Statistica Applicata - Italian Journal of Applied Statistics Vol. 34 (2)
doi.org/10.26398/IJAS.0034-011

Evaluation of distance measures for nonparametric classification in small-scale educational settings

Angelos Markos, Evripidis Themelis [1]
Department of Primary Education, Democritus University of Thrace, Alexandroupoli, Greece

Stratos Moschidis
Hellenic Statistical Authority, Greece

Abstract *Cognitive diagnostic models (CDMs) are a type of latent class models that link observable data, such as questionnaire responses, to categorical latent variables, often dichotomous in nature. CDMs are frequently used in educational testing to provide diagnostic information by identifying an examinee's level of mastery in a set of predetermined skills or attributes. However, parametric CDM estimation requires a relatively large sample size, which is often not feasible in assessments designed to inform classroom learning. To address this issue, non-parametric or algorithmic approaches have been developed that classify examinees by minimizing the distance between observed responses and expected responses for a given mastery profile. This study aims to evaluate a set of L_2 distance measures for use in non-parametric classification in small-scale educational settings. Results from simulations showed that the squared χ^2 distance outperformed or performed equally well to the commonly used Euclidean squared distance. Additionally, test length was found to be a crucial factor in classification performance, with tests containing more items and/or fewer attributes being preferred to compensate for small sample sizes. An analysis of fraction subtraction data is provided as an example. Recommendations for the use of non-parametric CDMs in the classroom are provided.*

Keywords: *educational testing, cognitive diagnosis, non-parametric classification, GNPC*

1. Introduction

Cognitive diagnostic testing has emerged in educational assessment to bridge the gap between psychometric theories of latent trait measurement and cognitive theories for problem-solving (Leighton and Gierl, 2007; Zhang et al., 2023). The cognitive diagnostic testing process assumes that an individual's response to a question is related to their mastery of specific skills or knowledge, referred to

[1] Angelos Markos, amarkos@eled.duth.gr

as attributes. The goal of this testing is to classify individuals into discrete attribute mastery classes based on their responses to a series of carefully crafted assessment items. Cognitive diagnostic tests are designed to assess the specific knowledge a learner has or has not acquired, rather than quantifying their overall level of knowledge with a single score. This discrete characterization is beneficial in various ways, such as identifying errors (e.g., misconceptions or lack of skill/knowledge) that impede problem-solving, creating diagnostic profiles of an individual's mastery level of specific skills, and providing guidance for both learners and teachers on areas of improvement and remediation (Zhang et al., 2023).

A class of statistical models known as cognitive diagnostic models (CDMs; Leighton and Gierl, 2007) or diagnostic classification models (DCMs; Rupp et al., 2010) have been developed to generate categorical classifications for multiple latent attributes using scored assessment responses. CDMs are distinct from Classical Test Theory (CTT) methods or Item Response Theory (IRT) models in that the latent variables in CDMs are discrete or categorical (i.e., indicating mastery or non-mastery) rather than continuous. In other words, CDMs assign test-takers to multidimensional attribute profiles by classifying them as masters versus nonmasters of each atrribute involved in any given test, whereas CTT/IRT models assign them scores on continuous scales representing more general abilities. Therefore, CDMs may be preferred in situations where a more detailed assessment of specific skills is desired.

CDMs are confirmatory in nature, meaning that the skills required to perform well on the test are specified in advance according to a substantive theory, then tested against the real data. Let's assume that ability in a given domain is a composite of K latent binary attributes. There are then 2^K distinct attribute profiles composed of these K attributes representing 2^K distinct proficiency classes. CDMs can be understood as restricted latent class models or probabilistic confirmatory multidimensional models with categorical latent attributes, where the model parameters are constrained by pre-assumed relationships between the test items and the latent attributes they assess. These relationships are specified through a binary matrix known as the $\mathbf{Q}$-matrix. This matrix is similar to a loading matrix in factor analysis with a complex loading structure, meaning there are items that load to more than one factor. The correct specification of the $\mathbf{Q}$-matrix is crucial for providing accurate diagnostic results for each test taker. Therefore, the matrix is usually provided by experts such as those who developed the items, and the attributes should be defined at an adequately detailed level.

Numerous studies have focused on detecting distinct patterns of attribute

mastery with CDMs (see Sessoms and Henson, 2018, for a review of the literature on specific applications of CDMs). In mathematics education, Bradshaw et al. (2014) and Izsák et al. (2019) employed CDMs to diagnose middle-grades teachers' multiplicative reasoning and proportional reasoning, respectively. Additionally, an online individualized tutor program based on the cognitive diagnostic reports outperformed a traditional remedial instruction program (Wu, 2019). Ravand (2016) and Ravand and Robitzsch (2018) illustrated the application of CDMs to the reading comprehension data of high-stakes tests. More recently, CDMs were used to assess students' progressions of understanding of energy in the physical science domain (Zhou and Traynor, 2022). There is some evidence that cognitive diagnostic feedback can promote students' learning and is more effective than correct-incorrect response feedback in promoting learning, especially in more challenging areas of knowledge (Tang and Zhan, 2021).

Several probabilistic models for cognitive diagnosis have been developed and widely applied in practice. Popular examples include the Deterministic Input Noisy Output "AND" gate (DINA) model (Junker and Sijtsma, 2001), the Deterministic Input Noisy Output "OR" gate (DINO) model (Templin and Henson, 2006), the Linear Logistic Model (LLM; Maris, 1999), the Additive CDM (ACDM; De La Torre, 2011), the Reduced Reparametrized Unified Model (RRUM; Hartz, 2002), the General Diagnostic Model (GDM; von Davier, 2005, 2008), the Loglinear Cognitive Diagnosis Model (LCDM; Henson et al., 2009), and the Generalized DINA (GDINA) model (De La Torre, 2011).

