British Journal of Mathematical and Statistical Psychology (2015), 68, 268–291

The British Psychological Society

www.wileyonlinelibrary.com

Heuristic cognitive diagnosis when the Q-matrix is unknown

Hans-Friedrich Köhn[1]*, Chia-Yi Chiu[2] and Michael J. Brusco[3]

[1]Department of Psychology, University of Illinois at Urbana-Champaign, Illinois, USA
[2]Department of Educational Psychology, Rutgers, the State University of New Jersey, New Brunswick, New Jersey, USA
[3]College of Business, Florida State University, Tallahassee, Florida, USA

Cognitive diagnosis models of educational test performance rely on a binary Q-matrix that specifies the associations between individual test items and the cognitive attributes (skills) required to answer those items correctly. Current methods for fitting cognitive diagnosis models to educational test data and assigning examinees to proficiency classes are based on parametric estimation methods such as expectation maximization (EM) and Markov chain Monte Carlo (MCMC) that frequently encounter difficulties in practical applications. In response to these difficulties, non-parametric classification techniques (cluster analysis) have been proposed as heuristic alternatives to parametric procedures. These non-parametric classification techniques first aggregate each examinee's test item scores into a profile of attribute sum scores, which then serve as the basis for clustering examinees into proficiency classes. Like the parametric procedures, the non-parametric classification techniques require that the Q-matrix underlying a given test be known. Unfortunately, in practice, the Q-matrix for most tests is not known and must be estimated to specify the associations between items and attributes, risking a misspecified Q-matrix that may then result in the incorrect classification of examinees. This paper demonstrates that clustering examinees into proficiency classes based on their item scores rather than on their attribute sum-score profiles does not require knowledge of the Q-matrix, and results in a more accurate classification of examinees.

1. Introduction

Cognitive diagnosis models of educational test performance decompose an examinee's overall ability into a set of specific discrete skills, called attributes, each of which he or she may or may not have mastered, thereby providing a detailed description, or attribute profile, of his or her strengths and weaknesses in the ability domain of the test. The entire set of possible attribute profiles for a given test defines classes of intellectual proficiency to which examinees can be assigned.

Current methods of fitting cognitive diagnosis models to educational test data use expectation maximization (EM) or Markov chain Monte Carlo (MCMC) (de la Torre, 2009, 2011; DiBello, Roussos, & Stout, 2007; von Davier, 2008) to obtain maximum likelihood estimates of the model parameters that are then used to assign examinees to proficiency classes. (For brevity, these parametric estimation procedures are collectively referred to as *MLE procedures*.) MLE procedures require that the Q-matrix underlying a given test be

*Correspondence should be addressed to Hans-Friedrich Köhn, Department of Psychology, University of Illinois at Urbana-Champaign, 603 E. Daniel St., Champaign, IL 61820, USA (email: hkoehn@cyrus.psych.uiuc.edu).

DOI:10.1111/bmsp.12044

known. The binary Q-matrix specifies the associations between each item in a test and the one or more attributes required (0 = not required, 1 = required) to respond correctly to that item (Tatsuoka, 1985). Unfortunately, the Q-matrix for most tests is not known and must be estimated to establish the associations between the items and the attributes, which might result in a misspecified Q-matrix and in the assignment of examinees to proficiency classes to which they do not belong. In addition, MLE procedures often encounter technical difficulties in practice (e.g., the possible misspecification of the cognitive diagnosis model that is supposed to underlie the data; the requirement of large samples of examinees that may not be available in small- or medium-sized testing programmes; sensitivity to the starting values; no guaranteed globally optimal solutions; CPU time issues; the need for high-quality software that is often proprietary).

In response to these technical difficulties, a number of researchers (Ayers, Nugent, & Dean, 2008; Chiu & Douglas, 2013; Chiu, Douglas, & Li, 2009; Park & Lee, 2011; Willse, Henson, & Templin, 2007) have proposed non-parametric classification techniques as heuristic or approximate alternatives to MLE procedures for assigning examinees to proficiency classes. (A heuristic uses clever computational shortcut strategies to obtain a solution that is very close, if not identical, to the optimal solution.) The heuristic classification procedures, as implemented to date, have used the Q-matrix and each examinee's set or profile of binary test item scores (0 = incorrect, 1 = correct) to compute for him or her a profile of aggregated attribute sum scores. These attribute sum-score profiles then serve as input to a clustering method that assigns examinees to clusters, which serve as proxies for the proficiency classes in cognitive diagnosis.

Like the MLE procedures, the heuristic classification techniques require that the Q-matrix underlying a given test be known and so risk the misclassification of examinees. This paper demonstrates that clustering examinees into proficiency classes based on their item-score profiles rather than on their attribute sum-score profiles does not require knowledge of the Q-matrix, and results in a more accurate classification of examinees. First, a brief review of cognitive diagnosis models and classification techniques adapted to cognitive diagnosis is provided in order to present relevant definitions and technical key concepts needed for the remainder of the paper. This review is followed by a detailed theoretical and empirical appraisal of the advantages and limitations of both attribute sum scores and item scores as input to classification techniques for cognitive diagnosis. The results of three simulation studies and a practical application to a subset of Tatsuoka's (1984) well-known fraction-subtraction data are then presented. Finally, the discussion addresses several questions regarding the theoretical and empirical implications raised by the findings and provides some guidelines for educational practitioners.

2. Technical background

To set the stage for the theoretical and empirical appraisal of attribute sum scores and item scores as input to heuristic classification techniques, necessary definitions and technical key concepts for cognitive diagnosis models and classification techniques as adapted for cognitive diagnosis models are reviewed next.

2.1. Cognitive diagnosis models

Models for cognitive diagnosis are constrained latent class models that are equivalent to a special form of finite mixture models. Let Y_{ij} denote the observed response of

examinee i, $i = 1, \ldots, N$, to binary item j, $j = 1, \ldots, J$. Consider N examinees who belong to K distinct latent classes of intellectual proficiency. The general latent class model defines the conditional probability of examinee i in proficiency class $\mathcal{C}_k$, $k = 1, \ldots, K$, answering correctly item j by the item response function, $P(Y_{ij} = 1|i \in \mathcal{C}_k) = \pi_{jk}$, where π_{jk} is constant for item j across all members i in proficiency class $\mathcal{C}_k$. For J items, the item response function is characterized by $J \times K$ parameters, π_{jk}. The proficiency-class membership of the examinees is estimated from the observed item responses, Y_{ij}, using MLE; the observed item responses are assumed independent conditional on proficiency-class membership (i.e., local independence). No further restrictions are imposed on the relation between the latent variable – proficiency-class membership – and the observed item response. In contrast, cognitive diagnosis models constrain the relation between the latent variable and the observed item response so that the mastery of cognitive attributes characteristic for distinct latent proficiency classes determines the observed response (correct or incorrect) to an item. (For reviews of cognitive diagnosis models, see DiBello *et al.*, 2007; Fu & Li, 2007; Haberman & von Davier, 2007; Henson, Templin, & Willse, 2009; Rupp, 2007; and Rupp, Templin, & Henson, 2010).

Suppose that A latent binary attributes constitute a certain ability domain; there are then 2^A distinct attribute profiles composed of these A attributes representing K distinct latent proficiency classes. (Note that an attribute profile for a proficiency class can consist of all zeros, because it is possible for an examinee not to have mastered any attributes at all.) Let the A-dimensional vector $\boldsymbol{\alpha}_k = (\alpha_1, \ldots, \alpha_A)'$ represent the binary attribute profile of proficiency class $\mathcal{C}_k$, where the ath entry indicates whether the respective attribute has been mastered. (For brevity, the attribute profile of examinee $i \in \mathcal{C}_k$, $\boldsymbol{\alpha}_{i \in \mathcal{C}_k}$, will often be written as $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iA})'$.) Consider a test of J items for assessing ability in that domain. Each individual item j is associated with a binary attribute profile that specifies or constrains the particular skills required for answering it correctly. Item-attribute profiles that consist entirely of zeros, however, are inadmissible, because they correspond to items whose answers require no skills at all; hence, given A attributes, there are at most $2^A - 1$ distinct item-attribute profiles. The entire set of constraints specifying the associations between J items and A attributes constitutes the Q-matrix, $\mathbf{Q} = \{q_{ja}\}_{(J \times A)}$, $a = 1, \ldots, A$, where $q_{ja} = 1$ if a correct answer to the jth item requires mastery of the ath attribute, and 0 otherwise (Tatsuoka, 1985); thus, the rows of $\mathbf{Q}$ represent the item-attribute profiles, $\mathbf{q}_j$.

Perhaps the most popular of the many available cognitive diagnosis models is the conjunctive non-compensatory Deterministic Input Noisy Output 'AND' Gate (DINA) model (Junker & Sijtsma, 2001; Macready & Dayton, 1977). (In a conjunctive non-compensatory model, an examinee cannot make up for a lack of mastery of a specific attribute or attributes by mastery of another attribute or attributes.) The item response function of the DINA model for item j and examinee i is

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

where $s_j = P(Y_{ij} = 0|\eta_{ij} = 1)$ and $g_j = P(Y_{ij} = 1|\eta_{ij} = 0)$ are item parameters formalizing the probabilities of 'slipping' (failing to answer item j correctly despite having the skills required to do so) and 'guessing' (answering item j correctly despite lacking the skills required to do so), respectively. The conjunction parameter η_{ij} indicates whether examinee i has mastered all the attributes needed to answer correctly item j and is defined as

$$\eta_{ij} = \prod_{a=1}^{A} \alpha_{ia}^{q_{ja}}. \quad (2)$$

Thus, η_{ij} represents the ideal response when neither slipping nor guessing occurs. (The entire vector of J ideal responses of examinee i is written as $\boldsymbol{\eta}_i$.) Estimation of the item parameters, s_j and g_j, and assignment of examinees to proficiency classes is typically accomplished using EM or MCMC.

