PSYCHOMETRIKA
2013
DOI: 10.1007/S11336-013-9372-Y

CONSISTENCY OF NONPARAMETRIC CLASSIFICATION IN COGNITIVE DIAGNOSIS

SHIYU WANG AND JEFF DOUGLAS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Latent class models for cognitive diagnosis have been developed to classify examinees into one of the 2^K attribute profiles arising from a K-dimensional vector of binary skill indicators. These models recognize that response patterns tend to deviate from the ideal responses that would arise if skills and items generated item responses through a purely deterministic conjunctive process. An alternative to employing these latent class models is to minimize the distance between observed item response patterns and ideal response patterns, in a nonparametric fashion that utilizes no stochastic terms for these deviations. Theorems are presented that show the consistency of this approach, when the true model is one of several common latent class models for cognitive diagnosis. Consistency of classification is independent of sample size, because no model parameters need to be estimated. Simultaneous consistency for a large group of subjects can also be shown given some conditions on how sample size and test length grow with one another.

Key words: cognitive diagnosis, nonparametric classification, large sample theory.

1. Introduction

The primary purpose of cognitive diagnosis is to accurately classify subjects according to the skills or attributes they possess. This paper provides a theoretical foundation for a nonparametric method of cognitive diagnosis studied in Chiu and Douglas (2013). In this method, classification of one subject is completely independent of classification of another, and can be conducted with a sample size as small as 1, just as accurately as with a larger sample size. This is an appealing feature, and may yield a greater variety of applications, for which calibrating a parametric model would not be practical or even possible. The theoretical results demonstrate that nonparametric classification is consistent, no matter which of several possible true models holds. We begin with a review of latent class models for cognitive diagnosis, focusing on three particular models that play a role in assumptions of consistency theorems.

1.1. Cognitive Diagnostic Models

Latent class models for cognitive diagnosis are generally restricted to reflect some assumptions about the underlying process by which examinees respond to items. We focus on a few such models that assume a conjunctive response process, and a more thorough review of cognitive diagnostic models can be found in Rupp and Templin (2007).

An important feature in the models we consider is a Q-matrix (Tatsuoka, 1985). This matrix records which attributes or skills are required to correctly respond to each item. Suppose that there are N subjects, J items, and K attributes to classify. Entry q_{jk} in the $J \times K$ matrix Q indicates whether item j requires the kth attribute. Let $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N$ be random item response vectors of N subjects, where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$. Let $\boldsymbol{\alpha}_i$ denote the attribute pattern for subject i, where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})'$ and each α_{ik} takes values of either 0 or 1 for $k = 1, 2, \ldots, K$. Specifically α_{ik} is an indicator of whether the ith subject possesses the kth attribute.

Requests for reprints should be sent to Jeff Douglas, University of Illinois at Urbana-Champaign, Champaign, IL, USA. E-mail: jeffdoug@uiuc.edu

Published online: 03 December 2013

Conjunctive latent class models for cognitive diagnosis express the notion that all attributes specified in Q for an item should be required to answer the item correctly, but allow for slips and guesses in ways that distinguish the models from one another.

The DINA model, an extension of the two-class model of Macready and Dayton (1977), and named in Junker and Sijtsma (2001), is one such example. Consider ideal response patterns, patterns that would arise if attribute possession entirely determined responses. Denote this ideal response pattern by $\boldsymbol{\eta}_i = (\eta_{i1}, \eta_{i2}, \ldots, \eta_{iJ})'$, where $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$. It denotes whether subject i has mastered all the attributes required by item j. The DINA allows for deviations from this pattern according to slipping parameters for each item, $s_j = P(Y_{ij} = 0|\eta_{ij} = 1)$, and guessing parameters for each item, $g_j = P(Y_{ij} = 1|\eta_{ij} = 0)$. The item response function of the DINA model is then

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

and a likelihood function may be constructed from these item response functions together with an assumption of independence of $\mathbf{Y}_i$ given the attribute vector $\boldsymbol{\alpha}_i$. Though it is a simple and practical model, the DINA has some strong restrictions (Roussos, Templin & Henson, 2007). In particular, it assumes that the probability of a correct item response, given non-mastery on at least one skill, does not depend on the number and type of required skills that are not mastered. The next model we consider differs in this regard, but has some restrictions of its own.

The NIDA model, introduced in Maris (1999), treats the slips and guesses at the subtask level. The ideal response patterns remain the same, but a subtask response $\eta_{ijk} = 0$ indicates whether subject i correctly applied attribute k to answer item j. In this model, $s_k = P(\eta_{ijk} = 0|\alpha_{ik} = 1, q_{jk} = 1)$, $g_k = P(\eta_{ijk} = 1|\alpha_{ik} = 0, q_{jk} = 1)$, and $Y_{ij} = 1$ only if all subtasks are correctly completed. By convention, let $P(\eta_{ijk} = 1|\alpha_{ik} = a, q_{jk} = 0) = 1$, no matter what the value of α_{ik} is. Then the item response function of the NIDA model is

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

A restriction of the NIDA model is that it implies items requiring the same set of attributes must have precisely the same item response functions. This can be viewed as a desired property in certain situations when the theory of the cognitive attributes is indeed correct, and the Q-matrix describes the cognitive processes to solve items of a certain type sufficiently well. It is also parsimonious, which is helpful for small data sets that may not afford estimation of a more general parametric model (Roussos et al., 2007). However, in many situations it can readily be seen to conflict with data, because it implies such strict conditions on observed proportion correct values of the items. For instance, the model implies that any two items with the same entry in Q must have the same expected proportion correct. A generalization of this that allows slipping and guessing probabilities to vary across items is a reduced version of the Reparameterized Unified Model (Hartz, Roussos, Henson, & Templin, 2005). In this model, the item response function is

$$P\left(Y_{ij} = 1|\boldsymbol{\alpha}_i, \pi^*, \mathbf{r}\right) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*q_{jk}(1-\alpha_k)}.$$

Here π_j^* is the probability a subject who possesses all of the required attributes answers item j correctly, and r_{jk}^* can be viewed as a penalty parameter for not possessing the kth attribute, and it is between 0 and 1.

These three specific models will be considered in the classification consistency theory of the next section. However, more general cognitive diagnostic models have been developed including

conjunctive, disjunctive, and compensatory models. For example, see the G-DINA framework (de la Torre, 2011), the log-linear cognitive diagnostic model (Henson, Templin, & Willse, 2009), and the general diagnostic model (von Davier, 2005).

