panel displays the results using the modified GNPC method. From left to right, the subfigures in each row illustrate the estimations for latent attributes, simulated under three different settings: uniform, low correlation setting, and high correlation setting.

In general, both the original and modified GNPC methods perform well under all the model settings. The modified GNPC method exhibits a slight edge in more complex scenarios, as demonstrated in Figures 4 and 8. A consistent trend across all figures is that as the test length J increases, the PARs improve, supporting our theory that the upper bound for classification error decreases with J. When the data are simulated from the GDINA models, there is a slight increase in classification errors for both methods compared to those generated using the DINA models. In addition, comparisons between

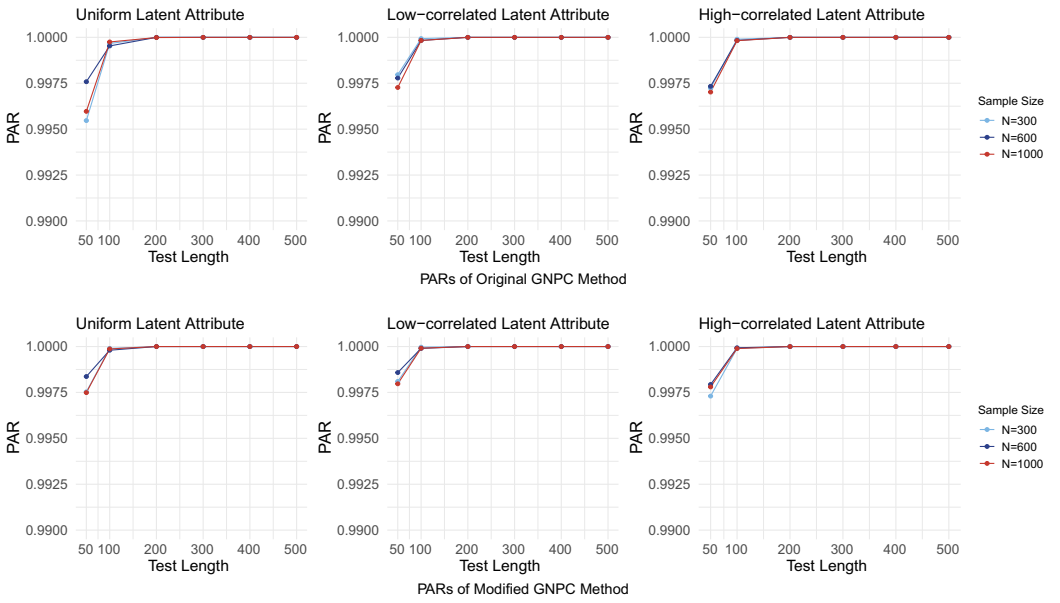

Figure 1. PARs when the data are generated using the DINA model with $K = 3$ and $r = 0.2$.

