Journal of Classification 30 (2013)
DOI: 10.1007/s00357-013-9132-9

A Nonparametric Approach to Cognitive Diagnosis by Proximity to Ideal Response Patterns

Chia-Yi Chiu

Rutgers, The State University of New Jersey, USA

Jeff Douglas

University of Illinois at Urbana-Champaign, USA

Abstract: A trend in educational testing is to go beyond unidimensional scoring and provide a more complete profile of skills that have been mastered and those that have not. To achieve this, cognitive diagnosis models have been developed that can be viewed as restricted latent class models. Diagnosis of class membership is the statistical objective of these models. As an alternative to latent class modeling, a nonparametric procedure is introduced that only requires specification of an item-by-attribute association matrix, and classifies according to minimizing a distance measure between observed responses, and the ideal response for a given attribute profile that would be implied by the item-by-attribute association matrix. This procedure requires no statistical parameter estimation, and can be used on a sample size as small as 1. Heuristic arguments are given for why the nonparametric procedure should be effective under various possible cognitive diagnosis models for data generation. Simulation studies compare classification rates with parametric models, and consider a variety of distance measures, data generation models, and the effects of model misspecification. A real data example is provided with an analysis of agreement between the nonparametric method and parametric approaches.

Keywords: Cognitive diagnosis; Nonparametric classification; Residual sum.

Author's Address: C-Y. Chiu, Rutgers, The State University of News Jersey, Graduate School of Education, 10 Seminary Place, Room 326, New Brunswick NJ, 08901, e-mail: chia-yi.chiu@gse.rutgers.edu; Jeff Douglas, University of Illinois at urbana-Champaign, Department of Statistics, 116E Illini Hall, 725 S. Wright St., Champaign IL 61820; e-mail: jeffdoug@illinois.edu.

Published online

1. Introduction

In educational testing research, specialized latent class models for cognitive diagnosis have been developed to classify mastery or non-mastery of each attribute in a set of attributes the exam is designed to assess. The ultimate goal of applying diagnostic models is to classify subjects into one of several different categories describing their attribute profiles. These attributes can take many forms, depending on the application, but often correspond one-to-one with specific skills needed to answer items on an exam. Classification according to these fine-grained skills is desired when specific information on knowledge states is required, and one important expectation is that it can lead to more efficient remediation.

Specialized latent class models for cognitive diagnosis are derived under assumptions on which attributes are needed for which items, and how the attributes are utilized to construct a response, and recognize that data generally do not correspond to the ideal response patterns. Ideal response patterns are the specific item response patterns that would be observed if item responses corresponded exactly to the attributes an examinee possesses and the attributes required for an item. In the case of conjunctive models, the ideal response for an item would be correct if the examinee possessed all the items required for the item and would be incorrect otherwise. In the disjunctive case, having at least 1 of the required attributes would suffice for a correct response. To address the fact that data do not generally appear to correspond exactly to such ideal response patterns, stochastic elements are built into the models, that allow deviations from purely deterministic models, giving rise to incorrect answers when a correct one would be expected, or correct answers when an incorrect answer would be expected. A criticism of cognitive diagnosis modeling is that these terms that allow for random deviations from what one would expect can sometimes be large, which raises validity concerns about the assumed model and cognitive structure. Ideally, the stochastic terms should allow for some departures from ideal responses, but not so much that the theory becomes doubtful. However, when this is the case, the ideal response pattern may well be the most likely response, according to the likelihood function of the model, and classification based on deviations from the ideal responses can be effective, without making a single assumption about the parametric form of the model. The practical implication is that no difficult model fitting is required and simple and fast software may be used for classification. This makes cognitive diagnosis feasible in a much wider array of settings, perhaps all the way down to the classroom instruction level, especially considering that the performance of the method is independent of sample size.

The objective of this research is to utilize measures of distance between observed item response vectors and ideal response patterns that can provide effective nonparametric classification, under a wide variety of possibilities for the underlying cognitive diagnosis model responsible for generating the data. Classification done in this manner depends upon how much noise is expected in the data, and we study the parameter values of underlying models for which this nonparametric technique is useful. Its behavior is perfect when item responses are deterministic functions of the attribute profile, and the question becomes how far can we stretch the stochastic elements of models that may be responsible for the data, while still achieving accurate nonparametric classifications.

In the next section, a review of cognitive diagnosis models is given, followed by a section detailing the proposed nonparametric technique. The fourth section concerns a simulation study comparing nonparametric classification to maximum likelihood estimates derived under the true model as well as misspecified models. A real data analysis of fraction subtraction data is then presented to show how nearly model-based and nonparametric classification agree, and we conclude with a section discussing the implications of the results on practice.

2. Cognitive Diagnosis Models

Though the method we present can be viewed as nonparametric, aside from identifying which attributes are required for each item, we present a review of cognitive diagnosis models, particularly those that are used in the simulation section for comparisons to the nonparametric technique. Let the attribute profile vector $\boldsymbol{\alpha}$ be a K-dimensional vector for which entry k, α_k, indicates mastery or non-mastery of attribute k, for $k = 1, 2, \cdots, K$. The cognitive diagnosis models we discuss require specification of a $J \times K$ Q-matrix (Tatsuoka 1985). Entry q_{jk} denotes if item j requires the k^{th} attribute. The 2^K possible values of $\boldsymbol{\alpha}$ are the latent classes for which classification is desired. Models differ according to how subjects utilize their attributes to create responses.

An example of a conjunctive model is the DINA model (Junker and Sijtsma 2001). The item response function of the DINA model is,

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i) = (1 - s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})},$$

where for the ith subject, $s_j = P(Y_{ij} = 0|\eta_{ij} = 1)$ and $g_j = P(Y_{ij} = 1|\eta_{ij} = 0)$ are the probabilities of slipping and guessing, respectively, for the j^{th} item. Parameter η_{ij} is the ideal response which associates the attribute pattern possessed by the i^{th} subject and the elements of $\boldsymbol{Q}$ according to

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}.$$

The ideal response pattern η_{ij} indicates whether the i^{th} subject possesses all the attributes needed for answering the particular item. In fact, ideal response patterns will be defined in this manner for all the cognitive diagnosis models we consider, aside from the disjunctive DINO model described below. This is because when we strip away all the parameters that account for random responses, all models collapse into the same deterministic model from which ideal responses are defined.

The NIDA model, introduced in Maris (1999), differs from the DINA by defining slips and guesses at the subtask level. Let η_{ijk} indicate whether the i^{th} subject correctly applied the k^{th} attribute in completing the j^{th} item. Slipping and guessing parameters are indexed by attribute, and are defined by

$$\begin{aligned} s_k &= P(\eta_{ijk} = 0 \mid \alpha_{ik} = 1, q_{jk} = 1) \text{ and} \\ g_k &= P(\eta_{ijk} = 1 \mid \alpha_{ik} = 0, q_{jk} = 1). \end{aligned}$$

In the NIDA model an item response Y_{ij} is 1 if all η_{ijk}'s are equal to 1, $Y_{ij} = \prod_{k=1}^{K} \eta_{ijk}$. The item response function has the form

$$\begin{aligned} P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i, \boldsymbol{s}, \boldsymbol{g}) &= \prod_{k=1}^{K} P(\eta_{ijk} = 1 \mid \alpha_{ik}, s_k, g_k) \\ &= \prod_{k=1}^{K} \left[(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}\right]^{q_{jk}}. \end{aligned}$$

The NIDA model is somewhat restrictive and implies that item response functions must be the same for all items sharing the same attributes. It seems unrealistic that this could apply to many datasets, because it implies that item difficulty levels would be exactly the same for many items, and is not something one expects to observe in practice. A generalization of this that allows parameters to differ item-by-item is a reduced version of the Reparameterized Unified Model, called the Reduced RUM (Hartz, Roussos, Henson, and Templin 2005). In the Reduced RUM, the item response function is

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*q_{jk}(1-\alpha_{ik})}.$$

Here π_j^* denotes the probability of answering correctly for someone who possesses all of the required attributes, and r_{jk}^* can be thought of as a penalty parameter and reduces the probability of a correct response by a factor somewhere between 0 and 1 for those not possessing the k^{th} attribute.