1.2. Nonparametric Classification for Cognitive Diagnosis

The restricted latent class models described above have become popular in cognitive diagnostic research. However, there are alternatives that do not assume any particular probability model. The rule space methodology (Tatsuoka, 1983; Tatsuoka, 1990; Tatsuoka and Tatsuoka, 1987) is a widely known and early approach to diagnostic testing that combines parametric modeling with the notion of an ideal response pattern. The idea behind rule space is to use Boolean descriptive functions to establish the relationship between the examinee's attribute pattern and the observed response pattern through the Q-matrix, after adjusting for a fitted item response model. Building on this method, but in a more nonparametric fashion, Barnes (2010) developed hill-climbing algorithms to build the Q-matrix and examinee classifications in a purely exploratory approach. Some recent research has attempted to classify attribute patterns by utilizing cluster analysis. 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. Park and Lee (2011) also examined a method of clustering attributes required to solve mathematics problems on the TIMSS by mapping item responses to an attribute matrix, and then conducting K-means and hierarchical agglomerative cluster analysis.

A direct approach to nonparametric classification is to match observed item response patterns to the nearest ideal response pattern. Chiu and Douglas (2013), studied this method, and found that accurate classification can be achieved when the true model is DINA and NIDA with slip and guess parameters considerably greater than 0. A step of the rule space method (Tatsuoka, 1983), is quite similar. However, rule space first requires calibration of the ability parameter based on an item response model, and cannot be viewed as a wholly nonparametric method. Rule space attempts to reduce the dimensionality of observed response patterns and the ideal response scores by mapping to a pair of new variables (θ, η) in a Cartesian product space, then calculates a Mahalanobis distance between the two patterns to identify the attribute pattern for each subject. By contrast, the Chiu and Douglas method requires fewer steps. The estimator of $\boldsymbol{\alpha}$ in this method would be perfect if all slip and guess parameters were 0, but still performs with good relative efficiency even when this is not the case. To formally define the estimator, first recall that $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N$ are random item response vectors of N subjects to J items, where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$. Define $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ to be the jth component of the ideal response pattern from the ith subject, and let $\boldsymbol{\eta}$ denote this pattern. Then all possible ideal response patterns, $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_{2^K}$, can be constructed from all 2^K possible values for $\boldsymbol{\alpha}_i$. Because the $\boldsymbol{\eta}_i$ is determined by $\boldsymbol{\alpha}_i$, we define the distance between the observed item response vector for the ith subject $\mathbf{y}_i$ and the ideal response pattern under attribute profile $\boldsymbol{\alpha}_m$ to be $D(\mathbf{y}_i, \boldsymbol{\alpha}_m)$, for $m = 1, 2, \ldots, 2^K$.

The nonparametric classification estimator $\hat{\boldsymbol{\alpha}}$ arises by minimizing some measure of distance over all possible ideal response vectors, and determining the $\boldsymbol{\alpha}$ associated with the nearest ideal response vector. It is natural to use Hamming distance for clustering with binary data, which

simply counts the components of $\mathbf{y}_i$ and $\boldsymbol{\eta}_m$ that disagree,

$$D(\mathbf{y}_i, \boldsymbol{\alpha}_m) = \sum_{j=1}^{J} |y_{ij} - \eta_{mj}| = \sum_{j=1}^{J} d_j^i(\boldsymbol{\alpha}_m). \tag{1}$$

Minimizing this distance over all possible values of the latent attribute vector produces the estimator,

$$\hat{\boldsymbol{\alpha}}_i = \arg \min_{m \in \{1,2,\ldots,2^K\}} D(\mathbf{y}_i, \boldsymbol{\alpha}_m). \tag{2}$$

This estimator can result in ties, especially in short exams. The probability of a tie converges to 0 as exam length increases, so ties play no role in the theory of the estimator. However, in practice one must decide how to break ties. This can be done by randomly choosing among the tied values, or by implementing a weighted version Hamming distance to reduce their frequency. The next section considers the asymptotic theory for $\hat{\boldsymbol{\alpha}}$, and consistency theorems are given when the true model is the DINA, NIDA, or Reparameterized Unified model.

2. Consistent Classification Theory

Though the nonparametric classifier based on minimizing Hamming distance to ideal response patterns is clearly consistent when slipping and guessing parameters are 0 and data are ideal response patterns, we show it also yields consistent classification under a variety of models when the stochastic terms differ considerably from 0. Chiu and Douglas (2013) gave a heuristic justification for the theoretical underpinnings of the classifier, and a formal analysis is given below. First we provide the assumptions and conditions for consistency under different latent class models, along with their justifications.

2.1. Assumptions and Conditions

First we make the standard assumptions of independent subjects and conditional independence.

Assumption 1. The item response vectors $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N$ for subjects $1, 2, \ldots, N$ are statistically independent.

Assumption 2. For subject i, the item response $Y_{i1}, Y_{i2}, \ldots, Y_{iJ}$ are statistically independent conditional on attribute vector $\boldsymbol{\alpha}_i$.

Let q_{jk} be the j,k element of the Q-matrix, and define $B_j = \{k | q_{jk} = 1\}$.

For some number $\delta \in (0, .5)$ we have the following conditions on parameters of the possible true model:

Condition (a.1). When data arise from the DINA model, parameters g_j and s_j satisfy $g_j < 0.5 - \delta$ and $s_j < 0.5 - \delta$.

Condition (a.2). When data arise from the NIDA model, $g_k < 0.5 - \delta$, for $k = 1, 2, \ldots, K$, and $\prod_{k \in B_j}(1 - s_k) > 0.5 + \delta$, for $j = 1, 2, \ldots, J$.

Condition (a.3). When data arise from the Reduced RUM model, $\pi_j^* > 0.5 + \delta$ for every j, and for some $k \in B_j$, $r_{jk}^* < 0.5 - \delta$.

Condition (b). Define $A_{m,m'} = \{j|\eta_{mj} \neq \eta_{m'j}\}$, where m and m' index different attribute patterns among the 2^K possible patterns. Card$(A_{m,m'}) \to \infty$ as $J \to \infty$.

Condition (c). The number of subjects and the test length satisfy the relationship that $\forall \varepsilon > 0$, $Ne^{-2J\varepsilon^2} \to 0$ as $J \to \infty$.