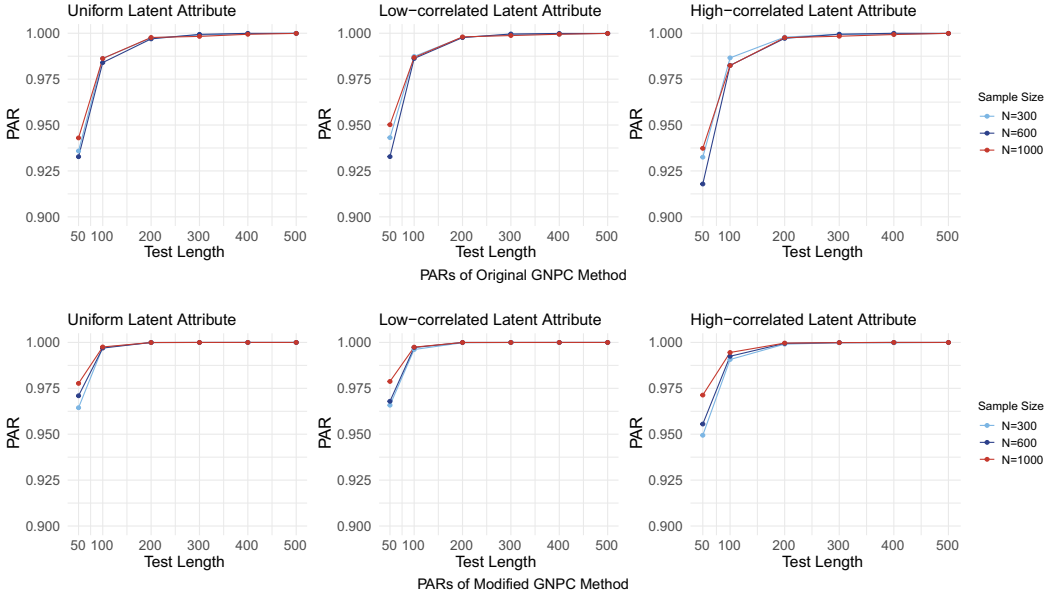

Figure 2. PARs when the data are generated using the DINA model with $K = 3$ and $r = 0.4$.

figures with lower noise levels (Figures 1, 3, 5, and 7) and those with higher ones (Figures 2, 4, 6, and 8) reveal lower classification errors with decreased noise. In particular, Figures 1 and 5 show nearly perfect classification results under low noise and $K = 3$ settings. Moreover, increasing the number of latent attributes typically results in less precise estimation, as evidenced by the comparisons between the settings of $K = 3$ (Figures 1, 2, 5, and 6) and $K = 5$ (Figures 3, 4, 7, and 8). Within each figure, a

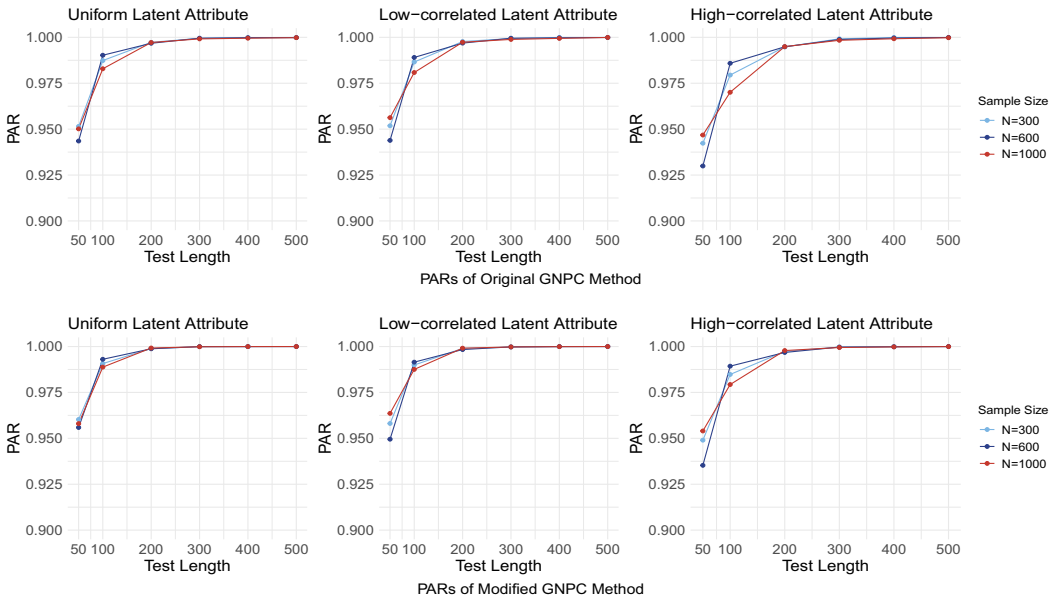

Figure 3. PARs when the data are generated using the DINA model with $K = 5$ and $r = 0.2$.