Conjunctive models require the intersection of a set of attributes. In contrast, disjunctive models require possession of at least one of the measured attributes. Templin and Henson (2006), introduced the DINO (Deterministic Input, Noisy Output "Or" gate) model. The item response function of the DINO model is expressed as

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i) = g_j^{(1-\omega_{ij})}(1 - s_j)^{\omega_{ij}}, \tag{1}$$

where $\omega_{ij} = 1 - \prod_{k=1}^{K}(1 - \alpha_{ik})^{q_{jk}}$ and indicates whether at least one of the attributes corresponding to the item is possessed.

These are some specific models, some of which will be considered in a heuristic theory for justifying the nonparametric method of the next section, and are used in the section on simulation. All of these models and an even greater variety of models can be represented in a log-linear model framework developed by Henson, Templin, and Willse (2009).

Parametric latent class modeling for cognitive diagnosis has some disadvantages. Estimation often requires highly specialized and proprietary software. In addition, software relies on the EM algorithm or Markov Chain Monte Carlo (MCMC) for fitting the parameters of the model. MCMC can consume excessive CPU time, and convergence is often difficult to establish. The EM algorithm often converges to locally-optimal extrema. In addition to problems with estimation, there is always the concern that parametric models are simply incorrect and do not fit. Motivated by these obstacles, some researchers have proposed to avoid likelihood-based parametric models, and use nonparametric techniques for assigning examinees to attribute profiles. These methods are less restrictive and often computationally more efficient. In addition, many nonparametric classification algorithms can be easily implemented in major statistical software packages.

The rule space methodology (Tatsuoka 1983, 1990; Tatsuoka and Tatsuoka 1987, 1997) is an early contribution to diagnostic testing. In this method, Boolean descriptive functions are utilized to establish the relationship between the examinee's attribute pattern and the observed response pattern through the Q-matrix. Inspired by this method, Barnes (2010) attempted to construct the Q-matrix by applying a hill-climbing algorithm to extract the Q-matrix based on examinees' responses. However, the resulting Q-matrix consists of entries in the interval from 0 to 1, unlike the other parametric and nonparametric methods we discuss.

A more recent stream of research in nonparametric cognitive diagnosis applies cluster analysis to classify examinees. Willse, Henson, and Templin (2007), for example, apply K-means clustering to cognitive diagnosis data generated by the reduced Reparameterized Unified Model (RUM). Ayers, Nugent, and Dean (2008) test the performance of various common

clustering methods in classifying examinees. Chiu, Douglas, and Li (2009) conducted a theoretical and empirical evaluation of hierarchical agglomerative and K-means clustering for grouping examinees into clusters having similar attribute patterns. They established conditions for clusters to match perfectly with corresponding latent classes with probability approaching 1 as test length increases.

3. Methods

In this section we propose classification based on examining proximity of observed response vectors to the ideal response vectors. To be specific, a classification $\hat{\boldsymbol{\alpha}}$ is made by minimizing some measure of distance over all possible ideal response vectors, and determining the $\boldsymbol{\alpha}$ associated with the nearest ideal response vector. This nonparametric method based on ideal response patterns makes no direct use of item parameters of any cognitive diagnosis model, and can be conducted just as well with any sample size.

To formally define the procedure, let $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$, be the j^{th} component of the ideal response pattern for the i^{th} subject, and let $\boldsymbol{\eta}_i$ denote this pattern. We see that $\boldsymbol{\eta}_i$ depends only on the Q-matrix, and is a function of the unobservable $\boldsymbol{\alpha}_i$. We can construct all possible ideal response patterns, $\boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)}, ..., \boldsymbol{\eta}^{(2^K)}$, for all 2^K possible values for $\boldsymbol{\alpha}_i$. The problem of obtaining an estimator $\hat{\boldsymbol{\alpha}}_{\boldsymbol{i}}$ amounts to minimizing the distance between the observed item response vector and the ideal response pattern, $d(\boldsymbol{y}_i, \boldsymbol{\eta}^{(m)})$, for $m = 1, 2, ..., 2^K$. That is, we define $\hat{\boldsymbol{\alpha}}_{\boldsymbol{i}}$ to be the value of $\boldsymbol{\alpha}$ that results in an ideal response pattern that is as similar as possible to the observed response pattern. Of course, determining the most appropriate distance measure $d(\cdot)$ is critical to this, and it should recognize the issues of slipping and guessing, as well as other issues.

With binary data, a very natural and widely used distance measure for clustering is Hamming distance, which simply counts the number of times two vectors disagree. Hamming distance could prove useful for this application, and is given by

$$d_h(\boldsymbol{y}, \boldsymbol{\eta}) = \sum_{j=1}^{J} |y_j - \eta_j|. \tag{2}$$

However, it is likely that the responses of some items will have more variability than others. If this is the case, a weighted version of Hamming distance that places more weight on terms associated with items having smaller variance might be more efficient. Let $\bar{p}_j$ denote the proportion correct on the jth item. Then we define weighted Hamming distance by weighting according to the inverse sample variance,

$$d_{wh}(\boldsymbol{y}, \boldsymbol{\eta}) = \sum_{j=1}^{J} \frac{1}{\bar{p}_j(1-\bar{p}_j)} |y_j - \eta_j|. \tag{3}$$

In our simulations, we found that weighted Hamming distance performs slightly better than Hamming distance, and does not result in nearly as many ties. It is the primary measure used in the simulation study to follow, and we also consider weighting differently for departures from the ideal response model that would result from slips versus guesses. For example, in the case of open-ended responses in which guessing may not be possible, it makes sense to penalize a guess, which is a 1 in the response vector corresponding to a 0 in the ideal response vector, more than the opposite form of disagreement. On a multiple-choice exam, perhaps guessing should not be penalized as much as a disagreement of the form $y_j = 0$ when $\eta_j = 1$, which is sometimes referred to as a slip. For a general family of distance measures, define w_g to be the weight assigned to a guess, and w_s the weight assigned to a slip. We define a distance measure, called penalized Hamming distance, such that when $w_g = w_s = 1$ it reduces to Hamming distance, but can assign more weight to guesses when $g < s$ and more weight to slips when $g > s$,

$$d_{gs}(\boldsymbol{y}, \boldsymbol{\eta}) = \sum_{j=1}^{J} w_g I[y_j = 1]|y_j - \eta_j| + \sum_{j=1}^{J} w_s I[y_j = 0]|y_j - \eta_j|. \quad (4)$$

To make this even more general, depending possibly on varied item types within an exam, one could let weights be specific to items, and replace w_g and w_s in Equation 4 with w_{gj} and w_{sj}, respectively.

We conclude this section with a heuristic justification for why and when this method should be successful. Though cognitive diagnosis models recognize that response patterns will differ from ideal response patterns, and parameterize them accordingly, it is worth recognizing that a person's most likely response pattern may still be the ideal response pattern. Thus, we would expect item response patterns to be nearer the modal value, the ideal response pattern corresponding to the correct value of $\boldsymbol{\alpha}$, than the ideal response pattern corresponding to an incorrect value of $\boldsymbol{\alpha}$. In fact, that is what the simulations in the next section reveal. A critical observation is that a model need not be nearly deterministic for this to work, and an analysis of the particular parameter values for which the ideal response is most likely is given for the DINA and NIDA models.

Let's first consider when the ideal response pattern is the most likely response pattern under the NIDA model. The same basic result holds for the Fusion model, which is like the NIDA model but lets slip and guess parameters vary item-by-item. First define ideal response pattern $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3, ... \eta_J)'$ with jth element $\eta_j = \prod_{k=1}^{K} \alpha_k^{q_{jk}}$. Now consider an arbitrary vector of binary responses $\boldsymbol{y} = (y_1, y_2, y_3, ..., y_J)'$ and consider sufficient conditions for $P(\boldsymbol{\eta}|\boldsymbol{\alpha}) > P(\boldsymbol{y}|\boldsymbol{\alpha})$ for all $\boldsymbol{y} \neq \boldsymbol{\eta}$. Certainly, by conditional indepen-

dence, $P(\boldsymbol{\eta}|\boldsymbol{\alpha}) > P(\boldsymbol{y}|\boldsymbol{\alpha})$ if $P(\eta_j|\boldsymbol{\alpha}) \geq P(y_j|\boldsymbol{\alpha})$ for all j in 1,2,...,J, with at least 1 of the inequalities being strict. Let $A_j = \{k : q_{jk} = 1\}$, the set of indices for the required attributes for the j^{th} item.