The aforementioned models mainly differ in the way they represent the relationship between attributes and item responses. Disjunctive models, such as the DINO model, assume that if an individual has mastered at least one of the attributes specified for an item, the probability of a correct response will be high. Conversely, conjunctive models, such as the DINA model, require mastery of all attributes for an item to achieve the highest probability of a correct response. General CDMs, such as the G-DINA, provide the most flexibility by relaxing the potentially restrictive assumptions of the DINA model. The current methods for fitting such models include marginal maximum likelihood estimation that utilizes the Expectation Maximization algorithm (MMLE-EM) and Markov Chain Monte Carlo (MCMC) techniques. However, parametric CDM estimation requires a large sample size, typically several hundred examinees, which is much larger than what is typical for assessments designed to guide classroom learning (For recent simulation studies, see Chiu and Chang, 2021; Chiu et al., 2018; Paulsen and Valdivia, 2021; Sen and Cohen, 2021). This makes parametric mod-

els less suitable for providing meaningful feedback about learning in small-scale educational settings, where formative guidance from CDMs may be particularly beneficial.

In response to these difficulties, non-parametric techniques have been developed to classify examinees into attribute profiles or groups (Chiu and Chang, 2021; Chiu and Douglas, 2013; Chiu and Köhn, 2019). Although they do not have the same flexible probabilistic background as parametric models, non-parametric CDMs require no statistical parameter estimation and they are often less computationally expensive and more efficient with small sample sizes. Based on an initial $I \times J$ student response matrix, where each value indicates if student $i \in 1,2,\ldots,I$ answered item $j \in 1,2,\ldots,J$ correctly, Ayers et al. (2008) first derived the so-called "capability matrix", a $I \times K$ matrix that shows for each attribute the proportion of correct answers for all items tried by each student involving that attribute. Then, they proposed to apply K-means clustering or the Gaussian mixture model to the capability matrix to group examinees into different clusters with the same attribute profiles. Similar approaches involve the application of hierarchical/K-means clustering (Chiu and Douglas, 2013) and spectral clustering (Guo et al., 2020) on the matrix with the summed scores of each student on the K attributes on rows. One limitation of all these approaches is that they involve an additional cluster labeling step to obtain the attribute profiles, because cluster analysis does not provide labels for the derived clusters.

To overcome this issue, Chiu and Douglas (2013) proposed the non-parametric classification method (NPC), which classifies examinees by minimizing the Hamming distance between observed examinee responses and the "ideal" or expected responses for a given attribute profile that would be implied by the $\mathbf{Q}$-matrix (assuming no measurement error). The NPC calculates the ideal responses based on either the DINA or the DINO model, making it less suitable when the model underlying the data is more general. The General Non-parametric Classification (GNPC; Chiu et al., 2018) improves upon the NPC by computing a weighted version of the ideal response profiles of the DINA and DINO, allowing it to be used with more general CDMs. Among all ideal item response profiles, an examinee's attribute profile is the one with the minimum squared Euclidean distance to the examinee's observed item response vector. The GNPC has been shown to produce higher classification accuracy than parametric CDMs estimated using the EM algorithm for small sample sizes (Chiu et al., 2018; Ma and Jiang, 2021) and was used effectively in an algorithm for computerized adaptive testing (Chiu and Köhn, 2019). Here, we limit our focus to non-parametric CDMs for binary

(e.g., incorrect-correct) items, but a method for polytomously scored items has also been recently proposed (Wang et al., 2022).

We note two fundamental contributions of this paper to the existing body of literature. First, the paper aims to improve the GNPC algorithm by investigating the effect of different distance metrics on its performance. The original GNPC algorithm calculates the squared Euclidean distance between observed responses and responses that are expected, based on a complex loading structure that specifies the attributes required to answer an item correctly. The question addressed is whether the squared Euclidean distance is the optimal choice among other compatible distances for the GNPC objective. Second, this is the first study to provide a comprehensive evaluation of the GNPC performance in samples smaller than 30, similar in size to those found in classroom settings. To achieve this, various experimental conditions were considered through a series of simulations using a full factorial design.

The following sections are included in this paper: Section 2 reviews the NPC and GNPC algorithms. In Section 3, alternative distance measures from the L_2 family that can be used in the GNPC objective criterion are presented. Section 4 presents the results of a simulation study designed to evaluate the impact of different distance measures. The proposed methodology is illustrated on a real data set in Section 5. Section 6 discusses the results and concludes the paper.

2. Non-parametric cognitive diagnosis methods

In this section, we describe in detail the non-parametric classification method (NPC; Chiu and Douglas, 2013) and the general NPC (GNPC; Chiu et al., 2018). To start, we introduce the following notation which will be used throughout this paper.