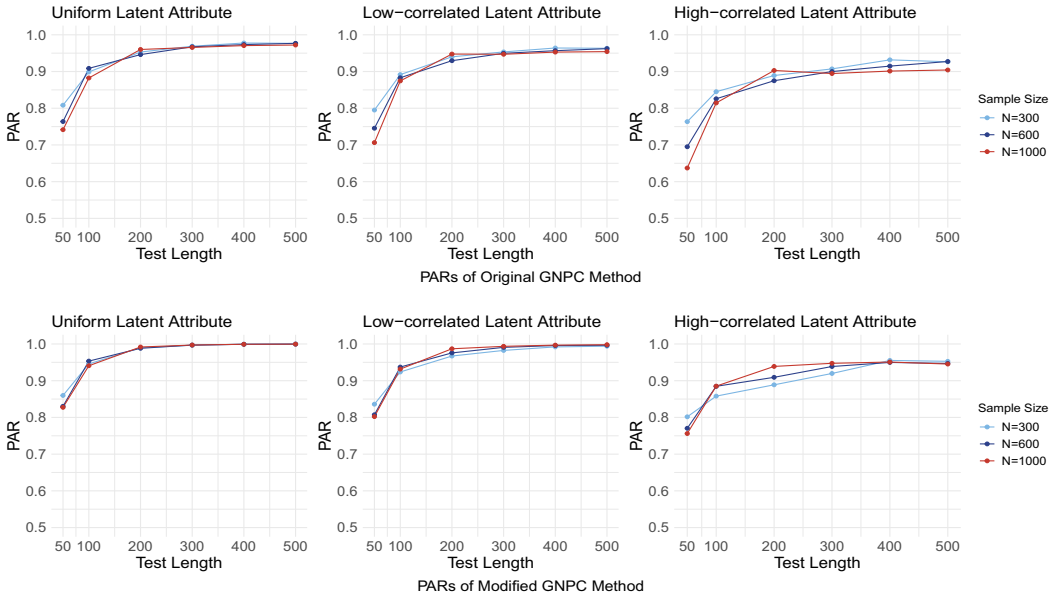

Figure 4. PARs when the data are generated using the DINA model with $K = 5$ and $r = 0.4$.

slight decrease in PARs is observed when the latent attributes exhibit a higher correlation. When the data are simulated under larger noises and more attributes (Figure 4 and 8), PARs from the original GNPC method appear not converging to 1 even when the sample size is 1,000 and test length is 500. This is likely attributable to the additional error term related to $\lambda_{N,J}/\delta_J$ in Theorem 3. Notably, $\lambda_{N,J}$ in (12) can become large when a significant proportion of $\theta_{j,\boldsymbol{\alpha}_i^0}$ fails to satisfy the constraint (7).

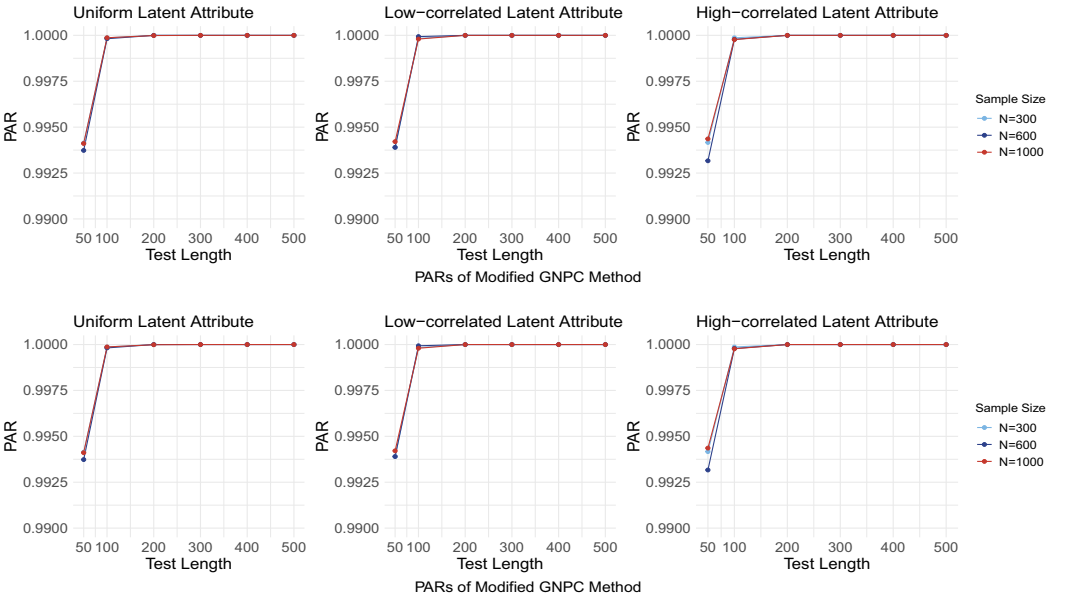

Figure 5. PARs when the data are generated using the GDINA model with small noises and $K = 3$.