2.2. Non-parametric classification adapted to cognitive diagnosis

For input to classification techniques, Ayers *et al.* (2008), Willse *et al.* (2007), and Chiu *et al.* (2009) aggregated each examinee's item scores, $\boldsymbol{Y}_i$, into an A-dimensional profile of attribute sum scores, $\boldsymbol{W}_i$, defined as $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iA})' = \boldsymbol{Y}_i\mathbf{Q}$, where $W_{ia} = \sum_{j=1}^{J} Y_{ij}q_{ja}$. Because each cell entry of $\mathbf{Q} = \{q_{ja}\}$ represents the association between an item and an attribute, each element of $\boldsymbol{W}_i$ consists of the sum of the correct answers of examinee i to all items requiring mastery of the ath attribute. (Items that require mastery of more than one attribute for their solution contribute to multiple elements of $\boldsymbol{W}_i$.) Across examinees, the attribute sum-score profiles, $\boldsymbol{W}_i$, form the rows of a rectangular $N \times A$ matrix, $\mathbf{W}$. (Score matrices are symbolized by non-italicized bold capital letters, for example, $\mathbf{W}$, but their rows by subscripted italicized bold capital letters, for example, $\boldsymbol{W}_i$, to emphasize that these are vectors of random variables. For brevity, the examinee index is omitted when the context permits).

Many techniques exist for non-parametric classification of a set of objects (such as the rows of a matrix). The principal objective shared by all of these technique is to identify maximally homogeneous groups ('clusters') that are maximally separated. To adapt one popular technique, hierarchical agglomerative cluster analysis (HACA), to cognitive diagnosis requires transforming the $N \times A$ matrix of examinees' attribute sum-score profiles into an $N \times N$ symmetric matrix of (squared) inter-examinee Euclidean distances. Popular HACA algorithms include single-link, complete-link, and average-link clustering (Johnson, 1967) and Ward's (1963) minimum-variance method. HACA algorithms all sequentially merge or agglomerate examinees (or groups of examinees) closest to each other at each step into an inverted tree-shaped hierarchy of nested classes that represents the relationship between examinees. The inter-examinee distances are updated after each merger to reflect the latest status of examinee/cluster cohesion as input for the next agglomeration step; the specific manner of recalculating these distances distinguishes the link algorithms. Ward's method uses a different strategy that does not rely upon inter-examinee distances but instead attempts to minimize the increase in total within-cluster variance after merging.

K-means clustering is undoubtedly the most popular classification technique for identifying an exhaustive disjoint (i.e., non-hierarchical) grouping of a data set (Bock, 2007; Forgy, 1965; Hartigan & Wong, 1979; MacQueen, 1967; Steinley, 2006). The number of clusters, K, to be extracted must be specified in advance. The grouping process attempts to minimize the loss function of within-cluster heterogeneity (which is equivalent to maximizing between-cluster heterogeneity). A collection of K (here, $K = 2^A$) mutually exclusive and exhaustive subsets of the entire set of N examinees, $\mathcal{C}_1, \ldots, \mathcal{C}_K$, is sought, so that the overall sum of squared within-cluster deviations

of examinees from the A-vector of their cluster centroids, $\mathrm{WCSS}(\mathbf{W}) = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} ||\boldsymbol{W}_i - \overline{\boldsymbol{W}}_k||^2$, is minimized. $\overline{\boldsymbol{W}}_k$ denotes the centroid (mean) vector of cluster $\mathcal{C}_k$ obtained by averaging the observed attribute sum-score profiles, $\boldsymbol{W}_{i \in \mathcal{C}_k}$, where the A elements, $\overline{W}_{ak}$, are defined as $\overline{W}_{ak} = \frac{1}{N_k} \sum_{i \in \mathcal{C}_k} W_{ia}$, with N_k indicating the number of examinees in $\mathcal{C}_k$. The typical K-means clustering algorithm starts by selecting an initial set of examinees as cluster centres ('seeds'). The distances of the remaining examinees to these seeds determine the initial value of the loss function. The algorithm follows an iterative improvement strategy by repeatedly relocating examinees to clusters according to minimum distance; cluster centroids are recalculated and examinees regrouped until no further decreases in the loss function can be realized (i.e., until each examinee is located closest to the centroid of the cluster to which he or she is assigned).

3. Theoretical and empirical appraisal of non-parametric classification: A comparison of attribute sum scores and item scores as input

Ideally, a non-parametric heuristic classification technique should be capable of assigning examinees to proficiency classes at a level of accuracy comparable to that of the MLE procedures. But a demonstration that a specific classification technique is a legitimate heuristic method for a particular cognitive diagnosis model should not depend solely on a comparison of its empirical results with those of the MLE procedures, but should also establish the theoretical legitimacy of that technique *vis-à-vis* the statistical properties of the cognitive diagnosis model in question. The asymptotic classification theory of cognitive diagnosis (ACTCD; Chiu *et al.*, 2009) provided a theoretical foundation for using HACA as a heuristic for assigning examinees to proficiency classes for educational data conforming to the DINA model (Junker & Sijtsma, 2001; Macready & Dayton, 1977). This theory consists of three lemmas, each of which specifies a condition necessary for a consistency theorem to hold; this theorem states that the probability that HACA assigns examinees correctly to their true proficiency classes approaches 1 as the length of a test (i.e., the number of test items) increases, provided that each examinee's item scores, $\boldsymbol{Y}$, have been aggregated into an attribute sum-score profile, $\boldsymbol{W}$ (consistency theorem of classification of the ACTCD; Chiu *et al.*, 2009, pp. 645–647). These lemmas are briefly described next; for additional details, including formal proofs, consult Chiu *et al.* (2009).

3.1. Theoretical advantages of attribute sum scores

Like MLE procedures, heuristic classification techniques using attribute sum-score profiles as input require that the Q-matrix be complete. A Q-matrix is said to be complete if it allows identification of all possible attribute profiles. Lemma 1 of the ACTCD states that a Q-matrix is complete if and only if each attribute is represented by at least one single-attribute item (Chiu *et al.*, 2009, p. 643). Let $\boldsymbol{T}(\boldsymbol{\alpha}) = E(\boldsymbol{W}\,|\boldsymbol{\alpha})$ be the conditional expectation of the attribute sum-score profile, $\boldsymbol{W}$, given attribute profile $\boldsymbol{\alpha}$, where the ath element of $\boldsymbol{T}(\boldsymbol{\alpha})$ is defined as

$$T_a(\boldsymbol{\alpha}) = E(W_a|\boldsymbol{\alpha}) = \sum_{j=1}^{J} E(Y_j|\boldsymbol{\alpha}) q_{ja} \tag{3}$$

and

$$E(Y_j|\boldsymbol{\alpha}, s_j, g_j) = (1 - s_j)^{\eta_j} g_j^{(1-\eta_j)} \tag{4}$$

is the expected response for the DINA model. Loosely speaking, $\boldsymbol{T}(\boldsymbol{\alpha})$ can be regarded as the centre of the proficiency class characterized by $\boldsymbol{\alpha}$. Consider two attribute profiles, $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$. Lemma 2 of the ACTCD states that, if the Q-matrix is complete, then $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^* \Rightarrow \boldsymbol{T}(\boldsymbol{\alpha}) \neq \boldsymbol{T}(\boldsymbol{\alpha}^*)$ always holds (Chiu *et al.*, 2009, pp. 643–644). Lemma 2 justifies using $\boldsymbol{W}$ as a statistic for $\boldsymbol{\alpha}$ because the centres of the different proficiency classes are guaranteed to be distinct. That is, the proficiency classes are well separated, which is a requirement for proving the consistency of $\boldsymbol{W}$. Thus far, Lemma 2 has been proved only for item responses conforming to the DINA model. If a finite mixture model with K latent classes underlies the item responses, then Lemma 3 of the ACTCD establishes that complete-link HACA accurately assigns examinees to their true proficiency classes provided that the resulting classification hierarchy is 'cut' at K clusters (Chiu *et al.*, 2009, pp. 644–645).

3.2. Theoretical advantages and limitations of item scores

The ACTCD does not address the relative advantages of assigning examinees to proficiency classes based on their item-score profiles, $\boldsymbol{Y}$, or their attribute sum-score profiles, $\boldsymbol{W}$, but focuses on only $\boldsymbol{W}$ because of its direct conceptual relation to the underlying constrained latent class model. Hence, the question arises whether $\boldsymbol{Y}$ is also consistent; that is, does the consistency theorem of classification also hold for $\boldsymbol{Y}$? As a condition necessary for consistency, $\boldsymbol{Y}$ must satisfy Lemma 2 of the ACTCD. Suppose that $\boldsymbol{Y}$ conforms to the DINA model, with the item response function given in equation (1). Assume that $0 < g_j, s_j < 0.5$ for all j. Let $\boldsymbol{S}(\boldsymbol{\alpha}) = E(\boldsymbol{Y}|\boldsymbol{\alpha})$ be the conditional expectation of an item-score profile, given attribute profile $\boldsymbol{\alpha}$, with the jth element defined as $S_j(\boldsymbol{\alpha}) = E(Y_j|\boldsymbol{\alpha})$ (see equation (4)). Assume that the Q-matrix is complete. Let j be a single-attribute item that requires mastery of the ath attribute; then $\mathbf{q}_j = \mathbf{e}_a$, where $\mathbf{e}_a$ is a $1 \times A$ vector, with the ath element, e_a, equal to 1, and all other entries equal to 0.

Proposition. $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^* \Rightarrow \boldsymbol{S}(\boldsymbol{\alpha}) \neq \boldsymbol{S}(\boldsymbol{\alpha}^*)$.