Let $\mathbf{Q}$ denote a matrix of size $J \times K$, where J is the number of dichotomous items (i.e., correct/incorrect or 0/1) in a cognitive diagnostic test and K is the number of attributes or skills. The elements of the $\mathbf{Q}$-matrix are 0 or 1, where $q_{jk} = 1$ if the j^{th} item requires the k^{th} skill and $q_{jk} = 0$ otherwise. The $\mathbf{Q}$-matrix is typically created by the test developer and needs to be properly structured (see Köhn and Chiu, 2015, for the definition of the $\mathbf{Q}$-matrix completeness). The general form of the $\mathbf{Q}$-matrix is therefore:

$$\mathbf{Q} = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1K} \\ \vdots & \ddots & & \vdots \\ q_{J1} & q_{J2} & \cdots & q_{JK} \end{bmatrix}.$$

Let $\mathbf{Y}$ be a matrix of size $I \times J$, where I denotes the number of students or examinees. The elements of the $\mathbf{Y}$ matrix are 0 or 1, where $y_{ij} = 1$ if the i^{th} student answered the j^{th} item correctly and $y_{ij} = 0$ otherwise. It has the general form:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1J} \\ \vdots & \ddots & & \vdots \\ y_{I1} & y_{I2} & \cdots & y_{IJ} \end{bmatrix}.$$

We use $M = 2^K$ to denote the total number of proficiency latent classes (i.e., binary attribute profiles) and $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_M$ are the distinct profiles in which examinees can be classified, i.e.,

$\boldsymbol{\alpha}_1 = (0,0,\ldots,0_K), \boldsymbol{\alpha}_2 = (1,0,\ldots,0_K), \boldsymbol{\alpha}_3 = (0,1,\ldots,0_K), \ldots, \boldsymbol{\alpha}_K = (1,1,\ldots,1)$, where the k^{th} entry indicates whether the respective attribute has been mastered.

2.1. Non-parametric classification method

In the NPC method, the so-called ideal response profiles are calculated based on either the DINA or the DINO model. These profiles express the ideal answers that students would give if they belong to a certain attribute profile $\boldsymbol{\alpha}_m$, where $m \in \{1,2,\ldots,M\}$. The ideal responses for each item j and each latent attribute profile $\boldsymbol{\alpha}_m$ are defined as follows (Chiu and Douglas, 2013):

$$\eta_{jm}^{DINA} = \prod_{k=1}^{K} \alpha_{mk}^{q_{jk}} \text{ and } \eta_{jm}^{DINO} = 1 - \prod_{k=1}^{K} (1 - \alpha_{mk})^{q_{jk}}$$

for the DINA and the DINO model, respectively.

Here, $\boldsymbol{\eta}_m = (\eta_{1m}, \eta_{2m}, \ldots, \eta_{Jm})$ denotes the ideal response vector for the m^{th} attribute profile, where η_{jm} can be the DINA or DINO ideal response.

Given the ideal response vector of each attribute profile, an examinee is classified to the closest profile that minimizes the distance between his/her observed response vector $\boldsymbol{y}_i$, and the ideal response vector:

$$\hat{\boldsymbol{\alpha}}_i = \underset{m \in \{1,2,\ldots,M\}}{\arg\min} \; d(\boldsymbol{y}_i, \boldsymbol{\eta}_m). \tag{1}$$

where $d(\cdot)$ is a distance function. Chiu and Douglas (2013) used the Hamming distance, where

$$d_H(\boldsymbol{y}_i, \boldsymbol{\eta}_m) = \sum_{j=1}^{J} |y_{ij} - \eta_{mj}|$$

Given that the ideal responses implied by the DINA or DINO model are binary in nature, it can be observed that the absolute difference between the observed response and the ideal response will be zero if they are equal, and one otherwise. Furthermore, it can be demonstrated that utilizing the Euclidean distance will yield identical results as using the Hamming distance. Lastly, it is relatively straightforward to demonstrate that the NPC is equivalent to the 1-nearest neighbor classifier when utilizing the ideal response vectors as the training data, with each vector corresponding to a distinct attribute profile and the observed response vectors as the test data.

2.2. General NPC

The NPC method may be constrained by its reliance on the assumptions of the DINA or DINO model, which establish two distinct extremes. Specifically, the DINA model requires that all item attributes be present and mastered in order to endorse an item and produce the correct response, while the DINO model stipulates that mastery of at least one item attribute is sufficient. To address this limitation, the General NPC (GNPC) method proposed by Chiu et al. (2018) considers a more general ideal response, which is a weighted average of the ideal responses of the DINA and DINO models, as represented by the following equation:

$$\eta_{jm}^{(w)} = w_{jm}\eta_{jm}^{DINA} + (1 - w_{jm})\eta_{jm}^{DINO} \tag{2}$$

where w_{jm} is the weight for the j^{th} item and the m^{th} attribute profile. The weighted ideal response vector for the m^{th} attribute profile is denoted as $\boldsymbol{\eta}_m^{(w)} = (\eta_{1m}, \ldots, \eta_{Jm})$.

To estimate the weights, Chiu et al. (2018) proposed minimizing the squared Euclidean distance between the responses to item j and the weighted ideal responses $\boldsymbol{\eta}_m^{(w)}$:

$$d_{jm} = \sum_{i \in C_m} (y_{ij} - \eta_{jm}^{(w)})^2, \tag{3}$$

where $\{C_m\}_{m=1}^{M}$ is the partition of the subjects into M attribute profiles. The following cases can be distinguished:

- First case $\rightarrow \eta_{jm}^{DINO} = \eta_{jm}^{DINA} = 0 \Rightarrow \hat{\eta}_{jm}^{(w)} = 0$
- Second case $\rightarrow \eta_{jm}^{DINO} = \eta_{jm}^{DINA} = 1 \Rightarrow \hat{\eta}_{jm}^{(w)} = 1$

- Third case $\rightarrow \eta_{jm}^{DINA} = 1$ *and* $\eta_{jm}^{DINO} = 0 \Rightarrow \hat{\eta}_{jm}^{(w)} = \bar{y}_{jm}$

where $\bar{y}_{jm} = \dfrac{\sum\limits_{i \in C_m} y_{ij}}{N_m}$ and N_m is the number of students in C_m. It should be noted that in all cases, the estimator $\hat{\eta}_{jm}^{(w)}$ is independent of the weight w_{jm}.