Conditions (a.1) and (a.2) bound slipping and guessing parameters in the DINA and NIDA models away from 0.5. These are reasonable assumptions for a valid model, because the probability of a subject answering an item correctly should be at least primarily determined by possession or nonpossession of the required attributes. If such assumptions are not met, diagnostic modeling will not be as useful, either with the nonparametric classifier or with the parametric model. The condition essentially tells us that the most likely response for someone who has mastered the attributes is 1, and the most likely response for someone who has not mastered the required attributes is 0. Certainly masters of the attributes should have a higher probability of success, though requiring it is at least 0.5 does make the conditional somewhat restrictive. Condition (a.3) expresses the same notion for the Reduced RUM model, which can be rewritten as a NIDA model in which slipping and guessing parameters can vary with the item. Condition (b) guarantees that for each pair of attribute patterns, there is an infinite amount of information to separate them as the number of items grows to infinity. Finally Condition (c) is established in order to get the simultaneous consistent classification of a whole sample of subjects, and is unnecessary when considering the consistent classification of a single subject.

2.2. *Asymptotic Results*

In this section, three propositions will be introduced first in order to prove the consistency results for a single subject and a sample of subjects.

Proposition 1. *Under Assumptions* 1, 2, *Conditions* (a.1), (a.2), (a.3), *and* (b), *for every* $i \in \{1, 2, \ldots, N\}$, *the true attribute pattern will minimize* $E[D(\mathbf{Y}_i, \boldsymbol{\alpha}_m)]$ $(m = 1, 2, \ldots, 2^K)$.

Proof: Suppose the real attribute pattern for a fixed item response vector $\mathbf{Y}_i$ is $\boldsymbol{\alpha}_1$. Let $\boldsymbol{\alpha}_2$ be another attribute pattern. Because $E[D(\mathbf{Y}_i, \boldsymbol{\alpha}_m)] = \sum_{j=1}^{J} E|Y_{ij} - \eta_{mj}|$, we just need to compare $E(|Y_{ij} - \eta_{1j}|)$ and $E(|Y_{ij} - \eta_{2j}|)$ for every j. Note that if $\boldsymbol{\alpha}_1 \neq \boldsymbol{\alpha}_2$, there must be some j such that $\eta_{1j} \neq \eta_{2j}$. Let $A_{1,2} = \{j|\eta_{1j} \neq \eta_{2j}\}$. Then for every $j \in A_{1,2}$, we have

$$\eta_{1j} = 1, \qquad \eta_{2j} = 0, \qquad E\big(|Y_{ij} - \eta_{1j}|\big) - E\big(|Y_{ij} - \eta_{2j}|\big) = 1 - 2P(Y_{ij} = 1),$$
$$\eta_{1j} = 0, \qquad \eta_{2j} = 1, \qquad E\big(|Y_{ij} - \eta_{1j}|\big) - E\big(|Y_{ij} - \eta_{2j}| = 2P(Y_{ij} = 1) - 1\big).$$

The problem then turns out to be deriving the specific $P(Y_{ij} = 1)$ under different models.

(1) When data arise from DINA model, $P(Y_{ij} = 1) = (1 - s_j)^{\eta_j} g_j^{1-\eta_j}$.
When $\eta_{1j} = 1$ and $\eta_{2j} = 0$,

$$E\big(|Y_{ij} - \eta_{1j}|\big) - E\big(|Y_{ij} - \eta_{2j}|\big) = 2s_j - 1.$$

When $\eta_{1j} = 0$ and $\eta_{2j} = 1$,

$$E\big(|Y_{ij} - \eta_{1j}|\big) - E\big(|Y_{ij} - \eta_{2j}| = 2g_j - 1\big).$$

From Condition (a.1) we know that $g_j < 0.5 - \delta, s_j < 0.5 - \delta$ for some positive number δ. So we can get $E(|Y_{ij} - \eta_{1j}|) < E(|Y_{ij} - \eta_{2j}|)$ under the DINA model.

(2) When data arise from the NIDA model, $P(Y_{ij} = 1) = \prod_{k=1}^{K} [(1 - s_k)^{\alpha_k} g_k^{1-\alpha_k}]^{q_{jk}}$.

When $\eta_{1j}=1$ and $\eta_{2j}=0$,

$$E\big(|Y_{ij}-\eta_{1j}|\big)-E\big(|Y_{ij}-\eta_{2j}|\big)=1-2\prod_{k\in B_j}(1-s_k).$$

When $\eta_{1j}=0$ and $\eta_{2j}=1$,

$$E\big(|Y_{ij}-\eta_{1j}|\big)-E\big(|Y_{ij}-\eta_{2j}|\big)=2\prod_{k\in B_j}g_k^{1-\alpha_k}(1-s_k)^{\alpha_k}-1.$$

According to Condition (a.2), for some positive number δ, $\prod_{k\in B_j}(1-s_k)>0.5+\delta$, so $1-2\prod_{k\in B_j}(1-s_k)<0$. Furthermore, when $\eta_{1j}=0$, there must be some $k'\in B_j$ such that $\alpha_{k'}=0$. Then $g_{k'}<0.5-\delta<0.5$, we can get $2\prod_{k\in B_j}g_k^{1-\alpha_k}(1-s_k)^{\alpha_k}-1<0$. So $E(|y_{ij}-\eta_{1j}|)<E(|y_{ij}-\eta_{2j}|)$ is also correct under the NIDA model.

(3) When data arise from the Reduced RUM model, $P(Y_{ij}=1)=\pi_j^*\prod_{k=1}^K r_{jk}^{*q_{jk}(1-\alpha_k)}$. When $\eta_{1j}=1$ and $\eta_{2j}=0$,

$$E\big(|Y_{ij}-\eta_{1j}|\big)-E\big(|Y_{ij}-\eta_{2j}|\big)=1-2\pi_j^*.$$

When $\eta_{1j}=0$ and $\eta_{2j}=1$,

$$E(|Y_{ij}-\eta_{1j}|)-E(|Y_{ij}-\eta_{2j}|)=2\pi_j^*\prod_{k\in B_j}r_{jk}^{*(1-\alpha_k)}-1.$$

Similar to the argument for the NIDA model, there must be some $k'\in B_j$ such that $\alpha_{k'}=0$. With Condition (a.3) that $\pi_j^*>0.5+\delta$ for each j, we can get $1-2\pi_j^*<0$. And $r_{jk'}<0.5-\delta$, then $\pi_j^*\prod_{k\in B_j}r_{jk}^{*(1-\alpha_k)}=r_{jk'}\pi_j^*\prod_{k\in B_j,k\neq k'}r_{jk}^*<0.5$.