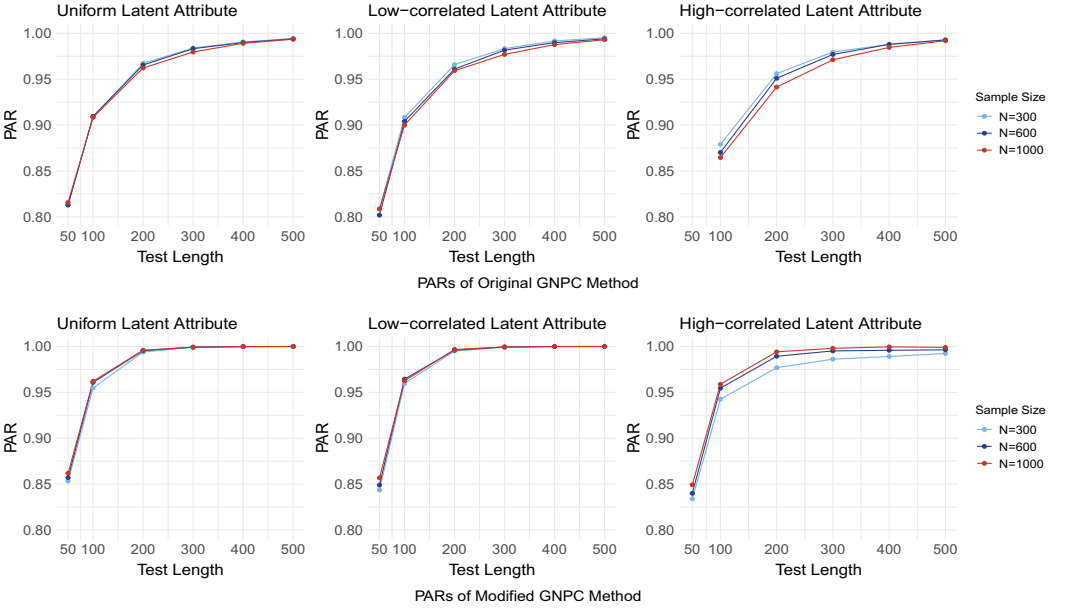

Figure 6. PARs when the data are generated using the GDINA model with large noises and $K = 3$.

5. Discussion

In this work, we revise the consistency results for the GNPC method, originally offered in Chiu and Köhn (2019a), under relaxed and more practical assumptions. We deliver finite sample error bounds for the considered two versions of the GNPC method. These bounds not only guarantee asymptotic consistency in estimating the latent profiles of subjects but also offer insights into the precision of these estimates in small sample situations. Furthermore, we derive uniform convergence of item response