The GNPC method begins with an initial partitioning of the examinees, typically provided by the NPC based on either the DINA or the DINO model (Chiu et al., 2018). An examinee is classified by minimizing the squared Euclidean distance between their observed responses and the ideal response vectors estimated from the previous step, $\hat{\boldsymbol{\alpha}}_i = \underset{m \in \{1,2,\ldots,M\}}{\arg\min}\ d(\boldsymbol{y}_i, \hat{\boldsymbol{\eta}}_m)$.

The GNPC algorithm is iterative in nature, beginning with initial values at step $t = 0$, and subsequently updating the estimates at the $(t+1)$-th step as follows:

$$\hat{\boldsymbol{\alpha}}_i^{(t+1)} = \underset{m \in \{1,2,\ldots,M\}}{\arg\min}\ d(\boldsymbol{y}_i, \hat{\boldsymbol{\eta}}_m^{(w)(t)}), \qquad \hat{\eta}_{jm}^{(w)(t+1)} = \bar{y}_{jm^{(t+1)}}.$$

The stopping criterion for the GNPC algorithm is typically $\frac{\sum_{i=1}^{N} I[\boldsymbol{\alpha}_i^{(t)} \neq \boldsymbol{\alpha}_i^{(t-1)}]}{N} < 0.001$, where $I[\cdot]$ is the indicator function.

3. The Squared L_2 distance family

As the GNPC is a distance-based approach, the selection of $d(\boldsymbol{y}_i, \hat{\boldsymbol{\eta}}_m^{(w)})$ is crucial for its performance, particularly in scenarios where it outperforms parametric methods, such as when dealing with small sample sizes characteristic of classroom assessments.

In this paper, we focus on the Squared L_2 distance family, also referred to as the χ^2 distance family, which encompasses a range of distance metrics including Squared Euclidean, Squared χ^2, Pearson's χ^2, Neyman's χ^2, Probabilistic Symmetric χ^2, Divergence, Clark, and Additive Symmetric χ^2 (Sung-Hyuk, 2007). It can be shown that all of these metrics have the same minimum in the context of the GNPC and therefore yield the same $\hat{\eta}_m^{(w)}$. This means that they are interchangeable and also satisfy the properties of stability, uniqueness, and minimization (see Chiu and Köhn, 2019, for a discussion of the statistical consistency of the GNPC). However, the Neyman χ^2 and the Additive Symmetric χ^2 distance cannot be used in the context of the GNPC, as they involve division by zero. Table 1 displays the expressions for each distance measure in the context of the GNPC. In other words, the distance metric in Eq. 3 can be replaced by any of these measures.

Table 1: Distances of Squared L_2 family

Distance metric	Formula
Squared χ^2	$\sum_{i \in C_m} \frac{(y_{ij} - \eta_{jm}^{(w)})^2}{y_{ij} + \eta_{jm}^{(w)}}$
Euclidean Squared	$\sum_{i \in C_m} (y_{ij} - \eta_{jm}^{(w)})^2$
Pearson χ^2	$\sum_{i \in C_m} \frac{(y_{ij} - \eta_{jm}^{(w)})^2}{\eta_{jm}^{(w)}}$
Probabilistic Symmetric χ^2	$2 \sum_{i \in C_m} \frac{(y_{ij} - \eta_{jm}^{(w)})^2}{y_{ij} + \eta_{jm}^{(w)}}$
Divergence	$2 \sum_{i \in C_m} \frac{(y_{ij} - \eta_{jm}^{(w)})^2}{(y_{ij} + \eta_{jm}^{(w)})^2}$
Clark	$\sqrt{\sum_{i \in C_m} \left(\frac{\lvert y_{ij} - \eta_{jm}^{(w)} \rvert}{y_{ij} + \eta_{jm}^{(w)}} \right)^2}$

The anticipated impact of various metrics on the process of attribute profile estimation is a topic of interest. It is noteworthy that the Pearson's χ^2 distance weighs the squared differences based on the inverse of the ideal responses. Consequently, the greater the discrepancy between the observed and ideal responses, the more significant the corresponding component in the distance calculation. The Squared χ^2 and Probabilistic Symmetric χ^2 distances, on the other hand, divide the squared differences between the observed and ideal responses by their absolute sum in order to emphasize similarities. Additionally, it is evident that Clark's distance employs a scaling mechanism in which smaller values are given greater weight when dividing by smaller values.

Table 2: Experimental factors considered in the simulation study

Factor	Values considered
Q-matrix	7×2, 10×2, 13×3, 20×3, 50×3, 100×3, 20×4, 50×4, 15×5, 30×5
model for data generation	GDINA, DINA, DINO, ACDM, LLM, RRUM, mixed scenario
N	10, 15, 20, 25, 30
attribute correlation	0.3, 0.5, 0.8

4. Simulation study

Data generation

A simulation study was conducted to investigate the effect of each of the six distance metrics in the Squared L_2 distance family on the performance of the GNPC. The subjects' true latent attribute patterns were generated using the multivariate normal threshold model, as described in Chiu et al. (2018). Each examinee's attribute profile was linked to a latent continuous ability vector $\boldsymbol{\theta} = \theta_1, \theta_2, \ldots, \theta_K \sim N(0, \boldsymbol{\Sigma})$, where the main diagonal of $\boldsymbol{\Sigma}$ was set to 1.00. The vectors were randomly sampled, and the kth entry of the attribute pattern $\boldsymbol{\alpha}_i$, α_{ik}, for examinee i was determined by

$$\alpha_{ik} = \begin{cases} 1 & \text{if } \theta_{ik} \geq \Phi^{-1}\frac{k}{K+1} \\ 0 & \text{otherwise} \end{cases},$$