Proof. Because $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$, the two attribute profiles must differ in at least one entry, a $\in \{1, \ldots, \mathrm{A}\}$, say, $\alpha_\mathrm{a} = 1$ and $\alpha_a^* = 0$. Also, because $\mathbf{Q}$ is complete, there exists some j $\in \{1, \ldots, \mathrm{J}\}$ such that $\mathbf{q}_j = \mathbf{e}_a$. Then $\eta_j = \prod_{a=1}^{A} \alpha_a^{q_{ja}} = \prod_{a=1}^{A} \alpha_a^{e_a} = 1$ and $\eta_j^* = \prod_{a=1}^{A} (\alpha_a^*)^{e_a} = 0$. Consequently, $S_j(\boldsymbol{\alpha}) = 1 - s_j$ and $S_j(\boldsymbol{\alpha}^*) = g_j$. But because $g_j < 1 - s_j$ for all j, $S_j(\boldsymbol{\alpha}) \neq S_j(\boldsymbol{\alpha}^*)$, implying that $\mathbf{S}(\boldsymbol{\alpha}) \neq \mathbf{S}(\boldsymbol{\alpha}^*)$. Therefore, if $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$, then $\mathbf{S}(\boldsymbol{\alpha}) \neq \mathbf{S}(\boldsymbol{\alpha}^*)$. □

So it can be proved that $\boldsymbol{Y}$ is covered by Lemma 2. That is, $\boldsymbol{Y}$ is a legitimate statistic for $\boldsymbol{\alpha}$, because the centres of the different proficiency classes are guaranteed to be distinct. But can the consistency theorem of classification also be proved for $\boldsymbol{Y}$? Unfortunately, the answer is 'no' (at least at present), because the dimensionality of $\boldsymbol{Y}$ depends on J: if J goes to infinity, then $\boldsymbol{Y}$ contradicts the fundamental assumption of any classification algorithm that the input to that algorithm be finite. This difficulty is elegantly avoided by $\boldsymbol{W}$, because its dimensionality depends on A (rather than J).

3.3. Empirical limitations of attribute sum scores

As indicated earlier, heuristic classification techniques that use attribute sum-score profiles, $\boldsymbol{W}$, as input require that the Q-matrix underlying a given test be known. In practice, however, the Q-matrix for most tests is not known, so the (fallible) judgement of educational experts is used to specify the associations between items and attributes, risking a misspecified Q-matrix that may then result in the incorrect classification of examinees.

Another difficulty is that aggregating each examinee's item scores, $\boldsymbol{Y}$, into an attribute sum-score profile, $\boldsymbol{W}$, can result in the representation of distinct $\boldsymbol{Y}$, possibly related to different proficiency classes, by identical $\boldsymbol{W}$, which may then lead to the misclassification of examinees. This difficulty is not due to the aggregation *per se*, but to the contamination of observed item scores with stochastic error. As demonstrated next, aggregating error-free item scores into an attribute sum-score profile never eliminates the distinction between proficiency classes, whereas aggregating error-contaminated item scores (such as those obtained in real testing situations) may do so.

3.3.1. Error-free item scores

In the case of error-free item scores, the vector of ideal item responses, $\boldsymbol{\eta}$, represents an item-score profile, and the vector of ideal attribute sum scores, defined as $\boldsymbol{W}(\boldsymbol{\eta}) = \boldsymbol{\eta}\mathbf{Q}$, represents an attribute sum-score profile. It can be proved that any proficiency class is uniquely determined not only by its attribute profile, $\boldsymbol{\alpha}$, but also by its corresponding $\boldsymbol{\eta}$ (shared by all examinees in that proficiency class) and associated $\boldsymbol{W}(\boldsymbol{\eta})$; and that aggregating $\boldsymbol{\eta}$ into $\boldsymbol{W}(\boldsymbol{\eta})$ completely preserves proficiency-class membership. Assume that the Q-matrix is complete.

Proposition. *For error-free item scores, the true proficiency-class membership of an examinee can always be identified directly from* $\boldsymbol{\eta}$ *or* $\boldsymbol{W}(\boldsymbol{\eta})$ *(say, by inspection): (i)* $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^* \Leftrightarrow \boldsymbol{\eta} \neq \boldsymbol{\eta}^*$ *and (ii)* $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^* \Leftrightarrow \boldsymbol{W}(\boldsymbol{\eta}) \neq \boldsymbol{W}(\boldsymbol{\eta}^*)$.

Proof. (i) ($\Rightarrow$) Because $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$, the two attribute profiles must differ in at least one entry, $a \in \{1, \ldots, A\}$, say, $\alpha_a = 1$ and $\alpha_a^* = 0$. Because the Q-matrix is complete, there exists some $j \in \{1, \ldots, J\}$ such that $\mathbf{q}_j = \mathbf{e}_a$. As shown previously, under this condition $\eta_j = 1$ and $\eta_j^* = 0$. Because $\eta_j \neq \eta_j^*$, $\boldsymbol{\eta} \neq \boldsymbol{\eta}^*$, regardless of whether the remaining entries in the two vectors are identical. Therefore, if $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$, then $\boldsymbol{\eta} \neq \boldsymbol{\eta}^*$.

(i) ($\Leftarrow$) For convenience, the equivalent statement $\boldsymbol{\alpha} = \boldsymbol{\alpha}^* \Rightarrow \boldsymbol{\eta} = \boldsymbol{\eta}^*$ will be proved. If $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$, then $\alpha_a = \alpha_a^*$ for all a. So $\eta_j = \prod_{a=1}^{A} \alpha_a^{q_{ja}} = \prod_{a=1}^{A} (\alpha_a^*)^{q_{ja}} = \eta_j^*$ for all j; thus, $\boldsymbol{\eta} = \boldsymbol{\eta}^*$. Therefore, $\boldsymbol{\eta} \neq \boldsymbol{\eta}^* \Rightarrow \boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$.

(ii) The ath entry of $\boldsymbol{W}(\boldsymbol{\eta})$ is

$$W_a(\boldsymbol{\eta}) = \sum_{j=1}^{J} \left(\prod_{a=1}^{A} \alpha_a^{q_{ja}} \right) q_{ja}. \tag{5}$$

($\Rightarrow$) Because $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$, the two attribute profiles must differ in at least one entry, say, $\alpha_a = 1$ and $\alpha_a^* = 0$. Because $\mathbf{Q}$ is complete, there exists some j such that $\mathbf{q}_j = \mathbf{e}_a$. Thus, $q_{ja} = 1$ and $\prod_{a=1}^{A} \alpha_a^{q_{ja}} = 1$, which implies that $W_a(\boldsymbol{\eta}) \geq 1$. On the other hand, because $\alpha_a^* = 0$, $W_a(\boldsymbol{\eta}^*) = \sum_{j=1}^{J} \left(\prod_{a=1}^{A} (\alpha_a^*)^{q_{ja}} \right) q_{ja} = 0$ regardless of whether $q_{ja} = 1$ or 0 for all a. Hence, $W_a(\boldsymbol{\eta}) \neq W_a(\boldsymbol{\eta}^*)$ always holds; therefore, $\boldsymbol{W}(\boldsymbol{\eta}) \neq \boldsymbol{W}(\boldsymbol{\eta}^*)$.

($\Leftarrow$) The equivalent statement, $\boldsymbol{\alpha} = \boldsymbol{\alpha}^* \Rightarrow \boldsymbol{W}(\boldsymbol{\eta}) = \boldsymbol{W}(\boldsymbol{\eta}^*)$, allows for a more elegant proof. If $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$, then $\alpha_a = \alpha_a^*$ for all a. Based on equation (5), $W_a(\boldsymbol{\eta}) = W_a(\boldsymbol{\eta}^*)$ for all a, which implies that $\boldsymbol{W}(\boldsymbol{\eta}) = \boldsymbol{W}(\boldsymbol{\eta}^*)$. Therefore, $\boldsymbol{W}(\boldsymbol{\eta}) \neq \boldsymbol{W}(\boldsymbol{\eta}^*) \Rightarrow \boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$. □

In summary, two examinees with identical attribute profiles must possess identical ideal item response vectors and identical ideal attribute sum-score vectors. However, two examinees with distinct attribute profiles can have neither identical ideal item response vectors nor identical ideal attribute sum-score vectors. Hence, in the case of error-free item scores, examinees will always be assigned to their true proficiency classes on the basis of either their item-score profiles or their attribute sum-score profiles. It will now be proved that this conclusion does not hold for error-contaminated item scores.

3.3.2. Error-contaminated item scores

Consider two attribute profiles, $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$, and their associated ideal item response vectors, $\boldsymbol{\eta}$ and $\boldsymbol{\eta}^*$. Conceptually, observed item scores can be viewed as ideal item scores perturbed by stochastic error, with the following possible outcomes:

Ideal item scores	Observed item scores
$\boldsymbol{\eta}=\boldsymbol{\eta}^*$	$\boldsymbol{Y}\neq\boldsymbol{Y}^*$ $\boldsymbol{Y}=\boldsymbol{Y}^*$
$\boldsymbol{\eta}\neq\boldsymbol{\eta}^*$	$\boldsymbol{Y}\neq\boldsymbol{Y}^*$ $\boldsymbol{Y}=\boldsymbol{Y}^*$

Hence, the relation between the observed item-score profiles of two examinees does not necessarily reflect the relation between their true proficiency classes. What are the implications for $\boldsymbol{W}$?

Proposition. *Aggregating identical observed item scores always results in identical observed attribute sum-score profiles, but aggregating distinct observed item scores can also result in identical observed attribute sum-score profiles: (i)* $\boldsymbol{Y} = \boldsymbol{Y}^* \Rightarrow \boldsymbol{W} = \boldsymbol{W}^*$*; but (ii)* $\boldsymbol{W} = \boldsymbol{W}^* \not\Rightarrow \boldsymbol{Y} = \boldsymbol{Y}^* \left(\Leftrightarrow \boldsymbol{Y} \neq \boldsymbol{Y}^* \not\Rightarrow \boldsymbol{W} \neq \boldsymbol{W}^* \right)$.