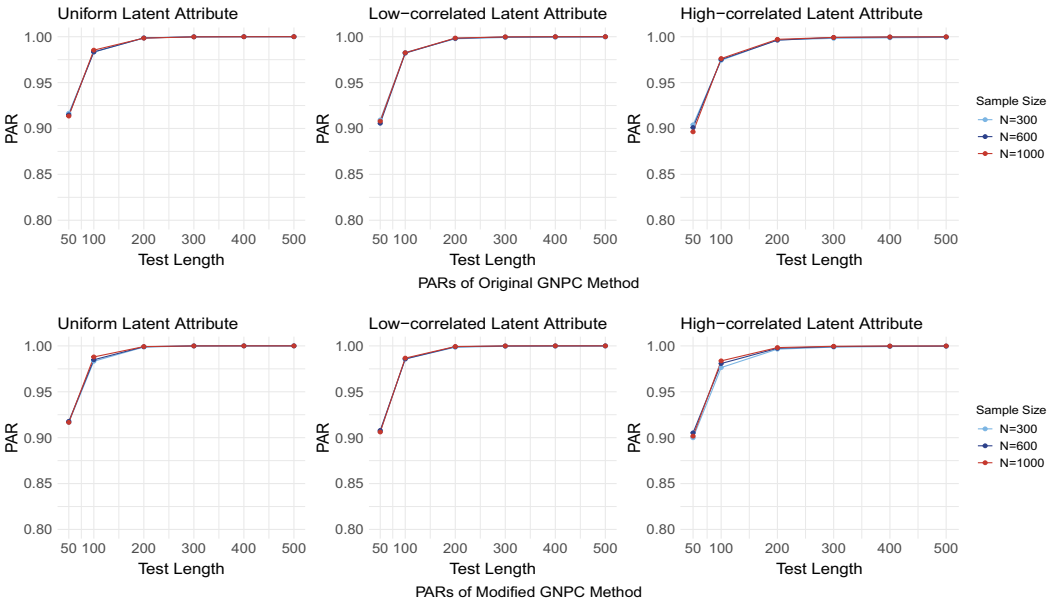

Figure 7. PARs when the data are generated using the GDINA model with small noises and $K = 5$.

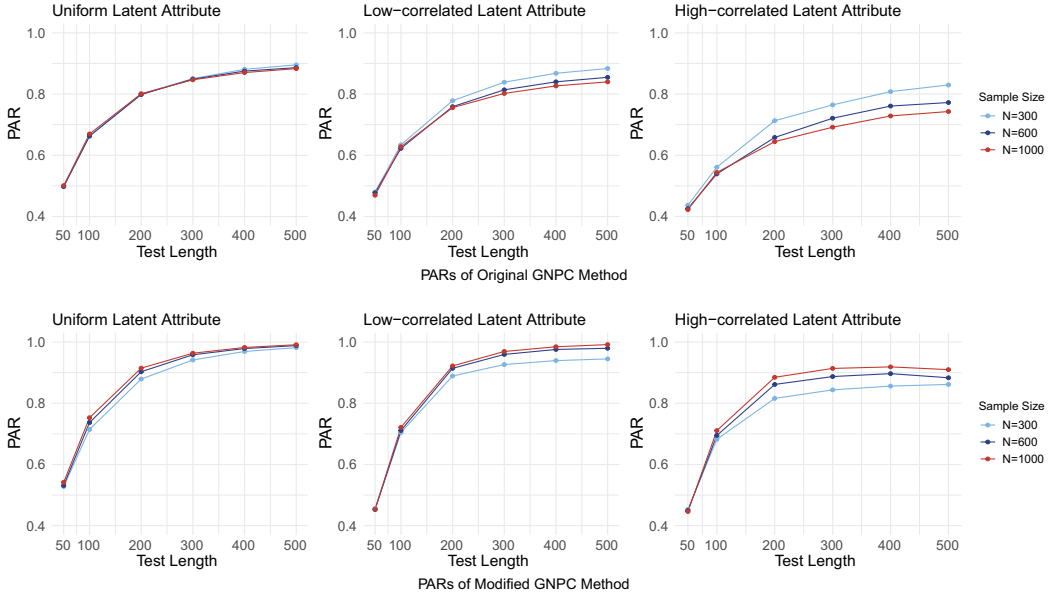

Figure 8. PARs when the data are generated using the GDINA model with large noises and $K = 5$.

parameters $\widehat{\boldsymbol{\Theta}}$ for the modified GNPC method. Notably, all of these advancements are achieved without the requirement for a calibration dataset.

The findings in this study open up several possibilities for future exploration. Using the consistency and finite sample error bounds established for estimating the discrete latent structure $\mathbf{A}$, future work can examine statistical inference on CDMs with a large number of test items and latent attributes. Additionally, it is important to note that in practical situations, the Q-matrix may not always be readily available. Various estimation techniques have been proposed in the literature (Chen et al., 2015, 2018;

Gu & Xu, 2023; Köhn et al., 2024; Li et al., 2022; Liu et al., 2012; Ma, Ouyang, et al., 2023; Xu & Shang, 2018). This leads to a potential future direction of developing theories and computational methods for CDMs estimation with an unknown Q-matrix within the nonparametric framework.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2025.9.