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

Four factors were systematically manipulated, as illustrated in Table 2. The attribute correlation levels were based on empirical levels of correlations found in multiple domain assessments (Paulsen and Valdivia, 2021). This resulted in a total of $10 \times 7 \times 5 \times 3 = 1{,}050$ distinct data scenarios. In addition, 50 replications were made of each scenario, yielding a total of $52{,}500$ data sets. The data were generated under each condition using the `simGDINA()` function of the R package `GDINA` (Ma and de la Torre, 2020). The modified GNPC algorithm was implemented in an R function, which is available upon request from the first author. The initial partitioning of examinees was based on the NPC using the DINO model. The recovery of attribute profiles was evaluated using the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) between the estimated and true examinee

Table 3: Agreement between distance metrics based on Pearson's correlation and mean ARI values

	(2)	(3)	(4)	(5)	(6)	Mean ARI
(1) Squared χ^2	0.951	0.983	0.812	0.812	−0.066	0.542
(2) Euclidean Sq.	-	0.951	0.784	0.783	−0.052	0.537
(3) Prob. Sym. χ^2		-	0.812	0.811	−0.065	0.542
(4) Divergence			-	0.966	−0.044	0.468
(5) Clark				-	−0.043	0.468
(6) Pearson χ^2					-	0.111

Note: All correlations are significant at $p < 0.01$

profiles. Intuitively, the ARI measures the degree of agreement between an estimated partition and a reference partition. A value of 1 for the ARI represents perfect agreement, while values close to 0 indicate almost random recovery.

Overall performance. Table 3 presents the correlations between the ARI values of the six GNPC variants and the overall mean ARI of each method. All the correlations are large enough (the largest being Cor(Squared χ^2, Probabilistic Symmetric χ^2) = 0.983, followed by Cor(Divergence, Clark) = 0.966 and Cor(Squared χ^2, Euclidean Squared) = 0.951). This allows the assertion that one method may serve as a substitution or predictor for another method (for instance, both Squared χ^2 and Euclidean Sq. account for 99.14% of the variance in the other. Also notice that the Probabilistic Symmetric χ^2 distance is the Symmetric χ^2 distance multiplied by 2. The best performing distance metrics are Squared χ^2, Euclidean Squared and Prob. Sym. χ^2, followed by Divergence and Clark, whereas Pearson χ^2 had by far the poorest performance.

Performance by factor level. After examining the overall performance of the Squared L_2 family of distances, it is informative to determine if distance metric performance is dependent upon specific situations, or if performance varies with the factor levels. To investigate this, a repeated-measures ANOVA was conducted on the true attribute profile recovery, as outlined in Table 4. The between-dataset effects can be thought of as the influence of design factors across all distance metrics. To simplify the discussion, only main effects are modeled and discussed. Furthermore, given the large sample size, it was expected that most factors would be statistically significant; therefore, all effects were evaluated with respect to their estimated effect sizes. Effect size was evaluated using partial eta squared

(partial η^2), calculated as the ratio of the sum of squares of the effects (SS_{Effect}) to the sum of squares of both the effects and the error (SS_{Error}).

The size of the **Q**-matrix had the largest effect on GNPC performance. As the number of items increased and the number of attributes decreased, attribute profile recovery significantly improved, from an ARI of 0.152 for 15 items and 5 attributes to an ARI of 0.759 for 100 items and 3 attributes. Additionally, the model used to generate the data had a moderate effect on attribute profile recovery. Interestingly, overall attribute profile recovery was better when the underlying model was the DINO (ARI = 0.547) or the DINA model (ARI = 0.446), whereas the worst performance was observed for the mixed model scenario (ARI = 0.416). The correlation between attributes and the number of examinees had a very small effect on attribute performance recovery.

Based on the lower half of Table 4 (within-dataset effects), the effectiveness of distance metrics under different conditions was determined. Table 5 shows the mean ARI values by factor level. The effect of the distance metric on attribute profile performance was large. The top-performing group of distances includes the Squared Chi-squared, Euclidean Squared, and Probabilistic Symmetric Chi-squared distances, with overall mean ARIs of 0.542, 0.537, and 0.542, respectively. Divergence and Clark formed a second group of distance metrics with lower mean ARIs than those of the first group (0.468 and 0.469, respectively). The Pearson Chi-squared distance had the lowest attribute profile recovery performance (ARI = 0.113), indicating solutions that were almost random. The interaction between the distance metric used and the size of the **Q**-matrix also had a large effect on performance. However, a careful examination of the values in Table 5 shows that this is due to the behavior of the Pearson Chi-squared distance, which tends to perform even worse when all other metrics perform much better, that is, for tests with many items (50 or 100). A moderate effect was also observed for the interaction between the distance metric and the model underlying the simulated data. The group of the three top-performing distances exhibited a similar trend, with better attribute profile recovery when the underlying model was the DINO or the DINA model, followed by the RRUM, the ACDM, the LLM, the GDINA, and the MIXED model scenario. The Divergence and Clark distance metrics did not follow this trend and performed better when the underlying model was the DINO model, followed by the MIXED scenario, the LLM, the GDINA, the ACDM, the RRUM, and the DINA model. Lastly, the degree of correlation between attributes and the number of examinees appeared to have similar effects on all methods.