Proof. (i) Because $\boldsymbol{Y} = \boldsymbol{Y}^*$, $Y_j = Y_j^*$ for all j. Thus, $W_a = \sum_{j=1}^{J} Y_j q_{ja} = \sum_{j=1}^{J} Y_j^* q_{ja} = W_a^*$ for all a, implying that $\boldsymbol{W} = \boldsymbol{W}^*$. Therefore, $\boldsymbol{Y} = \boldsymbol{Y}^* \Rightarrow \boldsymbol{W} = \boldsymbol{W}^*$, and $\boldsymbol{W} \neq \boldsymbol{W}^* \Rightarrow \boldsymbol{Y} \neq \boldsymbol{Y}^*$.

(ii) Consider a counterexample that shows that $\boldsymbol{W} = \boldsymbol{W}^* \Rightarrow \boldsymbol{Y} = \boldsymbol{Y}^*$ does not always hold. Suppose that A = 2, J = 5, and

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

If $\boldsymbol{Y}$=(00011) and $\boldsymbol{Y}^*$=(01110), then $\boldsymbol{W}$=$\boldsymbol{W}^*$=(12). Hence, $\boldsymbol{W}$=$\boldsymbol{W}^* \not\Rightarrow \boldsymbol{Y}$=$\boldsymbol{Y}^*$. □

In summary, aggregating the observed item scores of examinees who may belong to distinct proficiency classes can result in their having identical attribute sum-score profiles and therefore risks misclassification of those examinees. The proof suggests six possible scenarios:

Ideal item scores	Observed item scores	Attribute sum scores	Scenario
$\boldsymbol{\eta} = \boldsymbol{\eta}^*$	$\boldsymbol{Y} \neq \boldsymbol{Y}^*$	$\boldsymbol{W} \neq \boldsymbol{W}^*$	1
		$\boldsymbol{W} = \boldsymbol{W}^*$	2
	$\boldsymbol{Y} = \boldsymbol{Y}^*$	$\boldsymbol{W} = \boldsymbol{W}^*$	3
$\boldsymbol{\eta} \neq \boldsymbol{\eta}^*$	$\boldsymbol{Y} \neq \boldsymbol{Y}^*$	$\boldsymbol{W} \neq \boldsymbol{W}^*$	4
		$\boldsymbol{W} = \boldsymbol{W}^*$	5
	$\boldsymbol{Y} = \boldsymbol{Y}^*$	$\boldsymbol{W} = \boldsymbol{W}^*$	6

Scenarios 2 and 5 are problematical because the relation between $\boldsymbol{Y}$ and $\boldsymbol{Y}^*$ is not the same as the relation between $\boldsymbol{W}$ and $\boldsymbol{W}^*$; the first relation suggests that the two examinees with observed item-score profiles $\boldsymbol{Y}$ and $\boldsymbol{Y}^*$, respectively, should be assigned to distinct proficiency classes, whereas the second relation suggests that they should be assigned to the identical proficiency class, which may result in either the correct or the incorrect classification of the examinees, depending on their true proficiency classes.

The consequences of scenarios 2 and 5 are illustrated with two examples, each consisting of the simulated responses of 16 examinees to five items generated using the Q-matrix from the earlier proof; the examinees belong to four distinct proficiency classes (four examinees per proficiency class). For scenarios 2 and 5, Tables 1 and 2 respectively report each examinee's (1) true proficiency class; (2) attribute profile; (3) vector of ideal item responses, $\boldsymbol{\eta}$ (calculated using equation (2)); (4) vector of ideal attribute sum scores; (5) observed item-score profile, $\boldsymbol{Y}$ (sampled from a Bernoulli distribution, with π defined by equation (1), the item response function of the DINA model, with the slipping and guessing parameters drawn from a continuous uniform distribution $\mathcal{U}(0, 0.2)$); and (6) observed attribute sum-score profile, $\boldsymbol{W}$.

Scenario 2: $\boldsymbol{\eta} = \boldsymbol{\eta}^*$, $\boldsymbol{Y} \neq \boldsymbol{Y}^*$, $\boldsymbol{W} = \boldsymbol{W}^*$, In Table 1 for scenario 2, examinees 14, 15, and 16 all belong to the same true proficiency class, but examinees 14 and 16 have distinct observed item-score profiles, $\boldsymbol{Y}$, as do examinees 15 and 16; the observed attribute sum-score profiles, $\boldsymbol{W}$, are identical for all three examinees. Examinees' observed attribute sum-score profiles, $\boldsymbol{W}$, and observed item-score profiles, $\boldsymbol{Y}$, were grouped into four proficiency classes using four different clustering methods: (1) Hartigan and Wong's (1979) K-means clustering, implemented in the `kmeans` routine in R, with 10,000 random restarts and $K = 2^A$ as the known number of underlying clusters; (2) complete-link HACA (Johnson, 1967); (3) average-link HACA (Johnson, 1967); and (4) Ward's (1963) minimum-variance method, the latter three methods implemented in the `hclust` routine in R. Clustering $\mathbf{Y}$ using K-means changes the K-means loss function such that $\boldsymbol{W}$ is replaced by $\boldsymbol{Y}$, and $\overline{\boldsymbol{W}}_k$ by $\overline{\boldsymbol{Y}}_k$, the centroid vector of cluster $\mathcal{C}_k$ obtained by averaging the observed item-score profiles, $\boldsymbol{Y}_{i \in \mathcal{C}_k}$. For the three HACA methods, the (squared) inter-examinee Euclidean distances were computed from $\mathbf{Y}$ and collected into an $N \times N$ input proximity matrix.

The results of the four cluster analyses are reported in Table 3. All but one of the four clustering methods (Ward's method) assigned all 16 examinees to their true proficiency

Table 1. Scenario 2: Simulated responses of 16 examinees to five items underlain by two attributes

Examinee	True proficiency class	Attribute profiles		Ideal item responses					Ideal attribute sum scores		Observed item scores					Observed attribute sum scores	
		α_1	α_2	η_1	η_2	η_3	η_4	η_5	$W_1(\boldsymbol{\eta})$	$W_2(\boldsymbol{\eta})$	Y_1	Y_2	Y_3	Y_4	Y_5	W_1	W_2
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
4	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
5	2	1	0	1	1	0	0	0	2	0	1	1	0	0	0	2	0
6	2	1	0	1	1	0	0	0	2	0	1	1	0	0	0	2	0
7	2	1	0	1	1	0	0	0	2	0	1	1	0	0	0	2	0
8	2	1	0	1	1	0	0	0	2	0	1	1	0	0	1	3	1
9	3	0	1	0	0	1	1	0	0	2	0	0	1	1	0	0	2
10	3	0	1	0	0	1	1	0	0	2	0	0	1	1	0	0	2
11	3	0	1	0	0	1	1	0	0	2	0	0	1	1	0	0	2
12	3	0	1	0	0	1	1	0	0	2	0	0	0	1	0	0	1
13	4	1	1	1	1	1	1	1	3	3	1	1	1	1	1	3	3
→ **14**	4	1	1	1	1	1	1	1	3	3	0	1	1	1	1	2	3
→ **15**	4	1	1	1	1	1	1	1	3	3	0	1	1	1	1	2	3
→ **16**	4	1	1	1	1	1	1	1	3	3	1	0	1	1	1	2	3

Table 2. Scenario 5: Simulated responses of 16 examinees to five items underlain by two attributes

Examinee	True proficiency class	Attribute profiles		Ideal item responses					Ideal attribute sum scores		Observed item scores					Observed attribute sum scores	
		α_1	α_2	η_1	η_2	η_3	η_4	η_5	$W_1(\boldsymbol{\eta})$	$W_2(\boldsymbol{\eta})$	Y_1	Y_2	Y_3	Y_4	Y_5	W_1	W_2
1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
→ **2**	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	2
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	2	1	0	1	1	0	0	0	2	0	1	1	1	0	0	2	1
6	2	1	0	1	1	0	0	0	2	0	1	1	0	0	0	2	0
7	2	1	0	1	1	0	0	0	2	0	1	1	0	0	0	2	0
8	2	1	0	1	1	0	0	0	2	0	1	1	0	0	0	2	0
→ **9**	3	0	1	0	0	1	1	0	0	2	0	1	1	1	0	1	2
10	3	0	1	0	0	1	1	0	0	2	0	0	1	1	0	0	2
11	3	0	1	0	0	1	1	0	0	2	0	0	1	1	0	0	2
12	3	0	1	0	0	1	1	0	0	2	0	0	1	1	0	0	2
13	4	1	1	1	1	1	1	1	3	3	1	1	1	1	1	3	3
14	4	1	1	1	1	1	1	1	3	3	1	1	1	1	1	3	3
15	4	1	1	1	1	1	1	1	3	3	0	1	1	1	1	2	3
16	4	1	1	1	1	1	1	1	3	3	1	1	1	1	1	3	3

Table 3. Scenario 2: Proficiency-class assignments when attribute sum-score profiles, **W**, and item-score profiles, **Y**, serve as input to four clustering methods