From the above argument,we can see that no matter which of the models is true, $E(|Y_{ij}-\eta_{1j}|)<E(|Y_{ij}-\eta_{2j}|)$, when $j\in A_{1,2}$. Otherwise, for every $j\in A_{1,2}^C$, $E(|Y_{ij}-\eta_{1j}|)=E(|Y_{ij}-\eta_{2j}|)$. Then we can see that

$$E\left[\sum_{j=1}^J|Y_{ij}-\eta_{1j}|\right]<E\left[\sum_{j=1}^J|Y_{ij}-\eta_{2j}|\right].$$

□

The next proposition states that the true attribute pattern will be well separated from one another as test length goes to infinity.

Proposition 2. *Under the assumptions and conditions of Proposition 1, in addition to Condition (b), and suppose* $\boldsymbol{\alpha}_1$ *is the true attribute profile,* $\boldsymbol{\alpha}_2$ *is another different attribute profile,*

$$\lim_{J\to\infty}E\big[D(\mathbf{Y}_i,\boldsymbol{\alpha}_2)\big]-E\big[D(\mathbf{Y}_i,\boldsymbol{\alpha}_1)\big]=\infty.$$

Proof: Note that the difference between $E[D(\mathbf{Y}_i,\boldsymbol{\alpha}_1)]$ and $E[D(\mathbf{Y}_i,\boldsymbol{\alpha}_2)]$ is only determined by the values when $j\in A_{1,2}$. We first prove this proposition under the DINA model:

$$\begin{aligned}\lim_{J\to\infty}E\big[D(\mathbf{Y}_i,\boldsymbol{\alpha}_2)\big]-E\big[D(\mathbf{Y}_i,\boldsymbol{\alpha}_1)\big]&=\lim_{J\to\infty}\sum_{j\in J_1}1-2g_j+\lim_{J\to\infty}\sum_{j\in J_2}1-2s_j\\&>\lim_{J\to\infty}\sum_{j\in J_1}\delta+\lim_{J\to\infty}\sum_{j\in J_2}\delta=\infty.\end{aligned}$$

Here $J_1 = \{j \in A_{1,2} | \eta_{1j} = 0, \eta_{2j} = 1\}$, $J_2 = \{j \in A_{1,2} | \eta_{1j} = 1, \eta_{2j=0}\}$, and the infinite sum results from the cardinality of J_1 and J_2 going to infinity, as guaranteed by Condition (b).

The same argument can be applied to the NIDA model and the Reduced RUM. □

The third proposition investigates the relationship between the average of $d_j^i(\boldsymbol{\alpha}_m)$ and its expectation $E[d_j^i(\boldsymbol{\alpha}_m)]$, for fixed i and every $m \in \{1, 2, \ldots, 2^K\}$, when the test length goes to infinity.

Proposition 3. *Under Assumptions 1 and 2, $\forall \varepsilon > 0$ and a fixed $i \in \{1, 2, \ldots, N\}$, define $B_\varepsilon(J) = \{\max_{m \in \{1,2,\ldots,2^K\}} |\frac{1}{J} \sum_{j=1}^{J} (d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])| \geq \varepsilon\}$. Then*

$$\lim_{J \to \infty} P\big(B_\varepsilon(J)\big) = 0.$$

In order to prove Proposition 3, we need to apply Hoeffding's Inequality as it is stated below:

Hoeffding's inequality. *Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables such that $\forall i$, $0 \leq Z_i \leq 1$, Then*

$$P\left(\left\{\left|\frac{1}{n}\left(\sum_{i=1}^{n} Z_i - E[Z_i]\right)\right| \geq \varepsilon\right\}\right) \leq 2\exp\big(-2n\varepsilon^2\big).$$

Proof of Proposition 3: First we must show that for every $\varepsilon > 0$, $P(\{|\frac{1}{J} \sum_{j=1}^{J} (d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])| \geq \varepsilon\}) \leq 2\exp(-2J\varepsilon^2)$, and this is obtained by Hoeffding's Inequality.

Note that $\forall j \in 1, 2, \ldots, J$, $d_j^i(\boldsymbol{\alpha}_m) = |Y_{ij} - \eta_{mj}|$. For subject i, Y_{ij}, $j \in \{1, 2, \ldots, J\}$ are independent random variables, conditional on the true attribute pattern. This implies that $d_j^i(\boldsymbol{\alpha}_m)$, $j \in \{1, 2, \ldots, J\}$ are independent random variables, and $0 \leq d_j^i(\boldsymbol{\alpha}_m) \leq 1$ is obvious. So $d_1^i(\boldsymbol{\alpha}_m), d_2^i(\boldsymbol{\alpha}_m), \ldots, d_J^i(\boldsymbol{\alpha}_m)$ are independent random variables which satisfy the conditions of Hoeffding's inequality, which states that for every $\varepsilon > 0$,

$$P\left(\left\{\left|\frac{1}{J}\sum_{j=1}^{J}\big(d_j^i(\boldsymbol{\alpha}_m) - E\big[d_j^i(\boldsymbol{\alpha}_m)\big]\big)\right| \geq \varepsilon\right\}\right) \leq 2\exp\big(-2J\varepsilon^2\big).$$

Using this result we see that

$$\begin{aligned}
&P\left(\bigcup_{m=1}^{2^K}\left\{\left|\frac{1}{J}\sum_{j=1}^{J}\big(d_j^i(\boldsymbol{\alpha}_m) - E\big[d_j^i(\boldsymbol{\alpha}_m)\big]\big)\right| \geq \varepsilon\right\}\right) \\
&\leq \sum_{m=1}^{2^K} P\left(\left\{\left|\frac{1}{J}\sum_{j=1}^{J}\big(d_j^i(\boldsymbol{\alpha}_m) - E\big[d_j^i(\boldsymbol{\alpha}_m)\big]\big)\right| \geq \varepsilon\right\}\right) \\
&\leq 2^{K+1}\exp\big(-2J\varepsilon^2\big)
\end{aligned}$$