Table 4: Repeated-measures ANOVA for six distance metrics on ARI (factors are ordered by decreasing effect size, partial η^2).

Effect	Source	df	SS	F	partial η^2
	Q-matrix	9	9110.311	9082.436	0.614
Between	model	6	575.143	860.074	0.091
data sets	att. correlation	2	13.964	62.645	0.002
effects	N	4	6.485	14.548	0.001
	Error	5734.212			
	distance	5	7338.221	115337.445	0.692
Within	distance * matrix	45	2686.884	4692.301	0.451
data sets	distance * model	30	243.786	638.613	0.069
effects	distance * N	20	37.437	147.103	0.011
	distance * att. cor.	10	7.281	57.217	0.002
	Error	3273.451			

5. Empirical illustration

The data set at hand includes responses from 504 test-takers to 12 questions related to elementary probability theory (as cited in Heller and Wickelmaier (2013)). These questions assess the test-takers' ability to calculate the probability of the complement of an event (A1), the probability of two independent events (A2), the classic probability of an event (A3), and the probability of the union of two disjoint events (A4). The items range in difficulty, with some only requiring one attribute, while others require up to three attributes. The data set and corresponding **Q** matrix can be found in the R package `CDM` as `data.cdm05`.

As the true proficiency level of the test-takers is unknown, any evaluation must rely on relative standards. To establish such a standard, we used Additive CDM (ACDM) classification rates as a baseline, as this parametric model yielded the brst fit in a previous study (Philipp et al., 2018). The full data set was then split into 24 random subsets with 21 observations each, and re-analyzed with the GNPC based on the six different distance metrics. In this way, the full sample classification rates, which are expected to be more accurate, were used as a benchmark for evaluating the GNPC variants' performance with smaller samples.

The performance of the methods was assessed in terms of the patternwise agreement rate (PAR), which is essentially the percentage of correct examinee

Table 5: Attribute profile recovery of the GNPC based on the six different distance metrics by size of the Q-matrix, the underlying model used to generate the data, attribute correlation, and number of examinees (values of the Adjusted Rand Index)

Factor	Level	Sq. χ^2	Eucl. Sq.	Pr. S. χ^2	Pear. χ^2	Diverg.	Clark
	7×2	0.550	0.550	0.550	0.196	0.516	0.517
	10×2	0.585	0.581	0.585	0.157	0.500	0.499
	13×3	0.500	0.502	0.501	0.147	0.445	0.445
	15×5	0.171	0.172	0.171	0.105	0.145	0.146
Q	20×3	0.585	0.579	0.585	0.123	0.493	0.493
	20×4	0.346	0.344	0.345	0.138	0.292	0.292
	30×5	0.306	0.306	0.306	0.119	0.258	0.258
	50×3	0.840	0.820	0.840	0.049	0.693	0.693
	50×4	0.579	0.566	0.580	0.075	0.502	0.501
	100×3	0.957	0.948	0.957	0.023	0.834	0.834
	0.3	0.539	0.535	0.539	0.098	0.456	0.456
att.	0.5	0.542	0.537	0.542	0.108	0.465	0.466
correl.	0.8	0.544	0.538	0.545	0.128	0.483	0.482
	GDINA	0.514	0.503	0.514	0.109	0.446	0.447
	DINA	0.578	0.594	0.578	0.119	0.404	0.404
	DINO	0.635	0.630	0.635	0.135	0.625	0.624
model	ACDM	0.518	0.509	0.519	0.113	0.447	0.446
	LLM	0.518	0.509	0.518	0.109	0.454	0.453
	RRUM	0.534	0.530	0.534	0.111	0.431	0.430
	MIXED	0.496	0.482	0.496	0.081	0.47	0.471
N	10	0.517	0.510	0.517	0.118	0.479	0.479
	15	0.533	0.528	0.534	0.114	0.473	0.472
	20	0.547	0.543	0.547	0.111	0.467	0.466
	25	0.554	0.548	0.554	0.107	0.464	0.465
	30	0.558	0.554	0.558	0.104	0.458	0.458

classifications defined as

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

Table 6 reports the PAR values for each GNPC variant. The PAR values listed in the cells of the table were computed in comparing the classification results of ACDM on the full sample with those obtained for each of the six GNPC variants methods in the 24 subsets. For example, the PAR values in the first column resulted from comparing the classification of examinees obtained for each of the 24 subsets with GNPC (Squared χ^2 distance) with their classification when it was based on the full data set.

The findings suggest a high degree of average agreement (i.e., 0.91-0.92) for the Squared χ^2, Euclidean Squared, and Probabilistic Symmetric χ^2 distances. It should be noted that these were the three distance measures that performed the best according to the simulation study. Furthermore, a transition from large samples to small samples did not greatly affect the agreement between the GNPC and ACDM classifications for these distances. In concordance with the simulation study, the Divergence and Clark distances displayed significantly inferior performance, and the Pearson χ^2 distance had the lowest attribute profile recovery performance.