				HACA					
	True proficiency	*K*-mean		Complete-link		Average- link		Ward's method	
Examinee	class	***W***	***Y***	***W***	***Y***	***W***	***Y***	***W***	***Y***
1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1
5	2	2	2	2	2	2	2	2	2
6	2	2	2	2	2	2	2	2	2
7	2	2	2	2	2	2	2	2	2
8	2	2	2	2	2	2	2	2	2
9	3	3	3	3	3	3	3	3	3
10	3	3	3	3	3	3	3	3	3
11	3	3	3	3	3	3	3	1	1
12	3	3	3	3	3	3	3	3	3
13	4	4	4	4	4	4	4	4	4
→14	4	4	4	4	4	4	4	4	4
→15	4	4	4	4	4	4	4	4	4
→16	4	4	4	4	4	4	4	4	4
	ARI	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.82**	**0.82**

classes, regardless of whether their attribute sum-score profiles, **W**, or their item-score profiles, **Y**, served as input. For this simulated data set, the true proficiency-class membership of each examinee is known and provides a standard for quantifying the results of the cluster analyses. However, because these clustering methods do not label the clusters representing the proficiency classes with their respective attribute profiles, it is not possible to compute rates of correct classification. Instead, a measure of agreement between the true classification of the examinees and the classification assigned by each clustering method was computed using the Hubert–Arabie Adjusted Rand Index (ARI; Hubert & Arabie, 1985; Steinley, 2004), which has bounds 0 and 1 indicating perfect disagreement and perfect agreement, respectively. A major advantage of the ARI over the more common 'true–false' classification-rate indices is that it incorporates a sophisticated adjustment for random correct classifications. The bottom of Table 3 provides the ARIs for the four clustering methods when attribute sum-score profiles, **W**, and item-score profiles, **Y**, served as input. The ARIs confirm that using either **W** or **Y** as input to the cluster analyses assigned the examinees to their true proficiency classes (except for Ward's method, which consistently misclassified examinee 10, who is not one of the targeted examinees). For this small data set, the ARIs may appear trivial, but for data sets containing hundreds or thousands of examinees, measures such as ARI are indispensable if a conclusive evaluation of the results of competing clustering methods is desired.

Scenario 5: $\boldsymbol{\eta} \neq \boldsymbol{\eta}^*$, $\boldsymbol{Y} \neq \boldsymbol{Y}^*$, $\boldsymbol{W} = \boldsymbol{W}^*$, In Table 2 for scenario 5, examinees 2 and 9 belong to distinct true proficiency classes as reflected by their observed item-score profiles, **Y**, but

not their observed attribute sum-score profiles, **W**. Examinees' observed attribute sum-score profiles, **W**, and observed item-score profiles, **Y**, were grouped into four proficiency classes using the four clustering methods used for scenario 2. The results of the four cluster analyses are reported in Table 4. When item-score profiles, **Y**, served as input, the clustering methods (except for Ward's method) assigned all 16 examinees to their true proficiency classes, but when attribute sum-score profiles, **W**, served as input, examinee 2 was misclassified by all four clustering methods. The ARIs shown at the bottom of Table 4 confirm that the assignment of examinees to their true proficiency classes was superior when **Y** rather than **W** served as input to the cluster analyses (except for Ward's method).

3.4. Sensitivity of attribute sum scores and item scores

A remarkable conclusion can be drawn from the analyses for scenarios 2 and 5. The effects of the use of observed item-score profiles, $\boldsymbol{Y}$, and observed attribute sum-score profiles, $\boldsymbol{W}$, on examinee classification are not symmetric. Distinct observed item-score profiles, $\boldsymbol{Y} \neq \boldsymbol{Y}^*$, do not automatically lead to misclassification when the true proficiency classes are identical, as the analysis for scenario 2 demonstrates. In contrast, the analysis for scenario 5 shows that identical observed attribute sum-score profiles, $\boldsymbol{W} = \boldsymbol{W}^*$, do result in misclassification when the true proficiency classes are distinct.

Why is this so? One possibility is that item-score profiles are more sensitive than attribute sum-score profiles in preserving true proficiency-class membership. That is, the perturbation of two identical ideal item response vectors results in observed item-score profiles that are still so similar to one another that their shared true proficiency class can be identified by a clustering method. On the other hand, the perturbation of two distinct ideal

Table 4. Scenario 5: Proficiency-class assignments when attribute sum-score profiles, **W**, and item-score profiles, **Y**, serve as input to four clustering methods

				HACA					
		K-mean		Complete-link		Average-link		Ward's method	
Examinee	True proficiency class	***W***	***Y***	***W***	***Y***	***W***	***Y***	***W***	***Y***
1	1	1	1	1	1	1	1	1	1
→ **2**	1	3	1	3	1	3	1	3	1
3	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1
5	2	2	2	2	2	2	2	2	2
6	2	2	2	2	2	2	2	2	2
7	2	2	2	2	2	2	2	2	2
8	2	2	2	2	2	2	2	2	2
→ **9**	3	3	3	3	3	3	3	3	3
10	3	3	3	3	3	3	3	3	3
11	3	3	3	3	3	3	3	3	3
12	3	3	3	3	3	3	3	3	3
13	4	4	4	4	4	4	4	4	4
14	4	4	4	4	4	4	4	4	4
15	4	4	4	4	4	4	4	4	3
16	4	4	4	4	4	4	4	4	4
	ARI	**0.82**	**1.00**	**0.82**	**1.00**	**0.82**	**1.00**	**0.82**	**0.82**

item response vectors renders them still so dissimilar to one another that a clustering method is capable of detecting that they belong to distinct true proficiency classes.

What evidence exists to support this possibility? The concept of (dis)similarity invokes the concept of distance. Loosely speaking, the common goal of the clustering methods used in the examples for the two scenarios is to minimize the distances between examinees within a cluster. The results obtained from the cluster analyses for the two scenarios when attribute sum-score profiles, $\mathbf{W}$, and item-score profiles, $\mathbf{Y}$, served as input suggest that item-score profiles more accurately reflect the distances of examinees to the centres of their true proficiency classes than do attribute sum-score profiles. For $\boldsymbol{Y}$, the proficiency-class centres were defined as the expected conditional item response vectors, $\boldsymbol{S}(\boldsymbol{\alpha}) = E(\boldsymbol{Y}|\boldsymbol{\alpha})$ (see equation (4)). For $\boldsymbol{W}$, the proficiency-class centres were defined as the expected conditional attribute sum-score vectors, $\boldsymbol{T}(\boldsymbol{\alpha}) = E(\boldsymbol{W}|\boldsymbol{\alpha})$ (see equation (3)). The squared Euclidean distances of each examinee's attribute sum-score profile, $\boldsymbol{W}$, and item-score profile, $\boldsymbol{Y}$, to the four proficiency-class centres, $d\big(\boldsymbol{Y}, \boldsymbol{S}(\boldsymbol{\alpha})\big)$ and $d\big(\boldsymbol{W}, \boldsymbol{T}(\boldsymbol{\alpha})\big)$, respectively, were computed for all four clustering methods for each of the two scenarios. These distances are reported in Tables 5 and 6.

All four clustering methods should assign each examinee to the nearest proficiency-class centre. Comparing the known true proficiency-class membership with the proficiency-class assignment suggested by the distances of $\boldsymbol{W}$ and $\boldsymbol{Y}$ to the proficiency-class centres shows that the distances $d\big(\boldsymbol{Y}, \boldsymbol{S}(\boldsymbol{\alpha})\big)$ identify the correct proficiency class for all 16 examinees, but that the distances $d\big(\boldsymbol{W}, \boldsymbol{T}(\boldsymbol{\alpha})\big)$ do not. (See examinee 12 in Table 5 and examinee 2 in Table 6.)

3.5. Summary of theoretical and empirical appraisal

The results of the theoretical and empirical appraisal can be briefly stated: the use of item-score profiles, $\mathbf{Y}$, rather than attribute sum-score profiles, $\mathbf{W}$, as input to non-parametric heuristic classification techniques for assigning examinees to proficiency classes does not require knowledge of the Q-matrix, and results in a more accurate classification of examinees.

4. Simulation studies

Three simulation studies were conducted to compare the performance of item-score profiles, $\mathbf{Y}$, with that of attribute sum-score profiles, $\mathbf{W}$, when used as input to the four clustering methods for assigning examinees to proficiency classes under different testing conditions.

4.1. Study I: Equal attribute frequencies in the Q-matrix

4.1.1. Generation of item scores

Examinee item scores conforming to the DINA model were simulated according to the method described by Chiu *et al.* (2009). The experimental design included three variables: number of examinees, $N = 100, 500$; number of attributes, $A = 3, 4$; and number of items, $J = 20, 40, 80$. For the levels of variable A, $2^3-1 = 7$ and $2^4-1 = 15$ distinct binary item-attribute profiles were generated (profiles consisting of all zeros being omitted) to form template Q-matrices of tests containing $J = 20$ items (Chiu *et al.*, 2009, Table 2, p. 650). These Q-matrix designs guaranteed that all attributes occurred with equal frequencies. The Q-matrices for tests containing 40 or 80 items were created by stacking the 20-item template Q-matrices.