which implies that

$$
\begin{aligned}
&P\left(\left\{\max_{m\in\{1,2,\ldots,2^K\}}\left|\frac{1}{J}\sum_{j=1}^{J}\bigl(d_j^i(\boldsymbol{\alpha}_m)-E\bigl[d_j^i(\boldsymbol{\alpha}_m)\bigr]\bigr)\right|\geq\varepsilon\right\}\right)\\
&\quad=1-P\left(\left\{\max_{m\in\{1,2,\ldots,2^K\}}\left|\frac{1}{J}\sum_{j=1}^{J}\bigl(d_j^i(\boldsymbol{\alpha}_m)-E\bigl[d_j^i(\boldsymbol{\alpha}_m)\bigr]\bigr)\right|<\varepsilon\right\}\right)\\
&\quad=1-P\left(\bigcap_{m=1}^{2^K}\left\{\left|\frac{1}{J}\sum_{j=1}^{J}\bigl(d_j^i(\boldsymbol{\alpha}_m)-E\bigl[d_j^i(\boldsymbol{\alpha}_m)\bigr]\bigr)\right|<\varepsilon\right\}\right)\\
&\quad=P\left(\bigcup_{m=1}^{2^K}\left\{\left|\frac{1}{J}\sum_{j=1}^{J}\bigl(d_j^i(\boldsymbol{\alpha}_m)-E\bigl[d_j^i(\boldsymbol{\alpha}_m)\bigr]\bigr)\right|\geq\varepsilon\right\}\right)\\
&\quad\leq 2^{K+1}\exp\bigl(-2J\varepsilon^2\bigr).
\end{aligned}
$$

Note that 2^{K+1} and ε are constant, so this probability converges to 0 when J goes to infinity. □

The preceding propositions are now used to prove Theorem 1, along with a corollary to show that for a single subject, the estimate of $\hat{\boldsymbol{\alpha}}$ by the nonparametric method will converge to the true attribute vector almost surely when test length goes to infinity.

Theorem 1. *For a particular subject i with true attribute pattern $\boldsymbol{\alpha}_i$, under Assumptions 1 and 2 and Conditions (a.1), (a.2), (a.3), and (b), the estimator $\hat{\boldsymbol{\alpha}}_i$ derived from the nonparametric method of Equation (2) is a consistent estimator of $\boldsymbol{\alpha}_i$, provided one of the DINA model, NIDA model or Reduced RUM holds. Specifically,*

$$\lim_{J\to\infty}P(\hat{\boldsymbol{\alpha}}_i=\boldsymbol{\alpha}_i)=1.$$

Proof: For fixed subject i, $\forall\varepsilon>0$, let event $A_\varepsilon(J)=\{|\hat{\boldsymbol{\alpha}}_i-\boldsymbol{\alpha}_i|>\varepsilon\}$, and $B_\varepsilon(J)$ as defined in Proposition 3. Then we show that $A_\varepsilon(J)\subset B_\varepsilon(J)$. In order to prove this, we can prove $B_\varepsilon(J)^C\subset A_\varepsilon(J)^C$.

Suppose $B_\varepsilon(J)^C$ is true that is $\forall m\in\{1,2,\ldots,2^K\}$, $\forall\varepsilon>0$

$$\left|\frac{1}{J}\sum_{j=1}^{J}\bigl(d_j^i(\boldsymbol{\alpha}_m)-E\bigl[d_j^i(\boldsymbol{\alpha}_m)\bigr]\bigr)\right|<\varepsilon.$$

Then we find

$$\frac{1}{J}\sum_{j=1}^{J}E\bigl[d_j^i(\boldsymbol{\alpha}_m)\bigr]-\varepsilon<\frac{1}{J}\sum_{j=1}^{J}d_j^i(\boldsymbol{\alpha}_m)<\frac{1}{J}\sum_{j=1}^{J}E\bigl[d_j^i(\boldsymbol{\alpha}_m)\bigr]+\varepsilon.$$

If $\hat{\boldsymbol{\alpha}}_i\neq\boldsymbol{\alpha}_i$, then $\sum_{j=1}^{J}d_j^i(\hat{\boldsymbol{\alpha}}_i)<\sum_{j=1}^{J}d_j^i(\boldsymbol{\alpha}_i)$ and $\frac{1}{J}\sum_{j=1}^{J}d_j^i(\hat{\boldsymbol{\alpha}}_i)<\frac{1}{J}\sum_{j=1}^{J}d_j^i(\boldsymbol{\alpha}_i)$. These inequalities imply that

$$\frac{1}{J}\sum_{j=1}^{J}E\bigl[d_j^i(\hat{\boldsymbol{\alpha}}_i)\bigr]-\varepsilon<\frac{1}{J}\sum_{j=1}^{J}d_j^i(\hat{\boldsymbol{\alpha}}_i)<\frac{1}{J}\sum_{j=1}^{J}d_j^i(\boldsymbol{\alpha}_i)<\frac{1}{J}E\bigl[d_j^i(\boldsymbol{\alpha}_i)\bigr]+\varepsilon,\quad\forall\varepsilon>0.$$

Thus for small enough ε we have

$$\sum_{j=1}^{J} E\big[d_j^i(\hat{\boldsymbol{\alpha}}_i)\big] < \sum_{j=1}^{J} E\big[d_j^i(\boldsymbol{\alpha}_i)\big].$$

This is contradictory to Proposition 1 that shows the true attribute pattern will minimize $E[D(\mathbf{Y}_i, \boldsymbol{\alpha}_m)]$, and Proposition 2 that when $J \to \infty$, the difference of the expectation of the distance defined by (1) under the wrong attribute pattern with that of under the true attribute pattern will go to infinity. So we may conclude that $B_\varepsilon(J)^C \subset A_\varepsilon(J)^C$, and equivalently $A_\varepsilon(J) \subset B_\varepsilon(J)$.

By the above claim and Proposition 3, we have $\forall \varepsilon > 0$

$$P\big(|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i| > \varepsilon\big) \le P\big(B_\varepsilon(J)\big) \le 2^{K+1} \exp\big(-2J\varepsilon^2\big) \to 0, \quad \text{as } J \to \infty.$$

Thus we have proved that if Assumptions 1 and 2 and Conditions (a.1), (a.2), (a.3), and (b) are satisfied, $\lim_{J\to\infty} P(\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i) = 1$. □

Corollary 1. *Under Assumptions 1 and 2 and Conditions (a.1), (a.2), (a.3), and (b),*

$$\lim_{J\to\infty} P\big(|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i| > \varepsilon, \quad \text{i.o.}\big) = 0.$$

Proof: We only need to prove that

$$\sum_{J=1}^{\infty} P\big(B_\varepsilon(J)\big) < \infty$$

and the result will follow from the Borel–Cantelli Theorem. Note that

$$\sum_{J=1}^{\infty} P\big(B_\varepsilon(J)\big) < \sum_{J=1}^{\infty} 2^{K+1} \exp\big(-2J\varepsilon^2\big) = 2^{K+1} \sum_{J=1}^{\infty} \exp\big(-2J\varepsilon^2\big).$$

Define $f(J) = \exp(-2J\varepsilon^2)$.