Suppose $\eta_j = 0$ and $y_j = 1$. Because $P(0) = 1 - P(1)$, we need to see when $P(0) > .5$, or equivalently $P(1) < .5$.

$$P(1) = \prod_{k \in A_j} g_k^{(1-\alpha_k)}(1 - s_k)^{\alpha_k}.$$

The condition $\eta_j = 0$ implies that there must be at least one $k' \in A_j$ such that $\alpha_{k'} = 0$. Thus, $P(1) < P(0)$ holds when $g_{k'} < .5$, a very reasonable assumption for a valid model.

Now suppose $\eta_j = 1$ and $y_j = 0$. We need to see that $P(1) > .5$. When $\eta_j = 1$, we know that $\alpha_k = 1$ for all $k \in A_j$. Thus

$$P(1) = \prod_{k \in A_j} (1 - s_k) > .5.$$

This is the condition we need, the product of all $(1 - s_k)$s for required attributes must be larger than 0.5.

Now suppose the item response function follow a DINA model. Again, suppose $\eta_j = 0$ and $y_j = 1$. We need to see when $P(0) > .5$, or equivalently $P(1) < .5$. This is satisfied exactly when the guessing parameter g_j is less than 0.5. For $\eta_j = 1$ and $y_j = 0$, we must have $P(1) > .5$ and this holds precisely when the slipping parameter is less than 0.5. So, in the case of the DINA model, the ideal response pattern will always be the most like pattern, unless some slipping or guessing values exceed 0.5.

4. Simulation Studies

We investigated the performance of the nonparametric method under various conditions through a wide range of simulations. First, we examined the effect of weighted Hamming distances. Second, we studied to what extent guessing-slipping penalized Hamming distances can improve the classification of examinees in situations where one of the parameters in the given model is much smaller than the other, or even missing. The third simulation concerned the robustness of the nonparametric method when the given Q-matrix of a test is misspecified. In a fourth study, we compared the effect of conjunctive and disjunctive ideal response vectors on the classification of examinees; finally, we also explored the effect of a large number of skills. Classification results for the nonparametric method were evaluated by comparison with those obtained through maximum likelihood estimation (MLE) of class membership (i.e., $\boldsymbol{\alpha}$) that can be regarded as a "best case" sce-

nario because the true model underlying the data and all parameters must be known. We emphasize that this provides a rather conservative assessment of the relative efficiency of the nonparametric method, and mimics what might take place when item banks are calibrated using very large samples. For the robustness simulation (with misspecified Q-matrix), MLE does not provide the best possible classification; hence, we used the EM algorithm to obtain a standard of reference.

4.1 Simulation Design

The simulation conditions were formed by crossing test length, the number of attributes, the data generation model, and the expected departure from ideal response patterns governed by the level of random slips and guesses. For each condition, 25000 subjects were simulated using either the DINA or generalized NIDA model. The generalized NIDA model is equivalent to the Reduced RUM model, but employs a different parameterization and allows slipping and guessing parameters to vary for each item. In addition, the DINO model was adopted as the data generating model for one of the simulations to study the effectiveness of the nonparametric method when relying on disjunctive ideal response patterns. For each data set, K = 3 or 4 attributes were required and response profiles consisting of J = 20 or 40 items were sampled for N = 1000 examinees from a designated distribution (recall that we also conducted one simulation with K = 8 to evaluate the performance of the nonparametric method when K is large).

Examinees' attribute profiles were generated in two different ways. The first sampled attribute patterns, $\boldsymbol{\alpha}$, from a uniform distribution of 2^K possible values, each with probability $1/2^K$. The second approach, referred to as multivariate normal threshold model, was used to mimic a realistic situation where attributes are correlated and of unequal prevalence. The discrete $\boldsymbol{\alpha}$ were linked to an underlying multivariate normal distribution, $MVN(\mathbf{0}_K, \Sigma)$, with covariance matrix, Σ, structured as

$$\Sigma = \begin{pmatrix} 1 & & \rho \\ & \ddots & \\ \rho & & 1 \end{pmatrix},$$

and $\rho = 0.5$. Let $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \cdots, \theta_{iK})'$ denote the K-dimensional vector of latent continuous scores for examinee i. The attribute pattern $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{iK})$ was determined by

$$\alpha_{ik} = \begin{cases} 1, & \text{if } \theta_{ik} \geq \Phi^{-1}(\frac{k}{K+1}); \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Another critical factor affecting the classification results is the distribution of item parameters. The item parameters for the DINA and generalized NIDA models can both be set so that responses are nearly deterministic. In this case, using ideal response patterns will work nearly perfectly. However, from a practical point of view, it is important to investigate how far we can stretch these parameters and still obtain good nonparametric classifications. Therefore, except for the simulations taking specific distributions to generate guessing and slipping parameters for evaluating the use of the penalized weights, the guessing and slipping parameters in the other simulations were generated from uniform distributions with left endpoints of 0 and right endpoints, denoted as $Max.s$, either 0.1, 0.3, or 0.5, representing conditions of low, medium, and high perturbations.

For the simulations with penalized-weighted Hamming distance (i.e., non-equal weights for guessing and slipping parameters), as formulated in Equation 4, the data were generated with the assumption that one parameter was much less than the other or even missing. Specifically, to mimic the type of tests where guessing is close to 0 (e.g., open-ended test), the guessing parameters were set to 0 and the slipping parameters were generated from Unif(0, 0.4). Under the setup, a flip from ideal response 0 to observed response 1 represents the condition where the probability of success is contributed only by the guessing, and therefore deserves a larger penalty. In terms of Equation 4, w_g should take a value greater than w_s. w_g was thus set to 6 and w_s to 1 in the simulation. On the other hand, for conditions where guessing plays a major role to answer the items correctly, and when slipping rarely happens, the guessing parameters were generated from Unif(0, 0.4) and slipping parameters from Unif(0, 0.1).

The Q-matrices for tests of 20 items with K = 3 and 4 were designed as in Table 1, and those for tests of 40 items were obtained by doubling the length of the Q matrices in Table 1. For the simulation with misspecified Q-matrices, 10% or 20% of misspecified q entries were randomly arranged in the Q-matrix for each replication.

For $K = 8$, there are $2^8 = 256$ possible attribute combinations, and to include all of them in a single test is impractical. We therefore considered only include one-, two-, and three-skill items. Specifically, if $J = 40$, the Q-matrix consisted of all eight one-skill and 28 two-skill items; in addition, four three-skill items were randomly selected from the list of 56 possible three-skill items. Similarly, if $J = 60$, the Q-matrix included 16 one-skill items (each of the possible items repeated twice), 28 two-skill items and 16 randomly chosen three-skill items.

We also studied the effectiveness of the nonparametric method when relying on disjunctive ideal response patterns. For the simulation, data were generated from the DINO model, as described in Equation 1. The nonpara-

Table 1. Q-matrices for test of 20 items

K=3			K=4			
1	0	0	1	0	0	0
0	1	0	0	1	0	0
0	0	1	0	0	1	0
1	1	0	0	0	0	1
1	0	1	1	0	0	0
0	1	1	0	1	0	0
1	0	0	0	0	1	0
0	1	0	0	0	0	1
0	0	1	1	1	0	0
1	1	0	1	0	1	0
1	0	1	1	0	0	1
0	1	1	0	1	1	0
1	0	0	0	1	0	1
0	1	0	0	0	1	1
0	0	1	1	1	1	0
1	1	0	1	1	0	1
1	0	1	1	0	1	1
0	1	1	0	1	1	1
1	1	1	1	1	1	1
1	1	1	1	1	1	1

metric method links the observed response patterns to the disjunctive ideal response patterns $\boldsymbol{\omega}$, in which the element for the i^{th} examinee who took the j^{th} item is defined as

$$\omega_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_j k}. \tag{6}$$

Note that the expected results are independent of sample size because no item parameter was estimated and all classifications were statistically independent of the others. The large sample size of 25,000 used here is merely to eliminate any substantial Monte Carlo error.

4.2 Results

Results are summarized in terms of two indices reflecting the agreement between the obtained and the known true classification. The first we call pattern-wise agreement rate (PAR) denoting the proportion of attribute patterns accurately estimated: $PAR = \sum_{i=1}^{N} \frac{I[\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i]}{N}$. Attribute-wise agreement rate (AAR), defined as $AAR = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{I[\alpha_{ik} = \hat{\alpha}_{ik}]}{NK}$, refers to the proportion of individual attributes that were classified correctly. For

both, PAR and AAR, we also computed their "relative efficiency" as the ratio of nonparametric versus MLE indices. The tables below report the PAR, AAR, and their relative efficiencies for the nonparametric method and MLE. We would like to emphasize that any analysis of classification results should always take into consideration the level of accuracy that can be attained by chance alone (i.e., in ignoring the data) to provide a rationale frame of reference for the interpretation. For example, under the assumption of latent proficiency classes following a uniform distribution (i.e., classes are equally likely), for 8 classes (based on $K = 3$), a PAR of 0.125 can be obtained solely through random classification (or PAR = 0.0625 when $K = 4$, with 16 classes); the corresponding chance AAR equals 0.5.