An examinee's attribute profile was drawn from either a discrete uniform distribution (i.e., the attribute profile for each proficiency class, $\boldsymbol{\alpha}_k$, had the same probability, $1/K$,

Table 5. Scenario 2: Distances of attribute sum-score profiles, $\boldsymbol{W}$, and item-score profiles, $\boldsymbol{Y}$, to expected proficiency-class centres, $\boldsymbol{T}(_k)$ and $\boldsymbol{S}(_k)$, respectively

Distances between $\boldsymbol{W}$ and $\boldsymbol{T}(_k)$					
Examinee	$d(\boldsymbol{W}, \boldsymbol{T}(\boldsymbol{\alpha}_1))$	$d(\boldsymbol{W}, \boldsymbol{T}(\boldsymbol{\alpha}_2))$	$d(\boldsymbol{W}, \boldsymbol{T}(\boldsymbol{\alpha}_3))$	$d(\boldsymbol{W}, \boldsymbol{T}(\boldsymbol{\alpha}_4))$	True proficiency class
1	**0.38**	4.26	4.38	14.37	1
2	**0.38**	4.26	4.38	14.37	1
3	**0.29**	1.17	4.29	9.99	1
4	**0.29**	1.17	4.29	9.99	1
5	2.21	**0.08**	6.21	7.60	2
6	2.21	**0.08**	6.21	7.60	2
7	2.21	**0.08**	6.21	7.60	2
8	6.56	**1.42**	7.07	2.88	2
9	3.23	7.11	**0.29**	7.69	3
10	3.23	7.11	**0.29**	7.69	3
11	3.23	7.11	**0.29**	7.69	3
12	**0.80**	4.68	1.34	10.03	3
13	13.41	8.28	7.00	**0.21**	4
→ **14**	9.49	7.36	3.07	**0.59**	4
→ **15**	9.49	7.36	3.07	**0.59**	4
→ **16**	9.49	7.36	3.09	**0.59**	4
Distances between $\boldsymbol{Y}$ and $\boldsymbol{S}(_k)$					
Examinee	$d(\boldsymbol{Y}, \boldsymbol{S}(\boldsymbol{\alpha}_1))$	$d(\boldsymbol{Y}, \boldsymbol{S}(\boldsymbol{\alpha}_2))$	$d(\boldsymbol{Y}, \boldsymbol{S}(\boldsymbol{\alpha}_3))$	$d(\boldsymbol{Y}, \boldsymbol{S}(\boldsymbol{\alpha}_4))$	True proficiency class
1	**0.10**	1.75	1.77	4.09	1
2	**0.10**	1.75	1.77	4.09	1
3	**0.78**	0.92	2.45	3.26	1
4	**0.78**	0.92	2.45	3.26	1
5	1.42	**0.06**	3.09	2.40	2
6	1.42	**0.06**	3.09	2.40	2
7	1.42	**0.06**	3.09	2.40	2
8	2.02	**0.66**	3.69	1.71	2
9	1.93	3.57	**0.13**	2.45	3
10	1.93	3.57	**0.13**	2.45	3
11	1.93	3.57	**0.13**	2.45	3
12	0.96	2.60	**0.79**	3.11	3
13	3.85	2.48	2.41	**0.06**	4
→ **14**	3.17	3.31	2.41	**0.89**	4
→ **15**	3.17	3.31	1.57	**0.89**	4
→ **16**	3.21	3.34	2.41	**0.92**	4

Note. The minimum distance to a proficiency-class centre appears in bold type.

where $K = 2^A$) or from a more realistic and complex multivariate normal distribution ($\boldsymbol{\theta} = (\theta_1, \ldots, \theta_A)' \sim \mathcal{N}_A(\mathbf{0}, \boldsymbol{\Sigma})$, where $\mathbf{0}$ indicates the location vector and $\boldsymbol{\Sigma}$, the covariance matrix, with values along the main diagonal equal to 1.00 and off-diagonal entries set to either 0.25 or 0.50), so that each binary attribute, α_a, was linked to a latent continuous ability dimension, θ_a. For each examinee, a vector $\boldsymbol{\theta}_i$ was randomly sampled; if its component values exceeded a predetermined threshold, then the corresponding entry in the examinee's attribute profile, $\boldsymbol{\alpha}_i$, was set to 1:

Table 6. Scenario 5: Distances of attribute sum-score profiles, $\boldsymbol{W}$, and item-score profiles, $\boldsymbol{Y}$, to expected proficiency-class centres, $\mathbf{T}(\boldsymbol{\alpha}_k)$ and $\mathbf{S}(\boldsymbol{\alpha}_k)$, respectively

Distances between $\boldsymbol{W}$ and $\boldsymbol{T}(\boldsymbol{\alpha}_k)$					
Examinee	$d(\boldsymbol{W}, \boldsymbol{T}(\boldsymbol{\alpha}_1))$	$d(\boldsymbol{W}, \boldsymbol{T}(\boldsymbol{\alpha}_2))$	$d(\boldsymbol{W}, \boldsymbol{T}(\boldsymbol{\alpha}_3))$	$d(\boldsymbol{W}, \boldsymbol{T}(\boldsymbol{\alpha}_4))$	True proficiency class
1	**0.42**	4.43	1.00	10.06	1
→ **2**	2.90	3.50	**0.46**	3.36	1
3	**0.30**	4.30	3.90	14.34	1
4	**0.30**	4.30	3.90	14.34	1
5	3.13	**0.32**	3.71	3.20	2
6	3.00	**0.19**	6.60	7.47	2
7	3.00	**0.19**	6.60	7.47	2
8	3.00	**0.19**	6.60	7.47	2
→ **9**	2.90	3.50	**0.46**	3.36	3
10	2.55	6.55	**0.11**	7.79	3
1	2.55	6.55	**0.11**	7.79	3
2	2.55	6.55	**0.11**	7.79	3
3	13.73	7.51	8.27	**0.21**	4
4	13.73	7.51	8.27	**0.21**	4
5	9.37	6.56	3.92	**0.65**	4
6	13.73	7.51	8.27	**0.21**	4

Distances between $\boldsymbol{Y}$ and $\boldsymbol{S}(\boldsymbol{\alpha}_k)$					
Examinee	$d(\boldsymbol{Y}, \boldsymbol{S}(\boldsymbol{\alpha}_1))$	$d(\boldsymbol{Y}, \boldsymbol{S}(\boldsymbol{\alpha}_2))$	$d(\boldsymbol{Y}, \boldsymbol{S}(\boldsymbol{\alpha}_3))$	$d(\boldsymbol{Y}, \boldsymbol{S}(\boldsymbol{\alpha}_4))$	True proficiency class
1	**0.69**	2.45	0.78	3.22	1
→ **2**	**1.41**	3.17	1.50	2.55	1
3	**0.09**	1.85	1.67	4.10	1
4	**0.09**	1.85	1.67	4.10	1
5	2.53	**0.88**	2.58	1.61	2
6	1.72	**0.08**	3.31	2.33	2
7	1.72	**0.08**	3.31	2.33	2
8	1.72	**0.08**	3.31	2.33	2
→ **9**	2.29	2.34	**0.85**	1.58	3
10	1.50	3.25	**0.06**	2.49	3
11	1.50	3.25	**0.06**	2.49	3
12	1.50	3.25	**0.06**	2.49	3
13	3.85	2.20	2.41	**0.06**	4
14	3.85	2.20	2.41	**0.06**	4
15	3.00	3.05	1.57	**0.91**	4
16	3.85	2.20	2.41	**0.06**	4

Note. The minimum distance to a proficiency-class centre appears in bold type.

$$\alpha_{ia} = \begin{cases} 1 & \text{if } \theta_{ia} \geq \Phi^{-1}\left(\frac{a}{A+1}\right), \\ 0 & \text{otherwise.} \end{cases}$$

(Thus, under this condition, the attribute profiles of different proficiency classes, $\boldsymbol{\alpha}_k$, do not have equal probabilities.) The simulated item responses, Y_{ij}, were sampled from a Bernoulli distribution with $\pi_{ij} = P(Y_{ij} = 1)$ defined by equation (1), the item response function of the DINA model. The slipping and guessing parameters, s_j and g_j,

respectively, were drawn from the continuous uniform distribution $\mathcal{U}(0, 0.15)$, allowing only minor deviations from the ideal item responses, or $\mathcal{U}(0, 0.30)$, adding more noise. Completely crossing the three distributions for examinees' attribute profiles and the two distributions for the slipping and guessing parameters with the levels of the variables N, A, and J resulted in an experimental design with $3 \times 2 \times 2 \times 2 \times 3 = 72$ cells. Twenty-five paired ($\mathbf{W}$ and $\mathbf{Y}$) data sets were generated for each cell, for a total of 1,800 paired data sets.

4.1.2. Clustering attribute sum-score profiles and item-score profiles

Examinees' attribute sum-score profiles, $\boldsymbol{W}$, and item-score profiles, $\boldsymbol{Y}$, were grouped into $K = 2^A$ proficiency classes (which were assumed to be known) using the four different clustering methods: Hartigan and Wong's (1979) K-means clustering, implemented in the `kmeans` routine in *R*, with 10,000 random restarts and K as the known number of underlying clusters; complete-link HACA (Johnson, 1967); average-link HACA (Johnson, 1967); and Ward's (1963) minimum-variance method, the latter three methods implemented in the `hclust` routine in *R*.

It was suspected that the different length of the input vectors $\boldsymbol{W}$ and $\boldsymbol{Y}$ (A and J, respectively) might affect the computational efficiency of these clustering routines for large data sets. To provide an idea of the computational effort required, the average CPU times observed on a machine with 8 GB RAM, 2.30 GHz Intel Core processor, and 64-bit OS are reported for the most complex experimental condition. When $N = 500$, $J = 80$, $K = 4$, $s_j, g_j \sim U(0, 0.30)$, and the attribute profiles were generated from an underlying multivariate normal distribution, with $\rho = .50$ among the latent dimensions, the increase in average CPU time (in seconds) for HACA using $\mathbf{Y}$ as cluster input instead of $\mathbf{W}$ was marginal: mean CPU($\mathbf{W}$) = 0.45; mean CPU($\mathbf{Y}$) = 0.58. However, when using K-means with 10,000 random starts, the difference between the average CPU times of $\mathbf{W}$ and $\mathbf{Y}$ became substantial (but still within tolerable limits): mean CPU($\mathbf{W}$) = 6.33; mean CPU($\mathbf{Y}$) = 160.15.

The assignment of examinees to proficiency classes by each of the four clustering methods was evaluated using the Hubert–Arabie ARI (Hubert & Arabie, 1985; Steinley, 2004) described earlier; the ARI quantifies the agreement of the assignment of the examinees to proficiency classes by a clustering method with their true proficiency-class memberships.