According to the convergence rule of series,

$$\lim_{J\to\infty} \frac{f(J+1)}{f(J)} = \lim_{J\to\infty} \frac{\exp(-2(J+1)\varepsilon^2)}{\exp(-2J\varepsilon^2)} = \exp\big(-2\varepsilon^2\big) < 1$$

Then we have

$$\sum_{J=1}^{\infty} P\big(A_\varepsilon(J)\big) < \sum_{J=1}^{\infty} P\big(B_\varepsilon(J)\big) < \infty$$

which completes the proof of the corollary. □

Finally we investigate the joint consistency of a sample of N subjects. Essentially the same results hold, but there must be some control on the relative sizes of N and J as both go to infinity.

Theorem 2. *Under Assumptions* 1 *and* 2 *and Condition* (a.1), (a.2), (a.3), (b), *and* (c) *in Section* 2.1.1, *provided one of the DINA model, NIDA model or Reduced RUM holds,*

$$\lim_{J\to\infty} P\left(\bigcap_{i=1}^{N}\{\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i\}\right) = 1.$$

Proof: Note that the only difference between Theorem 1 and Theorem 2 is that Theorem 2 has sample size N but Theorem 1 has only one subject. So Proposition 1 and Proposition 2 still hold for every subject in Theorem 2. For Proposition 3, $\forall\varepsilon > 0$, we can define $B_\varepsilon(N, J) = \{\max_{i\in\{1,2,\ldots,N\}}\{\max_{m\in\{1,2,\ldots,2^K\}} |\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - Ed_j^i(\boldsymbol{\alpha}_m))| \geq \varepsilon\}\}$. Then under Condition (c) we can show that

$$\lim_{J\to\infty} P\big(B_\varepsilon(N, J)\big) = 0.$$

For every $i \in \{1, 2, \ldots, N\}$, $P(\{|\frac{1}{J}\sum_{j=1}^{J}(d_j^i(\boldsymbol{\alpha}_m) - E[d_j^i(\boldsymbol{\alpha}_m)])| \geq \varepsilon\}) \leq 2\exp(-2J\varepsilon^2)$ holds, so we see that

$$\begin{aligned}
&P\left(\bigcup_{i=1}^{N}\bigcup_{m=1}^{2^K}\left\{\left|\frac{1}{J}\sum_{j=1}^{J}\big(d_j^J(\boldsymbol{\alpha}_m) - E\big[d_j^i(\boldsymbol{\alpha}_m)\big]\big)\right| \geq \varepsilon\right\}\right)\\
&\leq \sum_{i=1}^{N}\sum_{m=1}^{2^K} P\left(\left\{\left|\frac{1}{J}\sum_{j=1}^{J}\big(d_j^i(\boldsymbol{\alpha}_m) - E\big[d_j^i(\boldsymbol{\alpha}_m)\big]\big)\right| \geq \varepsilon\right\}\right)\\
&\leq 2^{K+1}N\exp\big(-2J\varepsilon^2\big).\\
&\quad\Longrightarrow\\
&P\left(\left\{\max_{i\in\{1,2,\ldots,N\}}\left\{\max_{m\in\{1,2,\ldots,2^K\}}\left|\frac{1}{J}\sum_{j=1}^{J}\big(d_j^i(\boldsymbol{\alpha}_m) - E\big[d_j^i(\boldsymbol{\alpha}_m)\big]\big)\right| \geq \varepsilon\right\}\right\}\right)\\
&= 1 - P\left(\left\{\max_{i\in\{1,2,\ldots,N\}}\left\{\max_{m\in\{1,2,\ldots,2^K\}}\left|\frac{1}{J}\sum_{j=1}^{J}\big(d_j^i(\boldsymbol{\alpha}_m) - E\big[d_j^i(\boldsymbol{\alpha}_m)\big]\big)\right| < \varepsilon\right\}\right\}\right)\\
&= 1 - P\left(\bigcap_{i=1}^{N}\bigcap_{m=1}^{2^K}\left\{\left|\frac{1}{J}\sum_{j=1}^{J}\big(d_j^i(\boldsymbol{\alpha}_m) - E\big[d_j^i(\boldsymbol{\alpha}_m)\big]\big)\right| < \varepsilon\right\}\right)\\
&= P\left(\bigcup_{i=1}^{N}\bigcup_{m=1}^{2^K}\left\{\left|\frac{1}{J}\sum_{j=1}^{J}\big(d_j^i(\boldsymbol{\alpha}_m) - E\big[d_j^i(\boldsymbol{\alpha}_m)\big]\big)\right| \geq \varepsilon\right\}\right)\\
&\leq 2^{K+1}N\exp\big(-2J\varepsilon^2\big).
\end{aligned}$$

Note that 2^{K+1} is constant, so this probability converges to 0 provided the sample size and test length have the relationship $Ne^{-2J\varepsilon^2} \to 0$ as $J \to \infty$ (Condition (c)).

Now define $A_\varepsilon(N, J) = \{\max_{i\in\{1,2,\ldots,N\}} |\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i| > \varepsilon\}$, with the same argument as that in the proof of Theorem 1, we can prove that $A_\varepsilon(N, J) \subset B_\varepsilon(N, J)$. Then we find

$$\begin{aligned}
P\left(\bigcup_{i=1}^{N}\big\{|\hat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i| > \varepsilon\big\}\right) &\leq P\left\{\max_{i\in\{1,2,\ldots,N\}}\left\{\max_{m\in\{1,2,\ldots,2^K\}}\left|\frac{1}{J}\sum_{j=1}^{J}\big(d_j^i(\boldsymbol{\alpha}_m) - E\big[d_j^i(\boldsymbol{\alpha}_m)\big]\big)\right| \geq \varepsilon\right\}\right\}\\
&\leq 2^{K+1}N\exp\big(-2J\varepsilon^2\big) \to 0,
\end{aligned}$$

provided that $N\exp(-2J\varepsilon^2) \to 0$ as $J \to \infty$. This completes the proof of Theorem 2. □

Corollary 2. *Under Assumptions 1 and 2 and Conditions (a.1), (a.2), (a.3), (b), and (c), if the test length J and sample size N satisfy the relationship*

$$J^{n_1} \le N < J^{n_2}, \quad 1 < n_1 < n_2 < \infty.$$

Then $\lim_{J\to\infty} P(\bigcup_{i=1}^{N}\{\hat{\boldsymbol{\alpha}}_i \neq \boldsymbol{\alpha}_i\},\ \text{i.o.}) = 0.$