For the various conditions of the simulation study, Table 2 summarizes the average CPU times observed when classifying a data set consisting of 1000 examinees by the nonparametric method: even under the most extreme condition (i.e., (K, J, $Max.s$) = (4, 40, 0.5)), the nonparametric method needs only about 1 second per data set—a very encouraging finding.

Table 3 documents the effectiveness of the nonparametric method and MLE when applied to responses generated from the DINA model. In support of the theoretical prediction, both approaches produce nearly perfect classifications when slipping and guessing parameters are less than 0.1. Not too surprising, classification performance generally deteriorates at larger noise levels. Still, we should point out that the relative efficiency of PAR for the nonparametric method remains above 0.94 as long as the parameter settings for slipping and guessing do not exceed 0.3 (notice that the relative efficiency of AAR appears to be even less susceptible to the slipping and guessing settings)—an indication of the robustness of the proposed nonparametric method (to be discussed in greater detail below). As an aside, but similarly remarkable, the relative efficiency of PAR for the nonparametric method is larger when the data were generated by the multivariate normal threshold model (that incorporates a far more realistic scenario than the uniform distribution model). A heuristic explanation of the impressive performance of the nonparametric method for DINA data can be derived from the proof given in the previous section: as long as the slipping and guessing parameter settings do not exceed 0.5, the ideal response pattern suggests the most likely choice for the proficiency class. In summary, the nonparametric technique only requiring knowledge of the Q-matrix of a given test appears to be quite competitive for DINA data in comparison with MLE that rather represents a best-case scenario in its reliance on complete knowledge of all model parameters.

The results for the generalized NIDA model are presented in Table 4. The nonparametric and MLE classifications are excellent when slips and

Table 2. Average CPU time in seconds for nonparametric classification with datasets of N=1000

	$J = 20$		$J = 40$	
	$K = 3$	$K = 4$	$K = 3$	$K = 4$
$Max.s$ = 0.1	0.5340	0.9120	0.6996	1.0596
$Max.s$ = 0.3	0.5656	0.8844	0.7192	1.0704
$Max.s$ = 0.5	0.6264	0.9404	0.7500	1.0732

Table 3. Agreement of classification between the nonparametric method and MLE with data generated from the DINA model

			Nonparametric		MLE		Relative Efficiency (Nonpar./MLE)	
J	K	$Max.s$	PAR	AAR	PAR	AAR	PAR	AAR
			Uniform Attribute Patterns					
20	3	0.1	0.9926	0.9973	0.9928	0.9975	0.9998	0.9998
		0.3	0.9261	0.9715	0.9459	0.9801	0.9791	0.9912
		0.5	0.8400	0.9349	0.9222	0.9703	0.9109	0.9635
	4	0.1	0.9612	0.9895	0.9688	0.9916	0.9922	0.9979
		0.3	0.8324	0.9494	0.8766	0.9656	0.9496	0.9832
		0.5	0.6168	0.8669	0.7108	0.9083	0.8678	0.9544
40	3	0.1	0.9990	0.9997	0.9996	0.9999	0.9994	0.9998
		0.3	0.9796	0.9928	0.9859	0.9952	0.9936	0.9976
		0.5	0.8648	0.9469	0.9331	0.9753	0.9268	0.9709
	4	0.1	0.9942	0.9985	0.9979	0.9995	0.9963	0.9990
		0.3	0.9186	0.9773	0.9564	0.9881	0.9605	0.9891
		0.5	0.7334	0.9126	0.8451	0.9542	0.8678	0.9564
			Multivariate Normal Attribute Patterns					
20	3	0.1	0.9971	0.9990	0.9976	0.9992	0.9995	0.9998
		0.3	0.9250	0.9733	0.9284	0.9744	0.9963	0.9989
		0.5	0.6984	0.8724	0.7399	0.8983	0.9439	0.9712
	4	0.1	0.9772	0.9940	0.9801	0.9947	0.9970	0.9993
		0.3	0.8082	0.9411	0.8128	0.9432	0.9943	0.9978
		0.5	0.6138	0.8672	0.7236	0.9178	0.8483	0.9449
40	3	0.1	0.9986	0.9995	0.9994	0.9998	0.9992	0.9997
		0.3	0.9684	0.9890	0.9790	0.9927	0.9892	0.9963
		0.5	0.8932	0.9612	0.9340	0.9766	0.9563	0.984
	4	0.1	0.9940	0.9985	0.9979	0.9995	0.9961	0.9990
		0.3	0.9612	0.9899	0.9610	0.9897	1.0002	1.0002
		0.5	0.6971	0.9006	0.7901	0.9359	0.8823	0.9623

guesses are bounded by 0.1. If these bounds are increased, then the PAR scores of both methods tend to decline (although, the AAR scores remain high). In comparison with the application to DINA data, the nonparametric method appears generally less tolerant of larger slipping and guessing parameter settings in NIDA data, presumably, because in case of the latter

Table 4. Agreement of classification between the nonparametric method and MLE with data generated from the generalized NIDA model

			Nonparametric		MLE		Relative Efficiency (Nonpar./MLE)	
J	K	$Max.s$	PAR	AAR	PAR	AAR	PAR	AAR
			Uniform Attribute Patterns					
20	3	0.1	0.9875	0.9957	0.9927	0.9975	0.9948	0.9982
		0.3	0.8888	0.9575	0.9424	0.9792	0.9431	0.9778
		0.5	0.6868	0.8787	0.8653	0.9495	0.7937	0.9254
	4	0.1	0.9502	0.9866	0.9670	0.9914	0.9826	0.9952
		0.3	0.6648	0.8989	0.7836	0.9384	0.8484	0.9579
		0.5	0.4366	0.8095	0.6693	0.9044	0.6523	0.8951
40	3	0.1	0.9991	0.9997	0.9998	0.9999	0.9993	0.9998
		0.3	0.9417	0.9794	0.9837	0.9943	0.9573	0.9850
		0.5	0.6890	0.8796	0.9178	0.9705	0.7507	0.9063
	4	0.1	0.9937	0.9984	0.9987	0.9997	0.9950	0.9987
		0.3	0.8506	0.9604	0.9508	0.9870	0.8946	0.9730
		0.5	0.5391	0.8580	0.8066	0.9450	0.6684	0.9079
			Multivariate Normal Attribute Patterns					
20	3	0.1	0.9925	0.9974	0.9940	0.9980	0.9985	0.9994
		0.3	0.8436	0.9444	0.9089	0.9675	0.9282	0.9761
		0.5	0.6238	0.8526	0.7618	0.9109	0.8189	0.9360
	4	0.1	0.9556	0.9886	0.9653	0.9909	0.9900	0.9977
		0.3	0.7348	0.9268	0.8373	0.9557	0.8776	0.9698
		0.5	0.4777	0.8362	0.6939	0.9114	0.6884	0.9175
40	3	0.1	0.9996	0.9999	0.9998	0.9999	0.9998	1.0000
		0.3	0.9624	0.9874	0.9934	0.9978	0.9688	0.9896
		0.5	0.8179	0.9369	0.9582	0.9854	0.8536	0.9508
	4	0.1	0.9941	0.9985	0.9969	0.9992	0.9972	0.9993
		0.3	0.8157	0.9521	0.9388	0.9840	0.8689	0.9676
		0.5	0.4592	0.8284	0.7150	0.9170	0.6422	0.9034

slipping and guessing operate at the subtask level and can have a multiplicative effect. Hence, a one-to-one comparison of the classification results obtained from the nonparametric method for DINA and NIDA data does not appear particularly viable. Finally, we notice that the performance of the nonparametric method when applied to NIDA data seems also to depend on the size of K.