4.2. Study I: Results

Tables 7–9 (supporting information) present, separately for each distribution of the slipping and guessing parameters, the results of the simulation for each of the three d istributions used to generate examinees' attribute profiles. For each combination of N, A, and J, each table reports the average (mean) ARIs computed across the 25 simulated data sets when attribute sum-score profiles, $\mathbf{W}$, and item-score profiles, $\mathbf{Y}$, served as input to each of the four clustering methods, plus their ratio. Each ARI for K-means clustering is the average computed across the maxima obtained from 10,000 random restarts for each of the 25 simulated data sets. For comparison, the table column labelled 'EM' reports the average ARIs computed across 25 replications of examinee classification based on the DINA model parameters produced by EM. Each average ARI is multiplied by 1,000 for readability; each ratio of the average ARI for $\mathbf{W}$ to its corresponding ARI for $\boldsymbol{Y}$ remains in decimal form. Of course, because the simulated item responses conform perfectly to the DINA model, EM should outperform the four clustering methods.

The three tables show that, for the majority of the 72 cells of the experimental design, the performance of all four clustering methods in assigning examinees to their true proficiency classes was better when item-score profiles, **Y**, rather than attribute sum-score profiles, **W**, served as input; there are only 12 cases – two for *K*-means clustering and ten for average-link HACA – in which the average ARI for $\boldsymbol{W}$ exceeds the average ARI for $\boldsymbol{Y}$.

To facilitate further interpretation, Tables 7–9 (supporting information) were also summarized by boxplots as shown in Figure 1 (supporting information). The display presents separately for the two levels of error perturbation (as defined by the distributions of the slipping and guessing parameters) the results of *K*-means clustering and average-link HACA (which were chosen as representative examples due to space restrictions). Each of the four rows in Figure 1 (supporting information) shows the effect of the different levels of one of the experimental variables on cluster recovery when attribute sum-score profiles, **W**, and item-score profiles, **Y**, served as input. The first row refers to the number of examinees, N; the second to the number of items, J; the third to the number of attributes, A; and the fourth to the three distributions used to generate examinees' attribute profiles. The boxplots show that the assignment of the examinees to their true proficiency classes generally deteriorated when either the level of error perturbation increased from 0.15 to 0.30 or the number of attributes, A, was raised from 3 to 4. The number of examinees, N, and the specific distribution used to generate the examinees' attribute profiles do not appear to have an effect. The most important observation is that for both levels of error perturbation, a larger number of items, J, led to more accurate classification regardless of whether attribute sum-score profiles, **W**, or item-score profiles, **Y**, served as input; and the difference between the accuracy of examinee classification when **W** or **Y** served as input decreased for all four clustering methods as the number of items increased.

4.3. Study II: Unequal frequencies of attributes in the Q-matrix

In study I, all attributes occurred with equal frequencies in the template Q-matrices as part of the experimental control. However, this condition might appear too restrictive in comparison with real-world test applications, where it seems unlikely that the frequencies of all attributes in the Q-matrix of a test are balanced. Thus, the purpose of study II was to examine the effect of unequal attribute frequencies in the template Q-matrices on the performance of **W** and **Y** as cluster input. Except for this important modification, the design of study II was identical to that of study I. Specifically, for the template Q-matrices containing $J = 20$ items, the following imbalanced frequency patterns were used: when $A = 3$, the three attributes occurred with frequencies 13, 11, and 9; when $A = 4$, the four attributes occurred with frequencies 11, 8, 7, and 7. (The different frequency patterns were not crossed with A.) In addition, the number of examinees was fixed at $N = 500$, because the findings from study I suggested that the effect of different sample sizes on the performance of $\boldsymbol{W}$ and $\boldsymbol{Y}$ was negligible. Thus, study II consisted in total of $3 \times 1 \times 2 \times 2 \times 3 = 36$ cells. Twenty-five paired (**W** and **Y**) data sets were generated for each cell and served as input to the same four clustering methods as were used in study I. The recovery of the true proficiency classes was assessed by the ARI.

4.4. Study II: Results

The findings from study II are summarized in Tables 10–12 (supporting information) which are organized like Tables 7–9 (supporting information) of study I (note that N was

fixed at 500 examinees). The results are very similar to those obtained in study I. For the 36 cells of the experimental design, the performance of the four clustering methods in assigning examinees to their true proficiency classes is better when item-score profiles, **Y**, rather than attribute sum-score profiles, **W**, served as input. Tables 10–12 (supporting information) were also condensed into multiple boxplots (see Figure 2, supporting information), which support the observation that the results of the two studies are almost identical. In summary, a comparison of the findings of studies I and II suggests that the performance of **W** and **Y** as input to clustering does not depend on whether the overall frequencies of attributes in the template Q-matrices are balanced.

4.5. Study III: Large number of attributes underlying a test

As a second variant of a more realistic testing scenario, study III examined how the performance of **W** and **Y** as input to clustering was affected when a test involves a large number of attributes. Eight attributes were chosen as an extreme condition; because $A = 8$, there were $K = 2^8 = 256$ different proficiency classes. So 255 distinct item-attribute profiles can be used in the Q-matrix. To ensure that at least more than 10% of these candidate items were included in the template Q-matrices, only tests with $J = 40$ and $J = 80$ items were used in study III. The number of examinees was again set to $N = 500$. The remaining features of the study design were identical to those of study I. Hence, study III consisted of $3 \times 2 \times 1 \times 1 \times 2 = 12$ cells in total. Twenty-five paired (**W** and **Y**) data sets were generated for each cell and served as input to the four clustering methods previously used; the recovery of the true proficiency classes was assessed by the ARI.

4.6. Study III: Results

Tables 13–15 (supporting information) report – in the now familiar lay-out–the results of study III (note that now N and A were fixed). As in the previous two simulation studies, the use of the item-score profiles, **Y**, as input to the four clustering methods resulted in better recovery of the true proficiency classes than the attribute sum-score profiles, **W**. (In comparison with the parametric EM procedure, however, the non-parametric classification methods performed poorly in assigning examinees to their true proficiency classes when A was large.) The boxplots based on Tables 13–15 in supporting information (see Figure 3, supporting information) seem to indicate that, differently from studies I and II, the distributions used for generating the attribute profiles had a stronger effect on the recovery of the true proficiency classes when A was large. The classification of examinees appears to deteriorate when the attribute profiles come from a multivariate normal distribution. In summary, the findings of study III show that **Y** maintained its superiority over **W** as input to clustering but, in comparison with the previous studies, at a lower level.

5. Practical application

The real-world fraction-subtraction data set (Tatsuoka, 1984) is one of the most thoroughly studied in cognitive diagnosis research (e.g., Chiu, 2013; Chiu & Douglas, 2013; de la Torre, 2008, 2009; de la Torre & Douglas, 2008; DeCarlo, 2011; Mislevy, 1996; Tatsuoka, 2002; Templin, Henson, & Douglas, 2007). A subset of this data set consisting of the responses of 536 middle-school students to a collection of 15 test items (de la Torre,

2008) is used here to compare the ability of the four clustering methods to assign examinees to their true proficiency classes when attribute sum-score profiles, **W**, and item-score profiles, **Y**, serve as input. (As noted by DeCarlo, 2011, p. 24, a number of papers – including de la Torre, 2008 – have reported a sample size of 2,144 observations for the fraction-subtraction data; this is a mistake that arose when the original data were stacked four times.) The 15 fraction-subtraction test items and the five attributes required to answer them correctly are shown in the Q-matrix in Table 16 (supporting information).

Using this Q-matrix, de la Torre (2008) found that the DINA model fitted the fraction-subtraction data set subset well; with one exception, all item-parameter estimates were less than 0.30. Therefore, for illustrative purposes, the DINA model is considered here to be the 'true' model underlying the item responses, and the proficiency classes to which EM assigns the examinees are considered to be the 'true' proficiency classes. Note that with five attributes, there are 32 possible proficiency classes. Only 24 proficiency classes were identified using EM. They are used here as the standard of comparison for evaluating the proficiency-class assignments produced by the four clustering methods.

Table 17 (supporting information) reports the results of the four cluster analyses. The ARIs for the cluster analyses of the item-score profiles, $\boldsymbol{Y}$, always exceed those for the cluster analyses of the attribute sum-score profiles, $\boldsymbol{W}$. (The relatively low values of the ARIs are due to the small number of test items, as predicted by the ACTCD.)

6. Discussion

The ACTCD (Chiu *et al.*, 2009) provided a foundation for using non-parametric classification techniques in cognitive diagnosis modelling as heuristic or approximate methods for assigning examinees to proficiency classes. These methods require as input a statistic for $\boldsymbol{\alpha}$ that is based on the observed item responses. For educational data that conform to the DINA model, the theoretical properties and the empirical performance of two statistics for $\boldsymbol{\alpha}$, the item-score profile, $\boldsymbol{Y}$, and the attribute sum-score profile, $\boldsymbol{W}$, were compared. $\boldsymbol{W}$ is theoretically well supported by the ACTCD. The consistency theorem of classification states that the probability of assigning examinees correctly to their true proficiency classes based on their attribute sum-score profiles, $\boldsymbol{W}$, approaches 1 as the length of a test (i.e., the number of items, J) approaches infinity. But it was also shown that aggregating each examinee's item-score profile, $\boldsymbol{Y}$, into an attribute sum-score profile, $\boldsymbol{W}$, can result in the representation of distinct $\boldsymbol{Y}$ – where examinees possibly belong to different proficiency classes – by identical $\boldsymbol{W}$, which may then lead to the misclassification of those examinees. In fact, the simulation studies and the practical application demonstrated that assignment of examinees to their true proficiency classes was more accurate when their item-score profiles, **Y**, rather than their attribute sum-score profiles, **W**, served as input to four different clustering methods. These results create a dilemma, because the theoretically well-supported statistic $\boldsymbol{W}$ is outperformed empirically by the theoretically indeterminate statistic $\boldsymbol{Y}$. Four questions regarding the theoretical and empirical implications of these findings remain to be addressed here.