Proof: We only need to prove that

$$\sum_{N=1}^{J^{n_2}} \sum_{J=1}^{\infty} P\big(B_\varepsilon(N, J)\big) < \infty,$$

and the result will follow from the Borel–Cantelli Theorem like in Corollary 1. Note that

$$\begin{aligned}
\sum_{N=1}^{J^{n_2}} \sum_{J=1}^{\infty} P\big(B_\varepsilon(N, J)\big) &< \sum_{N=1}^{J^{n_2}} \sum_{J=1}^{\infty} 2^K N \exp\big(-2J\varepsilon^2\big) \\
&= 2^K \sum_{J=1}^{\infty} \left(\sum_{N=1}^{J^{n_2}} N \right) \exp\big(-2J\varepsilon^2\big) \\
&= 2^K \sum_{J=1}^{\infty} \frac{1 + J^{n_2}}{2} \exp\big(-2J\varepsilon^2\big).
\end{aligned}$$

Define $f(J) = \frac{1+J^{n_2}}{2}\exp(-2J\varepsilon^2)$. According to the convergence rule of series,

$$\begin{aligned}
\lim_{J\to\infty} \frac{f(J+1)}{f(J)} &= \lim_{J\to\infty} \frac{\frac{1+(1+J)^{n_2}}{2}\exp(-2(J+1))\varepsilon^2}{\frac{1+J^{n_2}}{2}\exp(-2J\varepsilon^2)} \\
&= \lim_{J\to\infty} \frac{1 + (J+1)^{n_2}}{1 + J^{n_2}} \exp\big(-2\varepsilon^2\big) \\
&= \exp\big(-2\varepsilon^2\big) < 1 \\
&\Rightarrow \sum_{J=1}^{\infty} \frac{1 + J^{n_2}}{2} \exp\big(-2J\varepsilon^2\big) < \infty.
\end{aligned}$$

Then we have

$$\sum_{N=1}^{J^{n_2}} \sum_{J=1}^{\infty} P\big(A_\varepsilon(N, J)\big) < \sum_{N=1}^{J^{n_2}} \sum_{J=1}^{\infty} P\big(B_\varepsilon(N, J)\big) < \infty$$

and the proof is complete. □

These theorems for consistency of the nonparametric method, assume that one of several possible true models hold, and we have focused on some of the common latent class models for cognitive diagnosis. The purpose was to show that the simple nonparametric method may be used without calibrating a model, no matter which of those models hold. However, essentially the same general condition was used in the proof of each particular model, and here we focus on

the most general condition that any cognitive diagnosis model must satisfy for the nonparametric technique to yield consistent classification. The key steps for the proofs of consistency require that Propositions 1, 2 and 3 hold, under proper regularity conditions. If we replace the conditions for the model parameters (Conditions (a.1), (a.2), and (a.3)) with the more general condition that involves no model parameters,

Condition (a′).

$$P(Y_j = 1|\eta_j = 1) > 0.5 + \delta, \quad P(Y_j = 1|\eta_j = 0) < 0.5 - \delta,$$
$$\text{for some positive number } \delta > 0,$$

the three propositions still hold. This means that the consistency results (Theorem 1, Corollary 1, Theorem 2, and Corollary 2) still hold for any models which satisfy the Condition (a′). The theory of the previous results essentially utilized this condition, but phrased it in terms of what it required of model parameters. More generally, these conditions on model parameters can be replaced by Condition (a′).

3. Numerical Studies

3.1. Study Design

In this section, we report simulated examples to illustrate finite test length behavior. The simulation conditions are similar to those in Chiu and Douglas (2013) and were formed by crossing test length, the data generation model, and the expected departure from ideal response patterns. For each condition, 1000 subjects were simulated using either DINA or NIDA model. For each data set, $K = 3$ attributes were required and response profiles consisting of $J = 20$ or 40 items were generated. For a much more thorough simulation study that covers more conditions and compares with competing parametric approaches, see Chiu and Douglas (2013).

Two methods were used to generate the attribute profiles. The first sampled attribute patterns, $\boldsymbol{\alpha}$, from a uniform distribution on the 2^K possible values. The second approach utilized a multivariate normal threshold method. Discrete $\boldsymbol{\alpha}$ were linked to an underlying multivariate normal distribution, MVN($\mathbf{0}_K$, Σ), with covariance matrix, Σ, structured as

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

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

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

The item parameters for the DINA and NIDA models were generated from uniform distributions with left endpoints of 0 and right endpoints, denoted as $\max(s, g)$, either 0.1, 0.3 or 0.5.

The Q-matrices for tests of 20 items with $K = 3$ were designed as in Table 1, and those for tests of 40 items were obtained by doubling the length of the Q-matrix 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.

TABLE 1.
Q-matrices for test of 20 items.

$K = 3$		
1	0	0
0	1	0
0	0	1
1	1	0
1	0	1
0	1	1
1	0	0
0	1	0
0	0	1
1	1	0
1	0	1
0	1	1
1	0	0
0	1	0
0	0	1
1	1	0
1	0	1
0	1	1
1	1	1
1	1	1

TABLE 2.
Classification rates for the nonparametric method with DINA data.

$\max(s, g)$	$J = 20$	$J = 40$
	Uniform attribute patterns, $K = 3$	
0.1	0.9925	1.000
0.3	0.9200	0.9825
0.5	0.8100	0.8850
	Multivariate normal attribute patterns, $K = 3$	
0.1	0.9900	0.9975
0.3	0.9500	0.9950
0.5	0.7050	0.8675

3.2. *Results*

Results are summarized by an index called pattern-wise agreement rate (PAR), denoting the proportion of attribute patterns accurately estimated according to $\text{PAR} = \sum_{i=1}^{N} \frac{I|\hat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i|}{N}$. The nonparametric estimator based on minimizing Hamming distance can result in some ties, and these ties were randomly broken, though there might be room for developing a more sophisticated technique.

In the first example, we investigate the impact of the model parameters and test length on the PAR of the nonparametric method based on Hamming distance, when actual data were generated from the DINA model and the NIDA model.

Table 2 documents the effectiveness of the nonparametric method when applied to responses generated from the DINA model. In support of the theoretical results, this approach produces nearly perfect classifications when the slipping and guessing parameter are less than 0.1. As the item parameters become close to 0.5, the classification become worse, but much better than ran-

TABLE 3.
Classification rates for the nonparametric method with NIDA data.