The simulations reported so far all used weighted Hamming distances for the nonparametric method; slipping and guessing parameters were drawn from the same uniform distribution. We also studied the effectiveness of the nonparametric method in comparison with MLE for alternative distance measures, weighted Hamming and penalized-weighted Hamming distances, when applied to data generated based on imbalanced guessing and slipping parameter settings. Tables 5 and 6 report the classification agreements for

Table 5. Agreement of classification between the nonparametric method with weighted and penalized-weighted Hamming distances and MLE with data generated from the DINA model. g=0, s=0.4, wg=6, ws=1

								Relative Efficiency	
		Weighted		Penalized-Weighted		MLE		(P-W/MLE)	
J	K	PAR	AAR	PAR	AAR	PAR	AAR	PAR	AAR
					Uniform Attribute Patterns				
20	3	0.9624	0.9873	0.9986	0.9995	0.9986	0.9995	1.0000	1.0000
	4	0.7468	0.9304	0.9732	0.9932	0.9763	0.9940	0.9968	0.9992
40	3	0.9572	0.9857	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	4	0.9326	0.9829	0.9998	1.0000	1.0000	1.0000	0.9998	1.0000
					Multivariate Normal Attribute Patterns				
20	3	0.9222	0.9734	0.9983	0.9994	0.9984	0.9995	0.9999	0.9999
	4	0.8633	0.9635	0.9878	0.9970	0.9880	0.9970	0.9998	1.0000
40	3	0.9477	0.9821	0.9999	1.0000	0.9999	1.0000	1.0000	1.0000
	4	0.9576	0.9892	0.9998	1.0000	0.9999	1.0000	0.9999	1.0000

Table 6. Agreement of classification for the nonparametric method with weighted and penalized-weighted hamming distances and MLE with data generated from the DINA model. g=0.4, s=0.1, wg=1, ws=2

								Relative Efficiency	
		Weighted		Penalized-Weighted		MLE		(P-W/MLE)	
J	K	PAR	AAR	PAR	AAR	PAR	AAR	PAR	AAR
					Uniform Attribute Patterns				
20	3	0.8967	0.9607	0.9187	0.9690	0.9569	0.9836	0.9601	0.9852
	4	0.7871	0.9353	0.8480	0.9521	0.8906	0.9675	0.9522	0.9642
40	3	0.9703	0.9908	0.9901	0.9966	0.9972	0.9991	0.9929	0.9975
	4	0.8608	0.9612	0.9519	0.9860	0.9662	0.9907	0.9852	0.9953
					Multivariate Normal Attribute Patterns				
20	3	0.8328	0.9396	0.8707	0.9523	0.9403	0.9781	0.9260	0.9736
	4	0.7734	0.9312	0.8307	0.9516	0.8786	0.9646	0.9455	0.9865
40	3	0.9847	0.9949	0.9973	0.9991	0.9982	0.9994	0.9991	0.9997
	4	0.9288	0.9816	0.9810	0.9951	0.9836	0.9958	0.9974	0.9993

data generated from the DINA model. In Table 5, the guessing parameter was set to 0, and the slipping parameter was drawn from Unif(0, 0.4). A large number, 6, was chosen as the weight for guessing, while the weight for slipping was set to 1. We observe that the weighted Hamming distance performs fairly well; however, the use of the penalized-weighted Hamming distance dramatically improves the classification agreement rate in comparison with the regular weighted Hamming distance across all simulation conditions. The difference between the two distance measures further increases if the length of the test is reduced and a larger number of required skills is used (e.g., for (J, K) = (20, 4), PAR(weighted Hamming) = 0.7468 versus

PAR(penalized-weighted Hamming) = 0.9732). Most notable, the nonparametric method with penalized-weighted Hamming distance performs almost as well as MLE (indices of relative efficiency all greater than 0.99).

Table 6 lists the results for the other conditions, with slipping and guessing parameters drawn from Unif(0, 0.1) and Unif(0, 0.4), respectively, and weights chosen as 2 and 1. Not surprisingly, in this condition, classification performance declines for all three techniques because the data contain more noise. However, the nonparametric method with penalized-weighted Hamming distances still outperforms weighted Hamming distances across all conditions. Note that the differences in agreement of classification are substantial when K is large. In addition, the nonparametric method combined with penalized-weighted Hamming distances attains a classification rate almost as good as MLE (indices of relative efficiency for PAR exceed 0.95 in 15 out of 16 conditions).

Tables 7 and 8 report the parallel results for test data generated from the generalized NIDA model. With the guessing parameter set to 0, and the slipping parameter generated from Unif(0, 0.4), using penalized-weighted Hamming distances leads to superior classification in comparison with weighted Hamming distances across all simulation conditions. Also, all PAR relative efficiency indices (versus MLE) for penalized-weighted Hamming distances exceed 0.98, while those for weighted Hamming distances are at most equal to 0.70 when $K = 4$, or 0.90 when $K = 3$. Similar results are obtained when slipping and guessing parameters are generated from Unif(0, 0.1) and Unif(0, 0.4), respectively.

The simulation results reported here strongly suggest that the proposed nonparametric method can compete with MLE-based classification provided an appropriate distance measure is chosen (reflecting the characteristics of the test). The weights of the slipping and guessing parameters were determined based on the inspection of the relationship between the agreement of classification and the actually chosen weight. The results reported in Table 5 imply that a large weight on the guessing parameter improves classification, while Table 6 rather suggests an inverted U-shaped relationship between weight and agreement of classification (increasing up to some value ranging between 1 and 3, but declining thereafter). Clearly, further studies are warranted to generate solid evidence that will help determine the weight for maximizing the effectiveness of the nonparametric method when used with penalized-weighted Hamming distances.

The next simulation investigates the robustness of the nonparametric method when some entries in the Q-matrix are misspecified. In each replication, 10% or 20% of the entries in a given Q-matrix were randomly changed. The misspecified Q-matrix was used for classifying examinees through the nonparametric method and DINA-EM. Tables 9 and 10 report the results

Table 7. Agreement of classification between the nonparametric method with weighted and penalized-weighted Hamming distances and MLE with data generated from the NIDA model. g=0, s=0.4, wg=6, ws=1

								Relative Efficiency	
		Weighted		Penalized-Weighted		MLE		(P-W/MLE)	
J	K	PAR	AAR	PAR	AAR	PAR	AAR	PAR	AAR
		Uniform Attribute Patterns							
20	3	0.8866	0.9617	0.9976	0.9992	0.9987	0.9996	0.9989	0.9996
	4	0.6952	0.9210	0.9748	0.9937	0.9956	0.9989	0.9791	0.9948
40	3	0.8684	0.9555	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	4	0.6760	0.9135	0.9932	0.9983	0.9992	0.9998	0.9940	0.9985
		Multivariate Normal Attribute Patterns							
20	3	0.7706	0.9196	0.9949	0.9983	0.9984	0.9995	0.9965	0.9988
	4	0.6463	0.9018	0.9677	0.9918	0.9730	0.9932	0.9946	0.9986
40	3	0.9018	0.9668	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	4	0.6980	0.9188	0.9978	0.9994	0.9990	0.9998	0.9988	0.9996

Table 8. Agreement of classification for the nonparametric method with weighted and penalized-weighted hamming distances and MLE with data generated from the NIDA model. g=0.4, s=0.1, wg=1, ws=2

								Relative Efficiency	
		Weighted		Penalized-Weighted		MLE		(P-W/MLE)	
J	K	PAR	AAR	PAR	AAR	PAR	AAR	PAR	AAR
		Uniform Attribute Patterns							
20	3	0.9310	0.9752	0.9382	0.9770	0.9748	0.9913	0.9625	0.9856
	4	0.7589	0.9303	0.7847	0.9335	0.8700	0.9624	0.9020	0.9700
40	3	0.9692	0.9895	0.9882	0.9959	0.9974	0.9991	0.9908	0.9968
	4	0.9090	0.9760	0.9276	0.9806	0.9808	0.9950	0.9458	0.9855
		Multivariate Normal Attribute Patterns							
20	3	0.8123	0.9361	0.9508	0.9836	0.9993	0.9998	0.9515	0.9838
	4	0.6864	0.9136	0.8490	0.9609	0.9821	0.9955	0.8645	0.9652
40	3	0.8660	0.9550	0.9882	0.9961	1.0000	1.0000	0.9882	0.9961
	4	0.7752	0.9410	0.9337	0.9834	1.0000	1.0000	0.9337	0.9834

for DINA data with attribute patterns generated from uniform distributions and multivariate normal threshold models, respectively. Table 9 shows that both methods result in low classification agreement rates when a misspecified Q-matrix is used. The nonparametric method does not perform well as the DINA-EM when the upper bound of guessing and slipping parameters and K are large, but does outperform the DINA-EM under some cases with 20% of misspecification. The impact of K and J on classification accuracy is similar to that found in other simulations. The relative efficiency of the nonparametric method versus DINA-EM ranges from 0.78 to 1.26 when 10% of the q-entries are misspecified, and from 0.73 to 1.05 with 20% of