First, does the inferior performance of the attribute sum-score profiles contradict the consistency theorem of classification of the ACTCD? The results of the simulation studies suggest that the answer to this question is 'no'. Recall that the ACTCD is an asymptotic theory: the probability of assigning examinees correctly to their true proficiency classes approaches 1 as the number of items, J, approaches infinity. In fact, under all experimental conditions, a larger number of items in a test led to more accurate classification regardless

of whether attribute sum-score profiles, **W**, or item-score profiles, **Y**, were used as input. Also, the difference between the accuracy of examinee classification when **W** or **Y** served as input decreased for all four clustering methods as the number of items increased. Therefore, the empirical findings support rather than contradict the consistency theorem of classification. (As an important aside, one should recall that MLE procedures also rely on asymptotic properties: generally the accuracy of the parameter estimates increases as the number of observations increases.)

Second, assigning examinees to proficiency classes based on $\boldsymbol{Y}$ does not require the Q-matrix of a test to be known. But the Q-matrix is the centrepiece of any cognitive diagnosis model because it specifies the item-attribute constraints. So, does assigning examinees to proficiency classes without **Q** not mean the elimination of the theoretical connection between individual items and the attribute profiles, $\boldsymbol{\alpha}$, that define the different proficiency classes? More to the point, does using **Y** as input to clustering not completely abandon the theoretical framework of cognitive diagnosis? Several observations suggest that the answer to this question is also 'no'. Recall that $\boldsymbol{Y}$ is linked to the attribute profiles, $\boldsymbol{\alpha}$, via the latent ideal item-response profile, $\boldsymbol{\eta}$; that is, $\boldsymbol{Y}$ is '$\boldsymbol{\eta}$ plus error' – similar to $\boldsymbol{W}$, which is '($\boldsymbol{\eta}$ plus error)**Q**'. In fact, it was proved that the item-response profile $\boldsymbol{Y}$, like the attribute sum-score profile $\boldsymbol{W}$, guarantees well-separated centres of the different proficiency classes and is, therefore, a legitimate statistic for $\boldsymbol{\alpha}$. Thus, theoretical considerations and the empirical evidence of the simulation studies suggest that the consistency theorem of classification also applies when item-score profiles, **Y**, serve as input to clustering. At present, however, this theorem cannot be proved for $\boldsymbol{Y}$, because the dimensionality of $\boldsymbol{Y}$ depends on J (a difficulty elegantly avoided by $\boldsymbol{W}$, because its dimensionality depends on A). In summary, further research is needed for a better understanding of the complex relation between $\boldsymbol{Y}$ and $\boldsymbol{\alpha}$ in all its detail.

Third, as noted in the introduction, software for fitting cognitive diagnosis models to educational data with MLE procedures tends to be proprietary and hence unavailable or expensive. But recently Robitzsch, Kiefer, George, and Uenlue (2014) have developed the package `CDM` ('Cognitive Diagnosis Modeling') for *R* that provides an implementation of the EM algorithm for fitting the DINA model and the DINO model (i.e., the Deterministic Input Noisy Output 'OR' gate model; Templin & Henson, 2006). So, is there any justification for further exploring the use of heuristic classification techniques given this development? The answer to this question is 'yes'. MLE procedures for assigning examinees to proficiency classes typically require that the Q-matrix of a test be known, whereas non-parametric classification techniques using item-score profiles, **Y**, as input do not. In practice, there is often no guarantee that the Q-matrix has been correctly determined (thus, risking the possible misclassification of examinees). This can become an issue especially when existing data that were originally collected with a test not based on the cognitive diagnosis framework are analysed *post hoc* with a cognitive diagnosis model (*retro-fitting*; Leighton & Gierl, 2007). As an additional complication, Liu, Xu, and Ying (2012, 2013) proved that for many practical applications the Q-matrix cannot be uniquely identified beyond a class of equivalent Q-matrices that all lead to an identical distribution of responses. Hence, one should be prepared for misspecified Q-matrices as the rule rather than the exception. Therefore, from a practitioner's view, the availability of procedures for assigning examinees to proficiency classes without prior knowledge of the Q-matrix would be highly desirable. Using item-score profiles, $\boldsymbol{Y}$, instead of $\boldsymbol{W}$ appears to be an option, but, as was discussed in the previous paragraph, further research is required for a more conclusive assessment.

Finally, what recommendations can be made to educational practitioners? First, they should be aware that non-parametric classification techniques were proposed as a heuristic alternative for assigning examinees to proficiency classes when MLE methods fail or are difficult to implement (Chiu *et al.*, 2009). The bottom line is that MLE methods should be used whenever they are feasible – that is, when the Q-matrix is well defined and the cognitive diagnosis model underlying the data is known. Under these regular conditions MLE methods typically outperform HACA in assigning examinees to proficiency classes, especially when a test is short. Also, by definition, non-parametric classification techniques cannot provide item parameter estimates and their standard errors. Thus, these methods should not be considered when item evaluation is the primary goal of the analysis. Second, the examinee clusters obtained from non-parametric classification methods serve as proxies for the proficiency classes. But non-parametric classification methods cannot estimate the attribute profiles underlying the clusters. Hence, the clusters must be interpreted or *labelled* – that is, their underlying attribute profiles must be reconstructed from the chosen input data, which can be tedious if the number of examinees is large. For $\mathbf{W}$ as input to clustering, Chiu *et al.* (2009) developed an automatic cluster labelling algorithm that seeks to identify an optimal match between examinees' within-cluster sum-score profiles and candidate attribute profiles potentially underlying this cluster. But this algorithm is 'experimental' and not available for routine applications; for $\mathbf{Y}$ as input to clustering, an automatic cluster labelling algorithm does not (yet) exist. Thus, for now, practitioners are recommended to be cautious when exploring $\mathbf{Y}$ as input to non-parametric classification techniques. Further research is required to deepen the understanding of the merits of $\boldsymbol{Y}$ and to develop procedures that facilitate automatic cluster labelling in terms of their underlying attribute profiles.

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. de Baker, T. Barnes, & J. E. Beck (Eds.), *Educational data mining 2008: Proceedings of the 1st International Conference on Educational Data Mining* (pp. 210–217). Montréal, Québec, Canada: International Data Mining Society.

Bock, H.-H. (2007). Clustering methods: A history of k-means algorithms. In P. Brito, P. Bertrand, G. Cucumel, & F. De Carvalho (Eds.), *Selected contributions in data analysis and classification* (pp. 161–172). Berlin, Germany: Springer.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement*, *37*, 598–618. doi:10.1177/0146621613488436

Chiu, C.-Y., & Douglas, J. A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, *30*, 225–250. doi:10.1007/s00357-013-9132-9

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633–665. doi:10.1007/s11336-009-9125-0

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343–362. doi:10.1111/j.17453984.2008.00069.x

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115–130. doi:10.3102/1076998607309474

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

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*, 595–624. doi:10.1007/s11336-008-9063-2

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8–26. doi:10.1177/0146621610377081

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Vol. 26, Psychometrics* (pp. 979–1030). Amsterdam, the Netherlands: North-Holland.

Forgy, E. W. (1965). Cluster analyses of multivariate data: Efficiency versus interpretability of classifications [Abstract #1130]. *Biometrics*, *21*, 768–769.

Fu, J., & Li, Y. (2007). An integrative review of cognitively diagnostic psychometric models. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skills diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Vol. 26, Psychometrics* (pp. 1031–1038). Amsterdam, the Netherlands: North-Holland.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A *K*-means clustering algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *28*, 100–108.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210. doi:10.1007/s11336-008-9089-5

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218. doi:10.1007/BF02289588

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*, 241–254. doi:10.1007/BF02289588

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. doi:10.1177/01466210122032064

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

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement*, *36*, 548–564. doi:10.1177/0146621612456591

Liu, J., Xu, G., & Ying, Z. (2013). Theory of the self-learning *Q*-matrix. *Bernoulli*, *19*, 1790–1817. doi:10.3150/12-BEJ430

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). Berkeley, CA: University of California Press.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *2*, 99–120. doi:10.3102/10769986002002099

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*, 379–416. doi:10.1111/j.17453984.1996.tb00498.x

Park, Y. S., & Lee, Y.-S. (2011). Diagnostic cluster analysis of mathematics skills. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *4*, 75–107.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2014). *CDM: Cognitive Diagnosis Modeling. R package version 3.1-14*. Retrieved from the Comprehensive R Archive Network [CRAN] website http://CRAN.R-project.org/package=CDM

Rupp, A. (2007). Unique characteristics of cognitive diagnosis models. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Rupp, A. A., & Templin, J. L., & Henson, R. A. (2010). *Diagnostic Measurement. Theory, Methods, and Applications*. New York, NY: Guilford.

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted rand index. *Psychological Methods*, *9*, 386–396. doi:10.1037/1082-989X.9.3.386

Steinley, D. (2006). *K*-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*, 1–34. doi:10.1348/000711005X48266

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

Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems* (Report NIE-G-81-0002). Urbana, IL: University of Illinois, Computer-based Education Research Library.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconception by the pattern classification approach. *Journal of Educational Statistics*, *10*, 55–73. doi:10.2307/1164930

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305. doi:10.1037/1082-989X.11.3.287

Templin, J. L., Henson, R. A., & Douglas, J. A. (2007). *General theory and estimation of cognitive diagnosis models: Using Mplus to derive model estimates*. Unpublished manuscript.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–301. doi:10.1348/000711007X193957

Ward, J. H. jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236–244.

Willse, J., Henson, R., & Templin, J. (2007). Using sum scores or IRT in place of cognitive diagnostic models: Can more familiar models do the job? Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Received 24 October 2012; revised version received 12 July 2014

Supporting Information

The following supporting information may be found in the online edition of the article:

Data S1. Supporting Information.