Acknowledgments. We thank the editor, associate editor, and three anonymous referees for their careful review and valuable comments.

Funding statement. This research was partially supported by NSF SES-2150601 and SES-1846747.

Competing interests. The authors declare none.

References

Chandía, E., Sanhueza, T., Mansilla, A., Morales, H., Huencho, A., & Cerda, G. (2023). Nonparametric cognitive diagnosis of profiles of mathematical knowledge of teacher education candidates. *Current Psychology*, *42*, 32498–32511.

Chen, Y., Culpepper, S. A., Chen, Y., & Douglas, J. (2018). Bayesian estimation of the DINA Q-matrix. *Psychometrika*, *83*, 89–108.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*, 850–866.

Chiu, C.-Y. & Chang, Y.-P. (2021). Advances in CD-CAT: The general nonparametric item selection method. *Psychometrika*, *86*, 1039–1057.

Chiu, C.-Y. & Douglas, J. A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, *30*, 225–250.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633–665.

Chiu, C.-Y. & Köhn, H.-F. (2015). Consistency of cluster analysis for cognitive diagnosis: The DINO model and the DINA model revisited. *Applied Psychological Measurement*, *39*(6), 465–479.

Chiu, C.-Y. & Köhn, H.-F. (2019a). Consistency theory for the general nonparametric classification method. *Psychometrika*, *84*, 830–845.

Chiu, C.-Y. & Köhn, H.-F. (2019b). Nonparametric methods in cognitively diagnostic assessment. In: M. von Davier & YS. Lee (Eds.), *Handbook of Diagnostic Classification Models. Methodology of Educational Measurement and Assessment*. (pp. 107–132). Springer.

Chiu, C.-Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika*, *83*, 355–375.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

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*, 281–296.

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). *Review of cognitively diagnostic assessment and a summary of psychometric models*. In C. Rao & S. Sinharay (Eds.), *Psychometrics, Volume 26 of Handbook of Statistics*, (pp. 979–1030). Elsevier.

George, A. C. & Robitzsch, A. (2015). Cognitive diagnosis models in R: A didactic. *The Quantitative Methods for Psychology*, *11*, 189–205.

Gu, Y. & Xu, G. (2019). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika*, *84*, 468–483.

Gu, Y. & Xu, G. (2023). A joint MLE approach to large-scale structured latent attribute analysis. *Journal of the American Statistical Association*, *118*, 746–760.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. PhD thesis, ProQuest Information & Learning.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210.

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.

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.

Köhn, H.-F., Chiu, C.-Y., Oluwalana, O., Kim, H., & Wang, J. (2024). A two-step Q-matrix estimation method. *Applied Psychological Measurement*, *49*, 3–28.

Li, C., Ma, C., & Xu, G. (2022). Learning large Q-matrix by restricted Boltzmann machines. *Psychometrika*, *87*(3), 1010–1041.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, *36*(7), 548–564.

Ma, C., de la Torre, J., & Xu, G. (2023). Bridging parametric and nonparametric methods in cognitive diagnosis. *Psychometrika*, *88*, 51–75.

Ma, C., Ouyang, J., & Xu, G. (2023). Learning latent and hierarchical structures in cognitive diagnosis models. *Psychometrika*, *88*(1), 175–207.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconception in the pattern classification approach. *Journal of Educational and Behavioral Statistics*, *12*, 55–73.

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

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–301.

von Davier, M. & Lee, Y.-S. (2019). *Handbook of diagnostic classification models*. Springer International Publishing.

Wang, D., Ma, W., Cai, Y., & Tu, D. (2023). A general nonparametric classification method for multiple strategies in cognitive diagnostic assessment. *Behavior Research Methods*, *56*, 723–735.

Wang, S. & Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika*, *80*, 85–100.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics*, *45*(2), 675–707.

Xu, G. & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association*, *113*, 1284–1295.

Cite this article: Cui, C., Liu, Y. and Xu, G. (2025). Consistency Theory of General Nonparametric Classification Methods in Cognitive Diagnosis. *Psychometrika*, 1–17. https://doi.org/10.1017/psy.2025.9