$\max(s, g)$	$J = 20$	$J = 40$
	Uniform attribute patterns, $K = 3$	
0.1	0.9950	1.000
0.3	0.8225	0.8725
0.5	0.6775	0.8023
	Multivariate normal attribute patterns, $K = 3$	
0.1	0.9925	0.9990
0.3	0.8900	0.9125
0.5	0.4650	0.4800

TABLE 4.
Classification results for the nonparametric methods with DINA data and a uniform distribution on $\boldsymbol{\alpha}$ when Q is misspecified.

$\max(s, g)$	$J = 20$	$J = 40$
	10 % misspecified q entries, $K = 3$	
0.1	0.9125	0.9650
0.3	0.7825	0.8775
0.5	0.4625	0.8021
	20 % misspecified q entries, $K = 3$	
0.1	0.7231	0.8024
0.3	0.6621	0.7642
0.5	0.4321	0.5861

dom assignment, which would have an expected classification rate of 0.125 when $K = 3$. Consistent with the asymptotic theory, classification rates clearly improve as test length increases. Table 3 presents the results when the responses were generated from the NIDA model. Note that the conditions for consistency are somewhat different for the NIDA (Condition (a.2)), and could be violated for several items, in the case where s and g parameters are allowed to be as large as 0.5.

Next we consider the robustness of the nonparametric method when several entries of 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 with the nonparametric method. Table 4 reports the results for the DINA data with attribute patterns generated from a uniform distribution. (The results when the attribute patterns generated from multivariate normal distribution have similar patterns thus omit here.) Table 4 shows that classification agreement decreases with the rate of misspecification. However, as the test length increases, the correct classification rate still increases. In all cases, classification rates are well above random assignment. Table 5 shows the similar results for NIDA data. Though it requires theoretical proof, we speculate that, for a certain range of the misspecified percent of entries in Q-matrix, the consistency theories may still hold.

4. Discussion

The consistency results of the previous section demonstrate that nonparametric classification can be effective under a variety of underlying conjunctive models. This can greatly expand potential applications, by allowing for conducting cognitive diagnosis when calibration of a parametric

TABLE 5.
Classification results for the nonparametric method with NIDA data and a uniform distribution on $\boldsymbol{\alpha}$ when Q is misspecified.

$\max(s, g)$	$J = 20$	$J = 40$
	10 % misspecified q entries, $K = 3$	
0.1	0.9125	0.9550
0.3	0.7123	0.8275
0.5	0.4232	0.4532
	20 % misspecified q entries, $K = 3$	
0.1	0.7135	0.7824
0.3	0.6213	0.6120
0.5	0.3346	0.4032

model is not feasible. Nonparametric classification based on minimizing the Hamming distance to ideal response patterns is simple and fast and can be used with a large number of attributes. One advantage over parametric modeling is that no model calibration is needed, and it can be performed with a sample size as small as 1. Requiring no large samples or calibration allows for small scale implementation, such as in the classroom setting, where diagnosis can be most important.

The appealing property of the nonparametric method, is that it is consistent under a variety of possible true parametric models, and can be viewed as robust in that sense. Here we studied the properties of the classifier under the DINA, NIDA, and RED-RUM models, but consistency is not be restricted to them, as a conjunctive response process and knowledge of the Q-matrix are the critical assumptions. As discussed following the theoretical results, the only general condition required of the underlying item response functions, is that the probability of a correct response for masters of the attributes is bounded above 0.5 for each item, and the probability for non-masters is bounded below 0.5. If the true model satisfies these simple conditions, nonparametric classification will be consistent as the test length increases.

Though consistency results were demonstrated, Chiu and Douglas (2013) show that maximum likelihood estimation with the correct parametric model is more efficient, which can be expected. Other advantages of using parametric statistical models is that one can use general statistical techniques for goodness-of-fit, model selection, and gain a sense of variability and the chance of errors. For instance, classification using parametric latent class models for cognitive diagnosis allows one to compute posterior probabilities for any attribute pattern, which cannot be done with the nonparametric classifier.

Nevertheless, the impressive relative efficiency (Chiu & Douglas, 2013) of the nonparametric classifier and its consistency properties suggest that the approach may be a useful alternative when calibration of a parametric model is not feasible. This approach can be implemented as soon as an item bank with a corresponding Q-matrix has been developed, and the computational simplicity allows one to construct reliable computer programs for classification that amount to exhaustively searching through all possible patterns, which is guaranteed to identify an optimal solution. Some promising directions for future research in nonparametric classification include development of fit indices and algorithms for computerized adaptive testing. Another ongoing issue for future research is identifying or validating the correct specification of the Q-matrix. In the numerical study we see the effect of misspecification rate on performance. Incorrect Q-matrix entries affect both parametric and nonparametric techniques, and suggest that robust methods could be a fruitful area of research.

References

Ayers, E., Nugent, R., & Dean, N. (2008). Skill set profile clustering based on student capability vectors computed from online tutoring data. In: R.S.J.d. Baker, T. Barnes, & J.E. Beck (Eds.), *Proceedings of the 1st international conference on education data mining*, Montreal (pp. 218–225).

Barnes, T. (2010). Novel derivation and application of skill matrices: the q-matrix method. In *Handbook on educational data mining* (pp. 159–172). Boca Raton: CRC Press.

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

Chiu, C., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnostic theory and applications. *Psychometrika*, *74*, 633–665.

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

Hartz, S., Roussos, L., Henson, R., & Templin, J. (2005). *The Fusion Model for skill diagnosis: blending theory with practicality*. Unpublished manuscript.

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

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

Macready, G.B., & Dayton, C.M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *2*, 99–120.

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

Park, Y., & Lee, Y. (2011). *IERI monograph series: issues and methodologies in large-scale assessments: Vol. 4. Diagnostic cluster analysis of mathematics skills*.

Roussos, L., Templin, J., & Henson, R. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement*, *44*(4), 293–311.

Rupp, A.A., & Templin, J.L. (2007). Unique characteristics of cognitive diagnosis models. In: *The annual meeting of the National Council for Measurement in Education*, Chicago, April 2007.

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: P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessments* (pp. 327–359). Hillsdale: Erlbaum.

Tatsuoka, K., & Tatsuoka, M. (1987). Bug distribution and pattern classification. *Psychometrika*, *52*, 193–206.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report RR-05-16). Princeton: Educational Testing Service.

Willse, J., Henson, R., & 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.

Manuscript Received: 12 JUN 2012