Table 9. Classification results for nonparametric classification and DINA-EM with DINA data and a uniform distribution on $\boldsymbol{\alpha}$ when Q is misspecified

			Nonparametric		DINA-EM		Relative Efficiency (Nonpar./DINA-EM)	
J	K	$Max.s$	PAR	AAR	PAR	AAR	PAR	AAR
			10% misspecified q entries					
20	3	0.1	0.9188	0.9717	0.9620	0.9870	0.9551	0.9845
		0.3	0.7910	0.9186	0.8576	0.9485	0.9223	0.9685
		0.5	0.6182	0.8398	0.6261	0.8477	0.9874	0.9907
	4	0.1	0.7836	0.9374	0.8868	0.9678	0.8836	0.9686
		0.3	0.6684	0.8915	0.7352	0.9165	0.9091	0.9727
		0.5	0.4566	0.8032	0.5837	0.8631	0.7823	0.9306
40	3	0.1	0.9698	0.9899	0.9973	0.9991	0.9724	0.9908
		0.3	0.9001	0.9651	0.9682	0.9888	0.9297	0.9760
		0.5	0.8149	0.9286	0.9209	0.9716	0.8849	0.9557
	4	0.1	0.8997	0.9738	0.9746	0.9935	0.9231	0.9802
		0.3	0.7999	0.9425	0.9112	0.9753	0.8779	0.9664
		0.5	0.5540	0.8497	0.4401	0.8081	1.2588	1.0515
			20% misspecified q entries					
20	3	0.1	0.7472	0.9045	0.8538	0.9476	0.8751	0.9545
		0.3	0.6715	0.8679	0.7810	0.9188	0.8598	0.9446
		0.5	0.4502	0.7472	0.4767	0.7760	0.9444	0.9629
	4	0.1	0.5321	0.8484	0.6323	0.8904	0.8415	0.9528
		0.3	0.3702	0.7468	0.3532	0.7477	1.0481	0.9988
		0.5	0.2531	0.6787	0.2502	0.6797	1.0116	0.9985
40	3	0.1	0.8124	0.9360	0.9652	0.9883	0.8417	0.9471
		0.3	0.7742	0.9157	0.8995	0.9649	0.8607	0.9490
		0.5	0.5967	0.8231	0.6602	0.8681	0.9038	0.9482
	4	0.1	0.6504	0.8967	0.8898	0.9695	0.7310	0.9249
		0.3	0.5680	0.8612	0.7290	0.9184	0.7791	0.9377
		0.5	0.4619	0.8042	0.6064	0.8736	0.7617	0.9206

misspecifications. Moreover, the degree of q-misspecification tends to deteriorate less in classification rate for the nonparametric method than for the DINA-EM. Under some conditions, relative efficiency increases along with the upper bound of the parameters, independent of the number of misspecified q-entries.

When examinees' attribute patterns were generated from the multivariate normal threshold model, the agreements of classification for the nonparametric method appear to be higher than those obtained from data with uniform attribute patterns, but are lower for the DINA-EM particularly when the upper bound of the parameters is moderate or small, as shown in Table 10. Relative efficiency of PAR (nonparametric method versus DINA-EM) ranges from 0.82 to 1.01 when 10% of the q-entries are misspecified, and from 0.78 to 1.03 when 20% are misspecified.

Table 10. Classification results for nonparametric classification and DINA-EM with DINA data and a threshold normal distribution on $\boldsymbol{\alpha}$ when Q is misspecified

			Nonparametric		DINA-EM		Relative Efficiency (Nonpar./DINA-EM)	
J	K	$Max.s$	PAR	AAR	PAR	AAR	PAR	AAR
			10% misspecified q entries					
20	3	0.1	0.9530	0.9839	0.9392	0.9795	1.0147	1.0045
		0.3	0.8354	0.9375	0.8364	0.9422	0.9988	0.9950
		0.5	0.6830	0.8744	0.7544	0.9083	0.9054	0.9627
	4	0.1	0.8679	0.9614	0.8700	0.9637	0.9976	0.9976
		0.3	0.5998	0.8612	0.6671	0.8983	0.8991	0.9587
		0.5	0.4689	0.8079	0.6017	0.8710	0.7793	0.9276
40	3	0.1	0.9752	0.9917	0.9953	0.9984	0.9798	0.9933
		0.3	0.9256	0.9736	0.9340	0.9776	0.9910	0.9959
		0.5	0.7578	0.9043	0.8196	0.9342	0.9246	0.9680
	4	0.1	0.9196	0.9789	0.9547	0.9883	0.9632	0.9905
		0.3	0.8514	0.9582	0.8957	0.9722	0.9505	0.9856
		0.5	0.6069	0.8675	0.7390	0.9225	0.8212	0.9404
			20% misspecified q entries					
20	3	0.1	0.8103	0.9334	0.8024	0.9331	1.0098	1.0003
		0.3	0.7206	0.8850	0.7173	0.8953	1.0046	0.9885
		0.5	0.5315	0.7852	0.5145	0.7943	1.0330	0.9885
	4	0.1	0.6366	0.8818	0.6648	0.8958	0.9576	0.9844
		0.3	0.4620	0.7943	0.5348	0.8399	0.8639	0.9457
		0.5	0.3487	0.7317	0.4458	0.8001	0.7822	0.9145
40	3	0.1	0.8910	0.9632	0.8820	0.9606	1.0102	1.0027
		0.3	0.8466	0.9417	0.8726	0.9568	0.9702	0.9842
		0.5	0.6820	0.8682	0.6778	0.8812	1.0062	0.9852
	4	0.1	0.7881	0.9390	0.8122	0.9495	0.9703	0.9889
		0.3	0.5796	0.8629	0.6453	0.8894	0.8982	0.9702
		0.5	0.5207	0.8332	0.6122	0.8691	0.8505	0.9587

Tables 11 and 12 summarize the findings for the data generated from the generalized NIDA model. However, as an important detail, we need to point out that we estimated examinees' proficiency class based on DINA-EM, trying to emulate a scenario where data are fitted by an inappropriate model (i.e., fitting NIDA-data by DINA-EM) to stress a different aspect of robustness of the nonparametric method; namely, its superiority in the specific case where an elaborate maximum likelihood approach is used that, however, relies on the wrong model.

As shown in the tables, the nonparametric classification method does not perform as well as the DINA-EM. The multiplicative effect of the NIDA model appears to have a severe impact on the nonparametric method in classification.

In addition to studying the effectiveness of the nonparametric method when relying on the conjunctive ideal pattern, we also investigated its per-

Table 11. Classification results for nonparametric classification and DINA-EM with NIDA data and a uniform distribution on $\boldsymbol{\alpha}$ when Q is misspecified

			Nonparametric		DINA-EM		Relative Efficiency (Nonpar./DINA-EM)	
J	K	$Max.s$	PAR	AAR	PAR	AAR	PAR	AAR
			10% misspecified q entries					
20	3	0.1	0.9202	0.9721	0.9704	0.9898	0.9483	0.9821
		0.3	0.7305	0.8943	0.8056	0.9281	0.9068	0.9636
		0.5	0.5310	0.8112	0.6740	0.8783	0.7878	0.9236
	4	0.1	0.7238	0.9203	0.8455	0.9566	0.8561	0.9621
		0.3	0.4567	0.8174	0.5216	0.8492	0.8756	0.9626
		0.5	0.3386	0.7651	0.3708	0.7955	0.9132	0.9618
40	3	0.1	0.9629	0.9876	0.9977	0.9992	0.9651	0.9884
		0.3	0.8505	0.9463	0.9463	0.9811	0.8988	0.9645
		0.5	0.5913	0.8405	0.8281	0.9391	0.7140	0.8950
	4	0.1	0.8853	0.9702	0.9805	0.9949	0.9029	0.9752
		0.3	0.6439	0.8979	0.8483	0.9580	0.7590	0.9373
		0.5	0.4142	0.8109	0.6027	0.8869	0.6872	0.9143
			20% misspecified q entries					
20	3	0.1	0.7176	0.8971	0.8729	0.9554	0.8221	0.9390
		0.3	0.6314	0.8536	0.7054	0.8915	0.8951	0.9575
		0.5	0.3475	0.7012	0.4388	0.7739	0.7919	0.9061
	4	0.1	0.4573	0.8183	0.6077	0.8793	0.7525	0.9306
		0.3	0.3588	0.7715	0.4168	0.8088	0.8608	0.9539
		0.5	0.2318	0.6879	0.2810	0.7365	0.8249	0.9340
40	3	0.1	0.7996	0.9323	0.9782	0.9926	0.8174	0.9393
		0.3	0.6290	0.8635	0.8737	0.9554	0.7199	0.9038
		0.5	0.4130	0.7506	0.5715	0.8434	0.7227	0.8900
	4	0.1	0.5950	0.8758	0.8608	0.9616	0.6912	0.9108
		0.3	0.4568	0.8217	0.7102	0.9180	0.6432	0.8951
		0.5	0.2770	0.7320	0.4254	0.8225	0.6512	0.8900