6. Discussion

In this paper, we conducted an extensive simulation study to evaluate the impact of distance metrics on the performance of the General Nonparametric Classification (GNPC) method. We chose to focus on the GNPC method as it has been shown to be the most effective algorithmic approach compared to parametric models, particularly for small samples where the latter do not converge or produce unstable results. The original GNPC algorithm calculates the squared Euclidean distance between observed responses and responses that would be expected based on a loading structure that specifies the attributes required to answer an item correctly. Our goal was to determine whether the squared Euclidean distance is the optimal choice among other compatible distances for the GNPC objective. Results indicated that at least one of the three distance metrics led to the highest attribute profile recovery. Similar findings were obtained in the analysis of a real data set. The Squared χ^2 distance and the Probabilistic Symmetric χ^2 distance are good alternatives and are recommended for use in the context of the GNPC, par-

Table 6: Pattern-wise agreement rates (PAR) between the results obtained for each of the six GNPC variants on 24 subsets and those of the ACDM on the full sample (treated as the true proficiency level).

	Sq. χ^2	Eucl. Sq.	Pr. S. χ^2	Diverg.	Clark	Pear. χ^2
1	0.9286	0.8810	0.8929	0.4286	0.6554	0.6429
2	0.8690	0.9286	0.9048	0.5952	0.6905	0.6905
3	0.8929	0.8810	0.8810	0.5357	0.6185	0.6310
4	0.9405	0.9167	0.9286	0.5238	0.7506	0.7381
5	0.9167	0.9286	0.9167	0.5238	0.6542	0.6667
6	0.9286	0.8571	0.9167	0.5119	0.5720	0.5595
7	0.9286	0.9405	0.9405	0.4762	0.7506	0.7381
8	0.9048	0.8929	0.9048	0.4524	0.6542	0.6667
9	0.9643	0.9286	0.9286	0.5833	0.7262	0.7262
10	0.9048	0.9643	0.9048	0.5357	0.7018	0.7143
11	0.9405	0.9405	0.9286	0.4524	0.7018	0.7143
12	0.9405	0.9048	0.8810	0.5357	0.7375	0.7500
13	0.9524	0.9286	0.9286	0.5595	0.8208	0.8333
14	0.9405	0.9048	0.9286	0.4643	0.6548	0.6548
15	0.9048	0.9167	0.8690	0.4643	0.5589	0.5714
16	0.9048	0.9405	0.9167	0.4762	0.6304	0.6429
17	0.9524	0.8929	0.9286	0.3690	0.6423	0.6548
18	0.9405	0.9286	0.9286	0.4524	0.8095	0.8095
19	0.8810	0.9048	0.9405	0.5595	0.7494	0.7619
20	0.9643	0.9167	0.9405	0.4881	0.7613	0.7738
21	0.9167	0.9167	0.9167	0.5595	0.6423	0.6548
22	0.8929	0.9167	0.9167	0.4881	0.7506	0.7381
23	0.9048	0.9762	0.9286	0.6071	0.7738	0.7738
24	0.9762	0.9524	0.9762	0.4286	0.8339	0.8214
Average	0.9246	0.9191	0.9187	0.5030	0.7017	0.7054

ticularly for small samples. To the best of our knowledge, this is the first study to evaluate the GNPC performance in samples smaller than 30, which are similar in size to classroom settings. In such settings, test length is another important factor to consider when using the GNPC for cognitive diagnostic assessment. Based on our study results, the GNPC performs adequately for tests with 50 items or more and 4 or fewer attributes. However, time constraints may prevent most educators from administering tests of such length.

Another constraint of the non-parametric approaches discussed in this paper pertains to the assumption that the **Q**-matrix is known and accurately specified by domain experts. It is well-known that a misspecified **Q**-matrix can negatively affect the classification of examinees (Kunina-Habenicht et al., 2012). While there are some methods for estimating the **Q**-matrix in the literature (Chen et al., 2018; Ren et al., 2021; Xu and Shang, 2018), developing methods for estimating CDMs with unknown **Q**-matrices is a next step that is left for future work. Additionally, in the CDM context, we assume that it is only the attributes involved that contribute to the difficulty of an item. Thus, given an item with a definite set of attributes associated with it, the difficulty of that item is also fixed, i.e., item difficulty is embedded in the **Q**-matrix. However, this static perspective may be unrealistic in the actual learning process, as human knowledge construction is dynamic and should be accounted for within cognitive diagnostic modeling (see, e.g., Gan et al., 2020, for a dynamic perspective). In addition, the GNPC can be easily extended to polytomous data (e.g., incorrect, partially correct, correct) or hierarchical structures among the latent attributes (Templin and Bradshaw, 2014), by considering suitable definitions of the ideal responses and appropriate distance measures between observed and ideal responses. Finally, given the relationship between the NPC and the 1-nearest neighbor classifier, we can consider a modified version of the GNPC based on efficient modifications of the KNN algorithm to deal with outliers (see, e.g., Liu and Chawla, 2011).

References

Ayers, E., Nugent, R., and Dean, N. (2008). Skill set profile clustering based on student capability vectors computed from online tutoring data. In EDM2008, ed., *1st International Conference on Educational Data Mining*. http://eprints.gla.ac.uk/47662/, Montreal, Canada.

Bradshaw, L., Izsák, A., Templin, J., and Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test

within the diagnostic classification framework. In *Educational measurement: Issues and practice*, 33 (1): 2–14.

Chen, Y., Culpepper, S.A., Chen, Y., and Douglas, J. (2018). Bayesian estimation of the dina q matrix. In *Psychometrika*, 83 (1): 89–108.

Chiu, C.Y. and Chang, Y.P. (2021). Advances in cd-cat: The general nonparametric item selection method. In *Psychometrika*, 86 (4): 1039–1057.

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

Chiu, C.Y. and Köhn, H.F. (2019). Consistency theory for the general nonparametric classification method. In *Psychometrika*, 85: 830–845.

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

De La Torre, J. (2011). The generalized dina model framework. In *Psychometrika*, 76 (2): 179–199.