formance in association with the disjunctive ideal pattern (defined in Equation 6). Weighted Hamming distances were computed between all possible disjunctive ideal response patterns and each observed response pattern generated from the DINO model. Class membership is determined by choosing the attribute pattern resulting in the shortest distance. Table 13 records the agreements of classification obtained for DINO data when applying the conjunctive and disjunctive nonparametric methods in comparison with DINO-MLE. The results show that the conjunctive nonparametric method can hardly classify disjunctive data correctly, as reflected by the low agreement rates. The disjunctive nonparametric method, however, performs almost as well as DINO-MLE (see the high relative efficiency rates). In summary, if the correct link between examinee's attribute patterns and the given item skill patterns is specified (i.e., conjunctive or disjunctive), then

Table 12. Classification results for nonparametric classification and DINA-EM with DINA data and a threshold normal distribution on $\boldsymbol{\alpha}$ when Q is misspecified

			Nonparametric		DINA-EM		Relative Efficiency (Nonpar./DINA-EM)	
J	K	$Max.s$	PAR	AAR	PAR	AAR	PAR	AAR
			10% misspecified q entries					
20	3	0.1	0.9126	0.9690	0.9132	0.9700	0.9993	0.9990
		0.3	0.7905	0.9231	0.8412	0.9442	0.9397	0.9777
		0.5	0.6028	0.8364	0.6884	0.8864	0.8757	0.9436
	4	0.1	0.7765	0.9363	0.8392	0.9571	0.9253	0.9783
		0.3	0.6236	0.8848	0.6345	0.8983	0.9828	0.9850
		0.5	0.3417	0.7641	0.4481	0.8290	0.7626	0.9217
40	3	0.1	0.9807	0.9935	0.9964	0.9988	0.9842	0.9947
		0.3	0.8664	0.9522	0.9599	0.9863	0.9026	0.9654
		0.5	0.6138	0.8496	0.7833	0.9256	0.7836	0.9179
	4	0.1	0.9315	0.9818	0.9714	0.9928	0.9589	0.9889
		0.3	0.6436	0.8967	0.8477	0.9589	0.7592	0.9351
		0.5	0.4419	0.8248	0.5784	0.8811	0.7640	0.9361
			20% misspecified q entries					
20	3	0.1	0.7852	0.9196	0.8051	0.9335	0.9753	0.9851
		0.3	0.6413	0.8537	0.6458	0.8719	0.9930	0.9791
		0.5	0.4322	0.7484	0.5552	0.8275	0.7785	0.9044
	4	0.1	0.6194	0.8702	0.6843	0.9024	0.9052	0.9643
		0.3	0.4592	0.8143	0.4896	0.8257	0.9379	0.9862
		0.5	0.2467	0.7093	0.4397	0.8013	0.5611	0.8852
40	3	0.1	0.8948	0.9645	0.9298	0.9765	0.9624	0.9877
		0.3	0.6813	0.8812	0.7952	0.9298	0.8568	0.9477
		0.5	0.5072	0.8087	0.6904	0.8929	0.7346	0.9057
	4	0.1	0.6608	0.9019	0.8196	0.9530	0.8062	0.9464
		0.3	0.5097	0.8436	0.7334	0.9254	0.6950	0.9116
		0.5	0.2756	0.7412	0.4442	0.8353	0.6204	0.8873

the nonparametric method can be an effective alternative to a wide range of CD models for classifying examinees.

The last part of the simulation concerns the effectiveness of the nonparametric method when test items require a large number of skills. Data were generated from the DINA model, with 40 or 60 items requiring 8 skills. Table 14 reports CPU times, the agreements of classification for the nonparametric method and MLE as well as their relative efficiency. The CPU times confirm the computational efficiency of the nonparametric method even when K is large. In addition, the relative efficiency indices demonstrate that the nonparametric method performs well in comparison with MLE when the data contain moderate or small amount of noise. But for noisy data, MLE proves superior, which is not too surprising because NIDA and DINA models actually share much similarities. Finally, when $Max.s$ is large, the

Table 13. Classification rates for the conjunctive and disjunctive nonparametric methods and MLE with DINO data

			Con-Nonpar.		Dis-Nonpar.		DINO-MLE		Relative Efficiency (Dis-Nonpar./MLE)	
J	K	$Max.s$	PAR	AAR	PAR	AAR	PAR	AAR	PAR	AAR
						Uniform Attribute Patterns				
20	3	0.1	0.2419	0.6818	0.9913	0.9969	0.9924	0.9974	0.9989	0.9995
		0.3	0.2864	0.7125	0.9336	0.9760	0.9524	0.9833	0.9803	0.9926
		0.5	0.1928	0.6443	0.7357	0.8858	0.8366	0.9334	0.8794	0.9490
	4	0.1	0.1635	0.6076	0.9622	0.9896	0.9732	0.9928	0.9887	0.9968
		0.3	0.1620	0.6435	0.8087	0.9417	0.8395	0.9536	0.9633	0.9875
		0.5	0.0938	0.6093	0.4718	0.8061	0.5489	0.8346	0.8595	0.9659
40	3	0.1	0.2692	0.6615	0.9994	0.9998	0.9998	0.9999	0.9996	0.9999
		0.3	0.2716	0.6908	0.9800	0.9930	0.9907	0.9968	0.9892	0.9962
		0.5	0.3720	0.7640	0.8469	0.9419	0.9084	0.9668	0.9323	0.9742
	4	0.1	0.1076	0.5480	0.9946	0.9986	0.9976	0.9994	0.9970	0.9992
		0.3	0.1399	0.6181	0.8970	0.9702	0.9315	0.9812	0.9630	0.9888
		0.5	0.0989	0.5985	0.7267	0.9144	0.8548	0.9581	0.8501	0.9544
						Multivariate Normal Attribute Patterns				
20	3	0.1	0.4119	0.7847	0.9904	0.9966	0.9926	0.9975	0.9978	0.9991
		0.3	0.4166	0.7897	0.8943	0.9599	0.9055	0.9631	0.9876	0.9967
		0.5	0.4184	0.7810	0.7544	0.9030	0.7895	0.9182	0.9555	0.9834
	4	0.1	0.2814	0.6752	0.9563	0.9881	0.9698	0.9920	0.9861	0.9961
		0.3	0.2193	0.6409	0.8522	0.9576	0.8732	0.9639	0.9760	0.9935
		0.5	0.1914	0.6627	0.5822	0.8514	0.6368	0.8738	0.9143	0.9744
40	3	0.1	0.4346	0.7893	0.9994	0.9998	0.9996	0.9999	0.9998	0.9999
		0.3	0.3977	0.7366	0.9759	0.9918	0.9864	0.9954	0.9894	0.9964
		0.5	0.4229	0.7699	0.9150	0.9692	0.9578	0.9855	0.9553	0.9835
	4	0.1	0.2784	0.6628	0.9937	0.9984	0.9949	0.9987	0.9988	0.9997
		0.3	0.2706	0.6766	0.9215	0.9791	0.9488	0.9864	0.9712	0.9926
		0.5	0.3426	0.7521	0.7973	0.9405	0.8577	0.9595	0.9296	0.9802

Table 14. Classification for the nonparametric method and MLE with data generated from the DINA model (N=1000; K=8)

		Nonparametric			MLE		Relative Efficiency (Nonpar./MLE)	
J	$Max.s$	CPU Time	PAR	AAR	PAR	AAR	PAR	AAR
40	0.1	6.9524	0.9580	0.9930	0.9702	0.9955	0.9874	0.9975
	0.3	6.9720	0.6666	0.9364	0.7454	0.9563	0.8943	0.9792
	0.5	6.9644	0.3492	0.8394	0.4683	0.8859	0.7457	0.9475
60	0.1	8.2652	0.9826	0.9976	0.9892	0.9986	0.9933	0.9990
	0.3	8.2756	0.8503	0.9764	0.8980	0.9850	0.9469	0.9913
	0.5	8.2984	0.5278	0.8997	0.6423	0.9288	0.8217	0.9687

agreement rates for both methods dramatically drop; thus, none of the two methods should be used when the level of noise approaches 0.5.