Gan, W., Sun, Y., Peng, X., and Sun, Y. (2020). Modeling learner's dynamic knowledge construction procedure and cognitive item difficulty for knowledge tracing. In *Applied Intelligence*, 50 (11): 3894–3912.

Guo, L., Yang, J., and Song, N. (2020). Spectral clustering algorithm for cognitive diagnostic assessment. In *Frontiers in Psychology*, 11: 944.

Hartz, S.M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. University of Illinois at Urbana-Champaign.

Heller, J. and Wickelmaier, F. (2013). Minimum discrepancy estimation in probabilistic knowledge structures. In *Electronic Notes in Discrete Mathematics*, 42: 49–56.

Henson, R.A., Templin, J.L., and Willse, J.T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. In *Psychometrika*, 74 (2): 191–210.

Hubert, L. and Arabie, P. (1985). Comparing partitions. In *Journal Of Classification*, 2 (2): 193–218.

Izsák, A., Jacobson, E., and Bradshaw, L. (2019). Surveying middle-grades teachers' reasoning about fraction arithmetic in terms of measured quantities. In *Journal for Research in Mathematics Education*, 50 (2): 156–209.

Junker, B.W. and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. In *Applied Psychological Measurement*, 25 (3): 258–272.

Köhn, H.F. and Chiu, C.Y. (2015). Conditions of completeness of the q-matrix of tests for cognitive diagnosis. In L.A. van der Ark, D.M. Bolt, W.C. Wang, J.A. Douglas, and M. Wiberg, eds., *Quantitative Psychology Research: The 80th Annual Meeting of the Psychometric Society*. Springer, Beijing: 255–264.

Kunina-Habenicht, O., Rupp, A.A., and Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. In *Journal of Educational Measurement*, 49 (1): 59–81.

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

Liu, W. and Chawla, S. (2011). Class confidence weighted knn algorithms for imbalanced data sets. In *Pacific-Asia conference on knowledge discovery and data mining*, 345–356. Springer.

Ma, W. and de la Torre, J. (2020). GDINA: an R package for cognitive diagnosis modeling. In *Journal of Statistical Software*, 93: 1–26.

Ma, W. and Jiang, Z. (2021). Estimating cognitive diagnosis models in small samples: Bayes modal estimation and monotonic constraints. In *Applied Psychological Measurement*, 45 (2): 95–111.

Maris, E. (1999). Estimating multiple classification latent class models. In *Psychometrika*, 64 (2): 187–212.

Paulsen, J. and Valdivia, D.S. (2021). Examining cognitive diagnostic modeling in classroom assessment conditions. In *The Journal of Experimental Education*, 1–18.

Philipp, M., Strobl, C., de la Torre, J., and Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models. In *Journal of Educational and Behavioral Statistics*, 43 (1): 88–115.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. In *Journal of Psychoeducational Assessment*, 34 (8): 782–799.

Ravand, H. and Robitzsch, A. (2018). Cognitive diagnostic model of best choice: a study of reading comprehension. In *Educational Psychology*, 38 (10): 1255–1277.

Ren, H., Xu, N., Lin, Y., Zhang, S., and Yang, T. (2021). Remedial teaching and learning from a cognitive diagnostic model perspective: Taking the data distribution characteristics as an example. In *Frontiers in Psychology*, 12.

Rupp, A.A., Templin, J., and Henson, R.A. (2010). Diagnostic measurement. In *Theory, Methods, and Applications, New York: Guilford*.

Sen, S. and Cohen, A.S. (2021). Sample size requirements for applying diagnostic classification models. In *Frontiers in psychology*, 4050.

Sessoms, J. and Henson, R.A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. In *Measurement: Interdisciplinary Research and Perspectives*, 16 (1): 1–17.

Sung-Hyuk, C. (2007). Comprehensive survey on distance/similarity measures between probability density functions. In *International journal of Mathematical models and methods in applied sciences*, 4: 300–307.

Tang, F. and Zhan, P. (2021). Does diagnostic feedback promote learning? evidence from a longitudinal cognitive diagnostic assessment. In *AERA Open*, 7: 23328584211060804.

Templin, J. and Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. In *Psychometrika*, 79 (2): 317–339.

Templin, J.L. and Henson, R.A. (2006). Measurement of psychological disorders using cognitive diagnosis models. In *Psychological methods*, 11 (3): 287.

von Davier, M. (2005). A general diagnostic model applied to language testing data. In *ETS Research Report Series*, 2005 (2): 1–35.

von Davier, M. (2008). A general diagnostic model applied to language testing data. In *British Journal of Mathematical and Statistical Psychology*, 61 (2): 287–307.

Wang, Y., Chiu, C.Y., and Köhn, H.F. (2022). Nonparametric classification method for multiple-choice items in cognitive diagnosis. In *Journal of Educational and Behavioral Statistics*, 10769986221133088.

Wu, H.M. (2019). Online individualised tutor for improving mathematics learning: a cognitive diagnostic model approach. In *Educational Psychology*, 39 (10): 1218–1232.

Xu, G. and Shang, Z. (2018). Identifying latent structures in restricted latent class models. In *Journal of the American Statistical Association*, 113 (523): 1284–1295.

Zhang, S., Liu, J., and Ying, Z. (2023). Statistical applications to cognitive diagnostic testing. In *Annual Review of Statistics and Its Application*, 10.

Zhou, S. and Traynor, A. (2022). Measuring students' learning progressions in energy using cognitive diagnostic models. In *Frontiers in Psychology*, 13.