5. Real Data Analysis

Next we analyze the well known fraction subtraction data introduced in Tatsuoka (1990). The data consist of responses to 20 items involving subtraction of fractions from 536 examinees, and were analyzed in de la Torre and Douglas (2004). Using the same Q-matrix as in de la Torre and Douglas (2004), 8 attributes are defined that amount to skills or operations needed to subtract fractions. These are: (1) Convert a whole number to a fraction, (2) Separate a whole number from fraction, (3) Simplify before subtracting, (4) Find a common denominator, (5) Borrow from whole number part, (6) Column borrow to subtract the second numerator from the first, (7) Subtract numerators, (8) Reduce answers to simplest form. Based on these definitions, the Q-matrix of attributes necessary to correctly answer each item are given in Table 15.

In this analysis, classifications are made using the nonparametric technique with weighted Hamming distance, and are also made by maximum likelihood estimation, using DINA parameters obtained by the Higher-Order DINA model published in de la Torre and Douglas (2004). We report the proportion of classifications that agree between these two techniques, both vector-wise and attribute-wise in Table 16. Note that roughly 45 percent of attribute pattern classifications were the same. Though this may not appear large at first glance, one must consider that with 8 attributes there are 256 different attribute patterns to choose from. Individual attributes were classified the same 87 percent of the time.

By averaging over the components of the classified vectors, we can tabulate the estimated proportion of mastery for each skill, as shown in Table 17. These marginal proportions appear quite similar for the two methods.

6. Conclusions

The educational importance of this research lies in its promise to make simple methods of cognitive diagnosis available without the need for calibrating parametric models. It can be performed with any sample size, and only requires a matrix that associates items with attributes. By specifying a distance measure that can be used to estimate attribute patterns by examining how far a response pattern lies from all of the ideal response patterns, one can produce rapid classifications with no information other than this matrix.

An aim of this paper was to investigate how much noise can be tolerated while still making good use of simple distance measures, rather than

Table 15. Q-Matrix for the fraction subtraction data

	Attribute							
Item	1	2	3	4	5	6	7	8
1	0	0	0	1	0	1	1	0
2	0	0	0	1	0	0	1	0
3	0	0	0	1	0	0	1	0
4	0	1	1	0	1	0	1	0
5	0	1	0	1	0	0	1	1
6	0	0	0	0	0	0	1	0
7	1	1	0	0	0	0	1	0
8	0	0	0	0	0	0	1	0
9	0	1	0	0	0	0	0	0
10	0	1	0	0	1	0	1	1
11	0	1	0	0	1	0	1	0
12	0	0	0	0	0	0	1	1
13	0	1	0	1	1	0	1	0
14	0	1	0	0	0	0	1	0
15	1	0	0	0	0	0	1	0
16	0	1	0	0	0	0	1	0
17	0	1	0	0	1	0	1	0
18	0	1	0	0	1	1	1	0
19	1	1	1	0	1	0	1	0
20	0	1	1	0	1	0	1	0

Table 16. Agreement between nonparametric classification and the Higher-Order DINA model

N	J	K	PAR	AAR
536	20	8	0.4552	0.8701

Table 17. Proportion of subjects possessing each of the attributes

Attribute	1	2	3	4	5	6	7	8
Nonparametric	0.49	0.80	0.70	0.67	0.59	0.73	0.83	0.81
H-O DINA	0.56	0.81	0.68	0.69	0.60	0.72	0.83	0.79

relying on a parametric model. The simulation study suggested that when data arise from the NIDA model or the DINA model, a moderate amount of noise can be tolerated, while maintaining high efficiency, relative to the best possible, but unrealistic case of knowing the true model and all of its parameters. Because the nonparametric technique worked well under both simulation models, as well as under the DINO in the disjunctive case, and assumes nothing about the model other than a Q-matrix, it is possible that it can actually outperform models in practice, if those models are somewhat misspecified. As such, there is a potential for the nonparametric technique

to be somewhat robust, provided the stochastic elements of the underlying model are not too dominant.

Several distance measures were suggested, and weighted Hamming distance was used in simulations and in the real data example. Distance measures the weight differentially for slipping and guessing were also studied, and can be utilized to improve performance. There is certainly room for investigating other distance measures that may produce higher classification rates. Though this method requires no specification of a probability model, which can be viewed as a strength in some ways, that also comes with inherent limitations. For example, it is not possible to calculate posterior probabilities of attribute mastery, which can be helpful as a measure of the certainty of a classification. However, the most serious limitation to the widespread use of cognitive diagnosis is in the calibration of cognitive diagnosis models, which can require large sample sizes, and methods of computation that are time-consuming. Determining when a method such as the one we have proposed can be effective is critical. The promising results suggest that there may be opportunities to implement cognitive diagnosis in small and medium sized testing programs, where it is simply not feasible to maintain reliably calibrated item banks for parametric cognitive diagnosis models.

References

AYERS, E., NUGENT, R., and DEAN, N. (2008), "Skill Set Profile Clustering Based on Student Capability Vectors Computed From Online Tutoring Data", in *Proceedings of the 1st International Conference on Education Data Mining*, eds. Baker and Beck, pp. 218–225. Montreal.

BARNES, T. (2010), "Novel Derivation and Application of Skill Matrices: The Q-Matrix Method", in *Handbook on Educational Data Mining*, Boca Raton FL: CRC Press, pp. 159–172.

CHIU, C., DOUGLAS, J., and LI, X. (2009), "Cluster Analysis for Cognitive Diagnosis: Theory and Applications", *Psychometrika, 74*, 633–665.

DE LA TORRE, J., and DOUGLAS, J. (2004), "Higher-order Latent Trait Models for Cognitive Diagnosis", *Psychometrika, 69,* 333–353.

HARTZ, S., ROUSSOS, L., HENSON, R., and TEMPLIN, J. (2005), *The Fusion Model for Skill Diagnosis: Blending Theory with Practicality*, unpublished manuscript.

HENSON, R., TEMPLIN, J., and WILLSE, J. (2009), "Defining a Family of Cognitive Diagnosis Models Using Log-Linear Models with Latent Variables", *Psychometrika, 74*, 191–210.

JUNKER, B.W., and SIJTSMA, K. (2001), "Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory", *Applied Psychological Measurement, 25*, 258–272.

MARIS, E. (1999), "Estimating Multiple Classification Latent Class Models", *Psychometrika, 64*, 187–212.

TATSUOKA, K. (1983), "Rule-Space: An Approach for Dealing with Misconceptions Based on Item Response Theory", *Journal of Educational Measurement, 20*, 34–38.

TATSUOKA, K. (1985), "A Probabilistic Model for Diagnosing Misconceptions in the Pattern Classification Approach", *Journal of Educational Statistics, 12*, 55–73.

TATSUOKA, K. (1990), "Architecture of Knowledge Structures and Cognitive Diagnosis: A Statistical Pattern Classification Approach", in *Cognitively Diagnostic Assessments*, eds. P. Nichols, S. Chipman, and R. Brennan, Hillsdale, NJ: Erlbaum, pp. 327–359.

TATSUOKA, K., and TATSUOKA, M. (1987), "Bug Distribution and Pattern Classification", *Psychometrika, 52*, 193–206.

TATSUOKA, K., and TATSUOKA, M. (1997), "Computerized Cognitive Diagnostic Adaptive Testing: Effect on Remedial Instruction as Empirical Validation", *Journal of Educational Measurement, 34*, 3–20.

TEMPLIN, J.L., and HENSON, R.A. (2006), "Measurement of Psychological Disorders Using Cognitive Diagnosis Models", *Psychological Methods, 11*, 287–305.

WILLSE, J., HENSON, R., and TEMPLIN, J. (2007), "Using Sum Scores or IRT in Place of Cognitive Diagnosis Models: Can Existing or More Familiar Models Do the Job?" paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois.