PSYCHOMETRIKA—VOL. 74, NO. 4, 633–665
DECEMBER 2009
DOI: 10.1007/S11336-009-9125-0

CLUSTER ANALYSIS FOR COGNITIVE DIAGNOSIS: THEORY AND APPLICATIONS

CHIA-YI CHIU

RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY

JEFFREY A. DOUGLAS

UNIVERSITY OF ILLINOIS

XIAODONG LI

MERCK & COMPANY, INC.

Latent class models for cognitive diagnosis often begin with specification of a matrix that indicates which attributes or skills are needed for each item. Then by imposing restrictions that take this into account, along with a theory governing how subjects interact with items, parametric formulations of item response functions are derived and fitted. Cluster analysis provides an alternative approach that does not require specifying an item response model, but does require an item-by-attribute matrix. After summarizing the data with a particular vector of sum-scores, K-means cluster analysis or hierarchical agglomerative cluster analysis can be applied with the purpose of clustering subjects who possess the same skills. Asymptotic classification accuracy results are given, along with simulations comparing effects of test length and method of clustering. An application to a language examination is provided to illustrate how the methods can be implemented in practice.

Key words: cluster analysis, cognitive diagnosis, latent class analysis.

1. Introduction

In educational testing research, specialized latent class models for cognitive diagnosis have been developed to classify mastery or nonmastery of each attribute in a set of attributes the exam is designed to assess. Like in item response theory, item parameters can be of interest. However, the ultimate goal of applying diagnostic models is to classify subjects into one of several different categories describing their attribute profiles. These attributes can take many forms, depending on the application, but often correspond one-to-one with specific skills needed to answer items on an exam or other psychological assessment. Classification according to these fine-grained skills is desired when specific information on knowledge states is required, and one important expectation is that it can lead to more efficient remediation. The cognitive diagnosis models that have recently gained attention share many similarities with other models or classification techniques that are familiar in psychometrics. General latent class modeling, without the structure imposed by cognitive diagnosis models, provides yet another way to search for homogeneous groups of subjects. However, these unrestricted latent class models do not make use of expert knowledge of the item and content matter, and consequently contain too many parameters to be efficient, and not enough structure to yield useful conclusions.

Classical techniques from exploratory multivariate data analysis such as cluster analysis, discriminant analysis, and multidimensional scaling are alternatives to these cognitive diagnosis

We would like to thank the English Language Institute at the University of Michigan for data and the National Science Foundation for funding (grant number 0648882).

Requests for reprints should be sent to Jeffrey A. Douglas, 101 Illini Hall, 725 S. Wright St., Champaign, IL 61820, USA. E-mail: jeffdoug@uiuc.edu

models, and other latent trait and latent class models, and do not require specifying models and corresponding likelihood functions. Cluster analysis and discriminant analysis can be promising tools for cognitive diagnosis, and share the goal of grouping objects in the correct way. In the context of cognitive diagnosis, this means placing subjects into groups or clusters that are homogeneous with regard to their attributes or skills.

The primary objective of this paper is to introduce methods for conducting cognitive diagnosis using only familiar techniques of cluster analysis, after first tailoring the input to recognize the assumed skill requirements of the items. This requires clustering on a well-chosen summary of the data, utilizing some of the same assumptions derived by expert opinion that are used in the latent class models for cognitive diagnosis. However, no further model assumptions are required. One limitation of the current status of cognitive diagnostic modeling is that it requires specialized software. By developing tailored methods of cluster analysis for this application, users can run familiar and widely available software to conduct cluster analysis, and depending on the method that is used, computer run time can be very short.

We begin with a review of latent class models and restricted latent class models for cognitive diagnosis. Following this in Section 3, cluster analysis is reviewed and methods of cluster analysis for cognitive diagnosis are introduced as alternatives. Section 4 studies the long test behavior of the proposed clustering methods under particular choices of the underlying model. A simulation study is provided in Section 5, and an analysis of real language testing data is given in Section 6 before concluding with a discussion of the results and remarks on some promising research directions.

2. Latent Class Models for Cognitive Diagnosis

In this section, a brief review of selected cognitive diagnosis models is provided. Though this is not a complete survey of such models, the aim is to provide examples that span many of the distinguishing characteristics that help define these models. Though many of the techniques discussed here can be extended to ordinal and even continuous data, binary responses are assumed throughout.

We begin by discussing the general latent class model for multiple binary responses, then consider more structured models that utilize assumptions made by experts concerning the attributes or skills that are required for each item and how these are combined to generate responses. This combination of attributes is often dictated by whether the skills or attributes operate in a compensatory, disjunctive, or conjunctive fashion. Examples of each of these are given. These models will be studied in later sections as potential underlying models for the data upon which the cluster analysis methods will be applied.

2.1. Unrestricted Latent Class Model

Unlike latent trait models which posit continuous latent variables, latent class models assume finite discrete latent variables. Under the assumption that the latent variable is nominal, the unrestricted model makes no assumptions about particular attributes required for items, nor does it assume how the attributes are utilized. In this regard, it seems simpler than more restricted models. However, the assumptions imposed by restrictive models actually reduce the number of parameters required. When a sufficient number of classes are included, the unrestricted model is more general and encompasses the restricted models that follow.

In the unrestricted model, the probabilities of correct responses to J items are parameterized by J different Bernoulli proportions which are distinct when taken as a collection, for each of

the M classes. Specifically, given a subject i in class m, the item response function of the model is specified as

$$P(Y_{ij} = 1 \mid m) = \pi_{jm}, \tag{1}$$

where π_{jm} is the probability of a subject in class m endorsing or answering item j correctly, depending on the context. This probability is constant across the class, but varies with the item.

The remaining feature of this model is conditional independence, which implies that knowledge of the correct latent class explains all of the dependence in the observable variables. Given the assumption of conditional independence, the likelihood function of the collection of N independent response vectors is expressed as

$$L(\boldsymbol{\pi}, \boldsymbol{\zeta} \mid \mathbf{y}) = \prod_{i=1}^{N} \left[\sum_{m=1}^{M} \zeta_m \prod_{j=1}^{J} \pi_{jm}^{y_{ij}} (1 - \pi_{jm})^{(1-y_{ij})} \right], \tag{2}$$

where ζ_m is the unknown population proportion of class m, so that $\sum_{m=1}^{M} \zeta_m = 1$. The convention, after fitting all of the parameters of the model, is to assign a subject to the class for which the posterior probability of membership is maximized. An obvious source of unidentifiability is that the class labels can be assigned arbitrarily. Once a particular permutation of these has been fixed, problems of local maxima of the likelihood function still arise. An in-depth discussion of this model including details for fitting and determining the number of classes can be found in Bartholomew (1987).

2.2. *Restricted Latent Class Models for Cognitive Diagnosis*

Specialized latent class models for cognitive diagnosis are derived under assumptions on which attributes are needed for which items, and how the attributes are utilized to construct a response. Let $\boldsymbol{\alpha}$ be a K-dimensional vector for which the kth entry α_k, indicates whether or not a subject possesses the kth attribute or skill, for $k = 1, 2, \ldots, K$. An attribute might refer to a clearly defined skill in some applications, or a more abstract psychological construct in another. All restricted latent class models for cognitive diagnosis that we consider require a $J \times K$ matrix $\boldsymbol{Q}$, referred to as a Q-matrix (Tatsuoka, 1985), with (j, k) entry q_{jk} denoting whether or not the jth item requires the kth attribute. The vector $\boldsymbol{\alpha}$ can take 2^K distinct values. These values index the 2^K latent classes in such models. What distinguishes models from one another are the assumptions that dictate how attributes are utilized to construct responses. We do not provide a complete survey of models, but do give examples that represent assumptions that attributes are required in conjunctive, disjunctive, or compensatory ways. A recent and thorough review of latent class models for cognitive diagnosis can be found in Rupp and Templin (2007).

A simple example of a conjunctive model is the DINA (Deterministic Input, Noisy Output "AND" gate) model (Junker & Sijtsma, 2001). The DINA model extends the work of Macready and Dayton (1977), which considers a two-class version of it for assessing mastery of a skill. The item response function of the DINA model is

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = (1 - s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})}, \tag{3}$$

where for all i, $s_j = P(Y_{ij} = 0 \mid \eta_{ij} = 1)$ and $g_j = P(Y_{ij} = 1 \mid \eta_{ij} = 0)$ are the probabilities of slipping and guessing, respectively, for the jth item, and η_{ij} is the ideal response which connects the attribute pattern possessed by a subject and the elements of $\boldsymbol{Q}$ in the following way:

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}. \tag{4}$$

The variable η_{ij} indicates whether the subject possesses all the attributes needed for answering the particular item. Therefore, the DINA model is characterized by its conjunctive feature that the probability of answering an item correctly will severely drop if any of the required attributes are not mastered or possessed. Fitting the DINA model can be done with the EM algorithm (Haertel, 1989), or by use of Markov chain Monte Carlo (de la Torre & Douglas, 2004; Tatsuoka, 2002). Templin, Henson, and Douglas (2007) discuss how to fit cognitive diagnosis models, including the DINA model as well as the remaining models in this section, using the software Mplus (Muthén & Muthén, 2006).

The NIDA (Noisy Input, Deterministic Output "And" gate) model, introduced in Maris (1999), and named in Junker and Sijtsma (2001), considers slips and guesses at the subtask level, where skills are applied in steps to construct an overall response. This can be viewed as a latent class version of the multicomponent item response model of Embretson (1997). Let η_{ijk} indicate whether the ith subject correctly applied the kth attribute in completing the jth item. Slipping and guessing parameters are indexed by attribute rather than by item, in the case of the DINA model, and are defined by $s_k = P(\eta_{ijk} = 0 \mid \alpha_{ik} = 1, q_{jk} = 1)$ and $g_k = P(\eta_{ijk} = 1 \mid \alpha_{ik} = 0, q_{jk} = 1)$. It is understood that $P(\eta_{ijk} = 1 \mid q_{jk} = 0)$ equals 1, regardless of the value of α_{ik}. In the NIDA model, an item response Y_{ij} is 1 if all η_{ijk}'s are equal to 1, $Y_{ij} = \prod_{k=1}^{K} \eta_{ijk}$. By assuming the η_{ijk}'s are independent conditional on $\boldsymbol{\alpha}_i$, the item response function has the form

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

The NIDA model is quite restrictive in that the slipping and guessing parameters for the different attributes are constant over items. A generalization of this that loosens that restriction is a reduced version of the Reparameterized Unified Model, called the Reduced RUM (Hartz, Roussos, Henson, & Templin, 2005). In the Reduced RUM, the item response function is

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*q_{jk}(1-\alpha_{ik})}, \tag{5}$$

where π_j^* denotes the probability of answering correctly for someone who possesses all of the required attributes, and r_{jk}^* is a parameter between 0 and 1 that represents the penalty for not possessing the kth attribute. Though the parameters of the Reduced RUM appear different than the slipping and guessing parameters of the NIDA, they amount to a reparameterization that helps identify the model when slipping and guessing probabilities are allowed to vary across items.

Whereas conjunctive models require the intersection of a set of attributes or successful implementations of these attributes, disjunctive models essentially replace "and" with "or." As an example, Templin and Henson (2006) introduced the DINO (Deterministic Input, Noisy Output "Or" gate) model. The item response function of the DINO model is expressed as

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

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

Compensatory models allow subjects to miss some attributes, or be low on them in the case of continuous latent traits, and make up for them with the remaining attributes. The basic characteristic of many of these is that the latent variables operate in an additive way under some link function that is monotonically related to the probability of a correct response. Compensatory models can allow for more equivalence classes of response probabilities per item than a strict

conjunctive model such as the DINA, but are not as specific in the way they model the response process. In our study, the general diagnostic model of von Davier (2005) is taken as the representative of a latent class compensatory model. Although the GDM generalizes to different item types, continuous as well as discrete latent variables, and through the use of interactions can even be made conjunctive, the version we consider here as an example of a conjunctive model and has the item response function,

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \frac{\exp[\beta_j + \sum_{k=1}^{K} \gamma_{jk} q_{jk} \alpha_{ik}]}{1 + \exp[\beta_j + \sum_{k=1}^{K} \gamma_{jk} q_{jk} \alpha_{ik}]}$$

where β_j serves as an intercept parameter and γ_{jk} plays the role of a discrimination parameter for the jth item and kth attribute. This version of the GDM is essentially a version of the logistic multidimensional item response model, only with binary latent variables. This model will be revisited in a later section, as a way of illustrating how sum-scores related to the attributes need to be treated as a vector rather than one at a time, due to the possibility of compensating for a lacking skill and falsely giving the impression that it is mastered.

3. Cluster Analysis for Cognitive Diagnosis

The latent variable models discussed in the previous section all require sophisticated software for fitting, either with the EM algorithm or by Markov chain Monte Carlo. The aim of this section is to describe how one can construct appropriate sum-scores that can be used to cluster subjects into what are essentially the correct latent classes. We first introduce the sum-score statistic, then discuss two common methods of cluster analysis that can be used to arrive at the classifications.

The method begins by first constructing a vector of sum-scores. For the ith subject, the variable $\boldsymbol{W}_i = (W_{i1}, W_{i2}, \ldots, W_{iK})'$ is defined as a vector of sum-scores with kth component,

$$W_{ik} = \sum_{j=1}^{J} Y_{ij} q_{jk}. \tag{6}$$

Thus, the components of $\boldsymbol{W}$ are just sum-scores for items corresponding to each attribute. Because many items will require more than 1 attribute, some items will contribute to more than just one component of $\boldsymbol{W}$.

The vector $\boldsymbol{W}$ is then taken as the input to a user-chosen method of cluster analysis, with a fixed number 2^K clusters, the same as the number of latent classes in the cognitive diagnosis models discussed above. The methods of cluster analyses we have investigated for this application are K-means (MacQueen, 1967), and hierarchical agglomerative cluster analysis (Hartigan, 1975), which will be referred to as HACA. The motivation behind these choices is that they are common and widely available procedures. Furthermore, the asymptotic theory for classification of subjects, to be presented later, assumes the use of HACA, either with single linkage or complete linkage. In some ways, the corresponding theory for K-means is less tractable and is still under investigation. Nevertheless, in simulation studies, it appears to outperform HACA in many realistic situations, and is retained for that reason. Next, we provide a review of these methods of cluster analysis in the context of the problem at hand.

3.1. Distance Measures

As mentioned previously, algorithms of cluster analyses are built up on a basis of similarity and dissimilarity measures. Different options of distance measures are taken into account according to the needs of the given task. Taking the K-means clustering as an example, the Euclidean

distance (or squared Euclidean distance) is usually the preferred measure, though it tends to produce ball-shaped clusters and this may not be desired. Several popular distance measures are discussed below.

Minkowski p-metric is a general class of distance metrics. For two K-dimensional data points $\mathbf{w}_i$ and $\mathbf{w}_{i'}$, the Minkowski distance d_{L_p} is defined as

$$d_{L_p}(\mathbf{w}_i, \mathbf{w}_{i'}) = \left[\sum_{k=1}^{K} \left(|w_{ik} - w_{i'k}|\right)^p\right]^{1/p}, \tag{7}$$

where K is the number of dimensions. If we take $p = 2$, the distance becomes the familiar Euclidean distance, and d_{E} is taken as its denotation.

Euclidean distance is the most common distance measure, and for data with different units, the distance is often applied to standardized data. Because it puts equal weight to each variable (dimension), all variables are equally important in determining the relative closeness of the objects. As mentioned previously, Euclidean distance tends to yield ball-shaped clusters. The Mahalanobis distance, on the other hand, adjusts distance for covariance between the variables. The distance is defined as

$$d_{\mathrm{M}}(\mathbf{w}_i, \mathbf{w}_{i'}) = (\mathbf{w}_i - \mathbf{w}_{i'})^T \boldsymbol{\Sigma}^{-1} (\mathbf{w}_i - \mathbf{w}_{i'}),$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the data matrix $\mathbf{W}$, and its inverse is taken as a weight to the data matrix. The selection of distance measure may not be trivial. It mostly depends on the distribution of data, the research purpose, and what types of interpretations are desired. With the distance measure being determined, cluster analysis is then performed by implementing the chosen algorithm. Next, we would like to introduce two common clustering techniques, K-means and HACA.

3.2. K-Means for Cognitive Diagnosis

K-means cluster analysis is a widely used exploratory procedure to group subjects according to a vector of data. The key idea of the K-means algorithm is to estimate the cluster centers based on the data with the number of clusters being predetermined. Note that the "K" in "K-means" does not refer to the number of attributes in the cognitive diagnosis model, in which the same letter has become the common notation. Once the centers are decided, data are sent to the closest cluster. Specifically, consider an $N \times K$ data matrix for N subjects and K variables, with rows $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_N$. Suppose the aim is to group the N subjects into M clusters based on the observations of the K variables. In our particular application $M = 2^K$, where K refers to the number of attributes to be diagnosed. However, as will be demonstrated in a data analysis that follows, one need not always fit this many clusters, because some combinations of attributes might be very infrequent in the population.

Taking Euclidean distance as an instance, the K-means method assigns data point $\mathbf{w}_i$ to the mth cluster if

$$m = \arg \min_{u \in \{1,\ldots,M\}} \|\mathbf{w_i} - \hat{\mathbf{c}}_\mathbf{u}\|^2, \tag{8}$$

where $\hat{\mathbf{c}}_\mathbf{u}$ is the estimated center of the uth cluster and is obtained by taking an average of the observations within the cluster. By searching for the centers and corresponding classifications that minimize the sum of the squared distances over all n subjects, the global K-means solution is found. The procedure is carried out by the following iterative steps:

1. Select M initial K-dimensional cluster centers.

2. Assign data points to clusters by the criterion (8).
3. Reset the cluster centers by calculating the average of assigned observations.
4. Repeat 2 and 3 until no observation can be relocated.

3.2.1. Selecting Initial Values for K-Means In our research, the greatest challenge in obtaining reliable results is deriving sufficient starting values. Having poor starting values can result in convergence to local optima (Steinley, 2003), and can yield solutions that are much poorer than the global solution. Many methods of initializing starting values for the K-means algorithm have been proposed. Some suggest incorporating other clustering methods to provide better initial values for K-means. For instance, one could first run HACA, which doesn't need starting values, and then average the $\boldsymbol{w}$'s in clusters obtained from HACA to initialize cluster centers for K-means. MacQueen (1967) suggested selecting the first M data points as the initial starting values. The shortcoming of this method occurs when data are ordered in some particular way. For example, for some achievement tests, data might be rearranged according to the order of scores. In this case, it is very likely that the first several data points actually belong to the same cluster. This will lead to inappropriate initializations. A modification of this method is to, instead of selecting the first M data points, randomly select M data points over the data set. Also, regarded as an adjusted method of random selection, Forgy (1965) proposed an initialization method of first randomly selecting M data points as seeds, assigning remaining data points to the cluster with the closest seed, computing the means of the clusters, and then taking these means as starting values for the K-means algorithm. Kaufman's and Rousseuw's (1990) method, with the feature of selecting the M starting centers successively according to a criterion function, has appeared in many applications.

An empirical study from Pena, Lozano, and Larranaga (1999) compared four popular initialization methods for K-means: the random, Forgy, Macqueen, and Kaufman methods. The results showed that the random and the Kaufman methods outperformed the other two in terms of effectiveness and independence of initial conditions.

Bradley and Fayyad (1998) estimated starting values several times, say P times, by finding the modes through the K-means procedures with P small random samples drawn from the original complete data set (the starting values for estimating the "working" starting values are drawn randomly from the data). These P estimated starting points are then used to run the K-means with the whole data for P times. The best center solution is selected from the $2P$ centers, including the P starting centers and the P final centers, by which can produce the minimal cluster distortion. For a more complete review of initialization methods, and for a thorough review of K-means in general, see Steinley (2006).

Although many methods for fixing the initial value problem have been developed, there is no absolute criterion to determine which is the best. The decision mostly depends on the experiment setting, features of data structure, and some other considerations.

All of the methods for initialization of cluster centers described above are for general cases, and when using cluster analysis for cognitive diagnosis, more specialized techniques can be used. For instance, solely for the purpose of initializing cluster centers, one of the parametric models from the previous section can be assumed along with assumed parameters when rough guesses are realistic to make, or with fitted parameters if software is available. As an example, one could assume the data arose from a DINA model with known values of the s and g parameters. One could set $s = 0.1$ and $g = 0.2$ for all items, say on a multiple-choice test with 5 options per item. Then by a very quick Monte Carlo simulation, these parameters together with $\boldsymbol{Q}$ easily determine the expected value of $\boldsymbol{W}$ for each attribute vector $\boldsymbol{\alpha}$, which become the initial cluster centers.

Interestingly, K-means can also be used as a method for finding a starting value for cognitive diagnosis modeling. Willse, Henson, and Templin (2007) used K-means to consider each separate component of the vector $\boldsymbol{W}$ and fitted two-cluster solutions. These were then combined

after K different runs to arrive at the 2^K clusters corresponding to the different values of $\boldsymbol{\alpha}$. This method showed promising results in simulation, but as will be seen in the next section, proceeding one attribute at a time can lead to inconsistent classification. Nevertheless, this method is quite practical, and is used effectively by Willse et al. (2007), for setting an initial value of $\boldsymbol{\alpha}$ for each subject in a clever model-free and iterative likelihood-based procedure for classification that follows the cluster analysis to refine the solution.

3.3. Hierarchical Agglomerative Cluster Analysis

In contrast to partitioning data into exclusive clusters like K-means does, hierarchical clustering forms a dynamic tree structure of clustering in which not only are the distances between data points taken into account, but also the distances between clusters are considered. Two common types of hierarchical clustering are the agglomerative method and the divisive method, which are different in the directions of setting up the clustering system. Agglomerative clustering constructs the tree system from the "leaves," meaning the clustering starts from assigning one cluster to each of the subjects and ends in collecting all subjects into one single cluster. The divisive method, on the other hand, starts from the "root." Here, we put more emphasis on the hierarchical agglomerative cluster analysis (HACA).

Compared with K-means, HACA is much simpler computationally, and does not require selecting initial values. HACA begins by defining a matrix of distances for all pairs of distinct observations, say $d_{ii'} = \sqrt{\sum_{k=1}^{K}(w_{ik} - w_{i'k})^2}$ in the case of Euclidean distance. Then each object, or subject in this case, begins as its own cluster. Defining the distance between two clusters C_l and $C_{l'}$ as $d^*_{ll'}$, the next step is to cluster the two objects i and i' for which $d^*_{ll'} = d_{ii'}$ is smallest. At each step thereafter, two clusters are joined to achieve the minimum distance by adjoining two of the existing clusters, and the cluster distances are updated after each merger. Defining these distances between clusters is what distinguishes different methods of linking clusters, which comprise the variations of HACA. Clusters are combined by using one of the following linkages to minimize the distance between the clusters that are joined in each step until the process is stopped at a fixed number of clusters or until only one cluster containing all of the objects remain. In our application, the process is stopped at the point where there are 2^K clusters.

3.3.1. Common Linkages in HACA We start with complete linkage, which will be taken to demonstrate the theorem in the next section. In the case of complete linkage, the distance between cluster C_l and $C_{l'}$, $d^*_{ll'}$ is just the maximum distance between two points, one from cluster C_l and one from $C_{l'}$,

$$d^*_{ll'} = \max_{i \in C_l, i' \in C_{l'}} d_{ii'}. \tag{9}$$

The above definition implies that every data point in the combined cluster would not be farther than $d^*_{ll'}$ away from every other data point in the cluster. Complete linkage clustering tends to produce homogeneous, but not necessarily separate, clusters.

Single linkage, on the other hand, defines distance according to the minimum distance resulting from taking a point from each cluster,

$$d^*_{ll'} = \min_{i \in C_l, i' \in C_{l'}} d_{ii'}. \tag{10}$$

The single linkage is known as very "myopic" (Lattin, Carroll, & Green, 2003), which means as long as a data point is close to any one of the other data points, it will be included into a cluster, no matter how far away it is from all the others in that cluster. Consequently, single linkage tends to produce long, stringy clusters and non-convex shapes, which is known as the chaining effect.

Instead of taking the two extreme distances into consideration, average linkage clustering uses the mean of distances between the data points in two different clusters as a measure. Specifically, the distance between two clusters is defined as

$$d^*_{ll'} = \frac{\sum_{\boldsymbol{w}_i \in C_l} \sum_{\boldsymbol{w}_{i'} \in C_{l'}} d_{ii'}}{N_l \times N_{l'}},$$

where N_l and $N_{l'}$ represent the numbers of data in clusters C_l and $C_{l'}$, respectively. This method has the tendency to produce ball-shaped clusters.

In addition to averaging distance of all possible pairs of the data points between two clusters as the average linkage clustering does, an alternative to calculate averaged distance is the centroid linkage, in which the data points in each cluster are taken on average first, and then the distance between two clusters is defined as the distance between the two centroids, as indicated in the following:

$$d^*_{ll'} = d_{\overline{\boldsymbol{w}}_l, \overline{\boldsymbol{w}}_{l'}}, \tag{11}$$

where $\overline{\boldsymbol{w}}_l$ is the centroid (average) of cluster C_l and the updated centroid is given by $\overline{\boldsymbol{w}}_{ll'} = (N_l \overline{\boldsymbol{w}}_l + N_{l'} \overline{\boldsymbol{w}}_{l'})/(N_l + N_{l'})$. The centroid method is known to be robust, but generally is outperformed by average linkage clustering.

When running the hierarchical agglomerative clustering algorithm, the above methods vary only in the ways they define distance. Every step remains the same except for replacing different methods for determining cluster distance. Ward's linkage (Ward, 1963), is a general hierarchical clustering method in which clusters are chosen to merge so that the updated within cluster sum of square errors are minimized. More specifically, instead of taking distances between clusters into account, Ward's method calculates *SSE*s, in which the *SSE* for the lth cluster is defined in the following:

$$SSE_l = \sum_{i=1}^{N_l} (\boldsymbol{w}_{li} - \overline{\boldsymbol{w}}_l)^T (\boldsymbol{w}_{li} - \overline{\boldsymbol{w}}_l). \tag{12}$$

At each clustering level, all possible pairs of clusters are considered and the pair which results in minimal overall *SSE* is selected to form a new cluster. The lth and the l'th clusters are merged at some clustering level if

$$SSE_{ll'} - (SSE_l + SSE_{l'}) \tag{13}$$

is the minimum among all pairs, where $SSE_{ll'}$ is the pooled *SSE* for cluster C_l and $C_{l'}$ with the updated center $\overline{\boldsymbol{w}}_{ll'} = (N_l \overline{\boldsymbol{w}}_l + N_{l'} \overline{\boldsymbol{w}}_{l'})/(N_l + N_{l'})$. In addition, it can be shown that Ward's linkage is related to centroid linkage by applying some simple algebra. If we insert (12) to (13), the latter equation becomes

$$\frac{N_l N_{l'}}{N_l + N_{l'}} (\overline{\boldsymbol{w}}_l - \overline{\boldsymbol{w}}_{l'})^T (\overline{\boldsymbol{w}}_l - \overline{\boldsymbol{w}}_{l'}). \tag{14}$$

By taking squared Euclidean distance as the similarity measure for (11), the centroid linkage and Ward's linkage differ only in the multiplier $N_l N_{l'}/(N_l + N_{l'})$ showing in (14). This relation implies that cluster size has an impact on Ward's linkage, but not on centroid linkage. It furthermore explains why Ward's method tends to produce nearly equal sized clusters which are convex and compact, and thus suffers from outliers (Milligan, 1980).

Punj and Stewart (1983) had a thorough review on applications of cluster analysis. In general, K-means outperformed HACA with starting values based on a priori knowledge, but did disappointingly with random starts (Milligan, 1980). Also, distance measure selection did not

seem to be critical as one might expect among a reasonable set of choices. The methods performed robustly across distance measures, and K-means was particularly less affected (Milligan, 1980). Among the selected linkages for HACA, findings indicated that the selection of which particular linkage method to be used mostly depends on the structure of the data (Cunnningham & Ogilvie, 1972; Everitt, Landau, & Leese, 2001). In some specific empirical investigations, it was found that Ward's linkage performs best with clusters of about equal sizes, given data generated from some particular distributions, and centroid and average linkages outperform others when the cluster sizes are unequal (Kuiper & Fisher, 1975; Blashfield, 1976; Hands & Everitt, 1987). However, the poor performance of single linkage makes it less considered in practice, though it seems to have the potential of being useful if its chaining property can be overcome (Everitt et al., 2001).

The two linkages we consider on mathematical grounds are complete linkage and single linkage. HACA, whether done with complete linkage or single linkage, turns out to be quite straightforward to study theoretically, as will be seen in the next section. However, the less theoretically tractable K-means appears to perform slightly better in simulations. One advantage of HACA for this application is that it is not so easily degraded when the 2^K population proportions for the different values of $\boldsymbol{\alpha}$ in the underlying latent class model vary greatly. K-means can suffer in that situation because the fitting criterion does not improve by wasting a cluster center on a cluster with a very low frequency.

4. Asymptotic Theory for Classification

In this section, we discuss some of the theory for classification using cluster analysis. Some results will be independent of the underlying latent class model generating responses, but other results need to be studied on a model-by-model basis. The DINA model is assumed for some of the main results that follow, though the only critical condition needed to generalize the consistency theorem to other models is that the mean of the vector $\boldsymbol{W}$ is distinct for different values of the binary latent vector $\boldsymbol{\alpha}$. In particular, let $\boldsymbol{T}(\boldsymbol{\alpha}) = E[\boldsymbol{W} \mid \boldsymbol{\alpha}]$ be the K-dimensional mean vector for the K sum-scores, which depends on each of the 2^K attribute patterns $\boldsymbol{\alpha}$. To show that $\boldsymbol{W}$ is an appropriate statistic to be utilized by either HACA or K-means, we first have to show that given two different attribute patterns $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$, the corresponding expectations of $\boldsymbol{T}(\boldsymbol{\alpha})$ and $\boldsymbol{T}(\boldsymbol{\alpha}^*)$ will be different. This will imply that with high probability when using the input $\boldsymbol{W}$, cluster analysis will group subjects correctly as the number of items becomes large.

The first step in establishing the separation of these expected values is to study what conditions must be met to identify attribute patterns when there is no stochastic component to the model. We begin by considering the ideal response patterns that different values of $\boldsymbol{\alpha}$ produce.

Definition 1. Call $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_J)'$ an ideal vector where $\eta_j = \prod_{k=1}^{K} \alpha_k^{q_{jk}}$, and let $\boldsymbol{e}_k$ be a $K \times 1$ vector with the kth entry being 1 and all other entries 0. A matrix $\boldsymbol{Q}$ is *complete* if it can identify all possible attribute patterns; that is, $\boldsymbol{\eta}(\boldsymbol{\alpha}) = \boldsymbol{\eta}(\boldsymbol{\alpha}^*)$ implies $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$.

Completeness refers to the ability of an exam to determine attribute patterns from one another, in a purely algebraic sense. Next, we consider two lemmas that are needed in a proof of the main consistency theorem to follow. The first concerns necessary and sufficient conditions for $\boldsymbol{Q}$ be complete. Completeness is generally needed for the identifiability of a model, which must be satisfied for correct classification. The first lemma implies that an exam must include items to measure each attribute alone, among its many items, if identifiability of attribute patterns is to be achieved.

Lemma 1. *A $J \times K$ matrix $\boldsymbol{Q}$ is complete if and only if it includes rows $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_K$, among its $K \leq J$ rows.*

Proof: First assume that some $\boldsymbol{e}_k$ is missing from $\mathbf{Q}$. For attribute patterns $\boldsymbol{\alpha} = (0, 0, \ldots, 0)'$ and $\boldsymbol{\alpha}^* = \boldsymbol{e}_k$,

$$\eta_j(\boldsymbol{\alpha}) = \eta_j(\boldsymbol{\alpha}^*) = \begin{cases} 1 & \text{if } \mathbf{q}_j = (0, 0, \ldots, 0), \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $\mathbf{Q}$ is not complete. Next, assume that $\mathbf{Q}$ has rows $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_K$, among its J rows. Reorder $\mathbf{Q}$ by moving the K items corresponding to these K vectors to the first K places. Then for a particular $\boldsymbol{\alpha}$, the first K entries of $\boldsymbol{\eta}(\boldsymbol{\alpha})$, denoted by $\boldsymbol{\eta}_{1:K}(\boldsymbol{\alpha})$, are identical to $\boldsymbol{\alpha}$ itself. Specifically, $\boldsymbol{\eta}_{1:K}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}$, for all $\boldsymbol{\alpha}$. If $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$, then $\boldsymbol{\eta}_{1:K}(\boldsymbol{\alpha}) \neq \boldsymbol{\eta}_{1:K}(\boldsymbol{\alpha}^*)$. Now considering all items in $\mathbf{Q}$, if $\boldsymbol{\eta}_{1:K}(\boldsymbol{\alpha}) \neq \boldsymbol{\eta}_{1:K}(\boldsymbol{\alpha}^*)$, then $\boldsymbol{\eta}(\boldsymbol{\alpha}) \neq \boldsymbol{\eta}(\boldsymbol{\alpha}^*)$, no matter if $\boldsymbol{\eta}_{(K+1):J}(\boldsymbol{\alpha})$ and $\boldsymbol{\eta}_{(K+1):J}(\boldsymbol{\alpha}^*)$ are identical or not. Therefore, if $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$, then consequently, $\boldsymbol{\eta}(\boldsymbol{\alpha}) \neq \boldsymbol{\eta}(\boldsymbol{\alpha}^*)$, and $\mathbf{Q}$ is complete. □

The next lemma assumes that data are generated according to the DINA model, and makes use of the previous lemma to show that the expectation of $\boldsymbol{W}$ is distinct for each of the 2^K attribute patterns.

Lemma 2. *Let* $\mathbf{W}$ *be defined as in* (6)*, and assume that responses are generated according to the DINA model with item response function given by* (3)*. For all items of the DINA model, we assume $0 \leq g_j < 1 - s_j \leq 1$ for $j = 1, 2, \ldots, J$. Also, assume that $\boldsymbol{Q}$ is complete. For attribute patterns $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$, if $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$, then $\mathbf{T}(\boldsymbol{\alpha}) \neq \mathbf{T}(\boldsymbol{\alpha}^*)$.*

Proof: The item response function of the DINA model is

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

Thus the corresponding expectation of a response is,

$$E[Y_{ij} \mid \boldsymbol{\alpha}_i] = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}, \tag{15}$$

where η_{ij} is defined by (4). Assume $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$. Then they must be different in some kth entry, say $\alpha_k = 1$ and $\alpha_k^* = 0$. In addition, since only nonzero entries of $\mathbf{Q}$ contribute to the components of $\mathbf{T}$, we define $\mathbf{B}_k = \{j \mid q_{jk} = 1\}$ to indicate items requiring the kth skill. For $j \in \mathbf{B}_k$, $\alpha_k^{q_{jk}} = 1$ and $(\alpha_k^*)^{q_{jk}} = 0$, so the ideal responses for $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$ are as follows:

$$\eta_j(\boldsymbol{\alpha}) = \begin{cases} 1 & \text{if all required skills are possessed,} \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

and

$$\eta_j(\boldsymbol{\alpha}^*) = 0.$$

From Lemma 1, we know that items of $\boldsymbol{e}_k$ must be included in set $\mathbf{B}_k$. This ensures that the subset for making $\eta_j(\boldsymbol{\alpha}) = 1$ in (16) is nonempty. We therefore partition $\mathbf{B}_k$ into $\mathbf{B}_{e_k}$ and $\mathbf{B}_{k-e_k}$, where $\mathbf{B}_{e_k} = \{j \mid \mathbf{q}_j = \boldsymbol{e}_k\}$ and $\mathbf{B}_{k-e_k} = \mathbf{B}_k \backslash \mathbf{B}_{e_k}$. Under this setting, the kth components of the expected

sum-scores vectors, $T_k(\boldsymbol{\alpha})$ and $T_k(\boldsymbol{\alpha}^*)$, can be expressed as

$$\begin{aligned} T_k(\boldsymbol{\alpha}) &= \sum_{j=1}^{J} E[Y_j \mid \boldsymbol{\alpha}] q_{jk} \\ &= \sum_{j \in \mathbf{B}_{e_k}} (1 - s_j) + \sum_{j \in \mathbf{B}_{k-e_k}} (1 - s_j)^{\eta_j(\boldsymbol{\alpha})} g_j^{(1-\eta_j(\boldsymbol{\alpha}))}, \end{aligned} \tag{17}$$

and

$$\begin{aligned} T_k\big(\boldsymbol{\alpha}^*\big) &= \sum_{j=1}^{J} E\big[Y_j \mid \boldsymbol{\alpha}^*\big] q_{jk} \\ &= \sum_{j \in \mathbf{B}_{e_k}} g_j + \sum_{j \in \mathbf{B}_{k-e_k}} g_j. \end{aligned} \tag{18}$$

With the assumption that $1 - s_j > g_j$ for all j, we know that $\sum_{j \in \mathbf{B}_{e_k}} (1 - s_j) > \sum_{j \in \mathbf{B}_{e_k}} g_j$ and $\sum_{j \in \mathbf{B}_{k-e_k}} (1 - s_j)^{\eta_j(\boldsymbol{\alpha})} g_j^{(1-\eta_j(\boldsymbol{\alpha}))} \geq \sum_{j \in \mathbf{B}_{k-e_k}} g_j$. These will lead to the result that $T_k(\boldsymbol{\alpha}) > T_k(\boldsymbol{\alpha}^*)$, which implies that $\mathbf{T}(\boldsymbol{\alpha}) \neq \mathbf{T}(\boldsymbol{\alpha}^*)$. □

The final lemma relates to the behavior of HACA when all of the observations are close to their expected values, which is what takes place with high probability for long exams when the data are summarized by the vector $\boldsymbol{W}/J$. We focus on HACA here because the same useful property is not quite so clear with K-means. In the case of K-means, population proportions for the different $\boldsymbol{\alpha}$ patterns can be manipulated so that tightness of observations around their expected values needs to be carefully balanced with these proportions to obtain the same result. Large sample theory for K-means has been worked out in some general cases, showing that cluster centers converge with certain rates (Hartigan, 1978; Pollard, 1981, 1982). However, these centers need not be the expected values for different latent classes, which is the interest here.

Lemma 3. *Let* $\mathbf{V}$ *be a random vector in K-dimensional space, with a mixture probability density function* $f(\boldsymbol{v}) = \sum_{m=1}^{M} f_m(\boldsymbol{v})\zeta_m$, *where* ζ_m *denotes the population proportion for the* m*th latent class, and* f_m *is a probability density function in* K*-dimensional Euclidean space with expected value* $\boldsymbol{\mu}_m$. *Assume that for some positive number* δ, $\min_{m \neq m'} \|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\| > \delta$. *Consider data* $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N$, *and let* $\boldsymbol{v}_i^{(m)}$ *denote that the* i*th observation arose from the* m*th component of the mixture density* f. *If there exists a small enough* $\varepsilon > 0$ *such that* $\max_{i=1,2,\ldots,N} \|\boldsymbol{v}_i^{(m)} - \boldsymbol{\mu}_m\| < \varepsilon$ *for all* $i = 1, \ldots, N$, *then the hierarchical agglomerative cluster analysis solution with either complete or single linkage, will place objects in clusters corresponding exactly with their latent class membership when the algorithm is cut at* M *clusters.*

Proof: The proof essentially follows from the definition of HACA. At any step the distance between clusters, known as the fusion coefficient, that are combined is kept as small as possible. Suppose $\boldsymbol{v}_i$ and $\boldsymbol{v}_{i'}$ arise from a common latent class m with corresponding mean $\boldsymbol{\mu}_m$. Then

$$\|\boldsymbol{v}_i - \boldsymbol{v}_{i'}\| \leq \|\boldsymbol{v}_i - \boldsymbol{\mu}_m\| + \|\boldsymbol{v}_{i'} - \boldsymbol{\mu}_m\| < 2\varepsilon.$$

Additionally, consider a third data point, $\boldsymbol{v}_{i^*}$, from a different cluster m^*. By selecting ε such that $0 < \varepsilon < \delta/4$, we know that

$$\begin{aligned} \|\boldsymbol{v}_i - \boldsymbol{v}_{i^*}\| &> \|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m^*}\| - 2\varepsilon \\ &> \delta - 2\varepsilon \\ &> 2\varepsilon. \end{aligned} \tag{19}$$

This means that adjoining clusters corresponding to the same latent class will always result in fusion coefficients less than 2ε and joining clusters containing members of different classes would result in fusion coefficients greater than 2ε, no matter whether distance between clusters is defined by complete linkage or single linkage. Thus, to minimize fusion coefficients, the algorithm must proceed by grouping clusters containing only data from the same latent class until there are just M clusters. □

These preceding lemmas are now used in a proof of the consistency of clustering, along with a corollary to show that the theorem holds in the case of the DINA model. Before proceeding, we define an *exact* cluster solution for data arising from a mixture model as one for which clusters correspond precisely with the components of the mixture, or synonymously the latent classes, from which the objects being clustered originated. For example, the previous lemma shows conditions under which an M-cluster HACA solution will be exact for a finite mixture model.

Theorem 1. *Let* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N$ *be item response vectors and* $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})$. *Define* $\boldsymbol{W}_i = \sum_{j=1}^{J} Y_{ij}\mathbf{q}_j$ *and* $\boldsymbol{V}_i = \boldsymbol{W}_i/J$, *where* $\mathbf{q}_j$ *is the* j*th row of a complete matrix* $\mathbf{Q}$ *with* J *rows and* K *columns. Assume that responses arise from a cognitive diagnosis model in which responses are conditionally independent given a* K*-dimensional vector with binary components* $\boldsymbol{\alpha}$, *and each of the* 2^K *values of* $\boldsymbol{\alpha}$ *is sampled with a probability greater than* 0. *Also, define* $E[\boldsymbol{V}_i \mid \boldsymbol{\alpha}^{(m)}] = \boldsymbol{\mu}_m$ *for* $m = 1, 2, \ldots, 2^K$, *and assume that for some positive number* δ, $\min_{m \neq m'} \|\boldsymbol{\mu}_m - \boldsymbol{\mu}_{m'}\| > \delta$. *Provided* $Ne^{-J} \to 0$ *as* $J \to \infty$, *the hierarchical agglomerative cluster analysis solution with either single linkage or complete linkage using* $\boldsymbol{V}$ *as input will be exact with probability converging to* 1. *Further, if* Ne^{-J} *is summable, the solution is inexact only finitely often with probability* 1.

Proof: Because the theorem assumes the mean vectors for the 2^K latent classes are separated by some constant δ it suffices to show that for any $\varepsilon > 0$, the probability that all observations are within ε of their expected values converges to 1. Then Lemma 3 may be invoked to obtain the result. First note that the statistic $\boldsymbol{V}$ is just a scaled version of $\boldsymbol{W}$, with an interpretation much like proportion correct, aside from the fact that we divide each component by J rather than by the number of items that require the attribute, which is certainly a reasonable alternative. Because $\boldsymbol{V}$ only amounts to rescaling, it yields the same cluster solution as $\boldsymbol{W}$ whether using K-means or HACA.

If we number the possible attribute patterns from $m = 1, 2, 3, \ldots, 2^K$, let $m(i)$ index the attribute pattern for the ith subjects, so that $\boldsymbol{V}_i$ has expected value $\boldsymbol{\mu}_{m(i)}$. Because $\|\boldsymbol{V}_i - \boldsymbol{\mu}_{m(i)}\| = \sqrt{\sum_{k=1}^{K}(V_{ik} - \mu_{m(i)k})^2} \leq \sum_{k=1}^{K} |V_{ik} - \mu_{m(i)k}|$, for $\varepsilon > 0$, if $\sum_{k=1}^{K} |V_{ik} - \mu_{m(i)k}| \geq \varepsilon$, there ex-

ists at least some k such that $|V_{ik} - \mu_{m(i)k}| \geq \varepsilon/K$. Therefore,

$$\begin{aligned} P\big[\|\boldsymbol{V}_i - \boldsymbol{\mu}_{m(i)}\| > \varepsilon\big] &\leq P\left[\sum_{k=1}^{K} |V_{ik} - \mu_{m(i)k}| \geq \varepsilon\right] \\ &= P\left[\sum_{k=1}^{K} \left|\frac{W_{ik}}{J} - \frac{T_k(\boldsymbol{\alpha}^{(m(i))})}{J}\right| \geq \varepsilon\right] \\ &= P\left[\sum_{k=1}^{K} \left|W_{ik} - T_k\big(\boldsymbol{\alpha}^{(m(i))}\big)\right| \geq J\varepsilon\right] \\ &\leq P\left[\bigcup_{\forall k}\left\{\left|W_{ik} - T_k\big(\boldsymbol{\alpha}^{(m(i))}\big)\right| \geq \frac{J\varepsilon}{K}\right\}\right] \\ &\leq \sum_{k=1}^{K} P\left[\left|W_{ik} - T_k\big(\boldsymbol{\alpha}^{(m(i))}\big)\right| \geq \frac{J\varepsilon}{K}\right]. \end{aligned} \tag{20}$$

Thus, the problem has been reduced to studying deviations of the components of $\boldsymbol{W}$ from their expected values. Due to conditional independence and the boundedness of the item response variables, the inequality above can be directly bounded by a result of Hoeffding (1963), and (20) can be bounded by

$$\begin{aligned} \sum_{k=1}^{K} P\left[\left|W_{ik} - T_k\big(\boldsymbol{\alpha}^{(m(i))}\big)\right| \geq \frac{J\varepsilon}{K}\right] &\leq \sum_{k=1}^{K} 2e^{-2J(\varepsilon/K)^2} \\ &= 2Ke^{-2J(\varepsilon/K)^2} \\ &< 2Ke^{-J\varepsilon^2/K^2}. \end{aligned} \tag{21}$$

Equations (20) and (21) lead to the result that for $i \in \{1, 2, \ldots, N\}$,

$$P\big[\|\boldsymbol{V}_i - \boldsymbol{\mu}_{m(i)}\| > \varepsilon\big] \leq 2Ke^{-J\varepsilon^2/K^2}. \tag{22}$$

Furthermore, it can be shown that

$$\begin{aligned} P\left[\max_i \|\boldsymbol{V}_i - \boldsymbol{\mu}_{m(i)}\| > \varepsilon\right] &= 1 - P\left[\max_i \|\boldsymbol{V}_i - \boldsymbol{\mu}_m\| \leq \varepsilon\right] \\ &= 1 - P\left[\bigcap_i \{\|\boldsymbol{V}_i - \boldsymbol{\mu}_{m(i)}\| \leq \varepsilon\}\right] \\ &= P\left[\bigcup_i \{\|\boldsymbol{V}_i - \boldsymbol{\mu}_{m(i)}\| > \varepsilon\}\right] \\ &\leq 2NKe^{-J\varepsilon^2/K^2}. \end{aligned} \tag{23}$$

If we set $N < e^{J\varepsilon^2/(2K^2)}$, then (23) is followed by

$$\begin{aligned} P\left[\max_i \left\|\boldsymbol{V}_i - \boldsymbol{\mu}_{m(i)}^{(V)}\right\| > \varepsilon\right] &\leq 2NKe^{-J\varepsilon^2/K^2} \\ &< 2Ke^{-J\varepsilon^2/(2K^2)}. \end{aligned} \tag{24}$$

Because K and ϵ are constants, this probability converges to 0 provided sample size and test length have the relationship $Ne^{-J} \to 0$ as $J \to \infty$, so that the probability that the cluster solution is exact converges to 1. This is a long test theory, though sample size is allowed to grow with J. If it grows slow enough so that Ne^{-J} is summable as N and J go to ∞, the Borel–Cantelli lemma gives the even stronger result that classification is inexact only finitely often with probability 1. □

Note that the conditions of this theorem are general and do not specify a particular cognitive diagnosis model. However, if the DINA model holds and we have a complete $\boldsymbol{Q}$, the condition on the never vanishing proportion of $\boldsymbol{e}_k$'s in $\boldsymbol{Q}$ together with Lemma 2 suffice to guarantee the assumption that the mean vectors of $\boldsymbol{V}$ are separated by some δ, and the consistency theorem holds for the DINA model.

Next, we provide some examples showing the importance of performing cluster analysis based on the entire vector $\boldsymbol{W}$ rather than just diagnose attribute mastery one component at a time, using the corresponding component of $\boldsymbol{W}$. The previous theorem showed that the critical element of classification is to have well-separated within-cluster expected values. However, when fitting two-clusters for a particular attribute, the remaining components of $\boldsymbol{\alpha}$ together with $\boldsymbol{Q}$ can alter the result. For example, assume that the number of attributes is $K = 2$, and the 4 attribute patterns are $\boldsymbol{\alpha}_1 = (0, 0)$, $\boldsymbol{\alpha}_2 = (1, 0)$, $\boldsymbol{\alpha}_3 = (0, 1)$, and $\boldsymbol{\alpha}_4 = (1, 1)$. For the DINA model, the expectations of $\boldsymbol{W}$ over all possible attribute patterns are as follows:

$$\mathbf{T}(\boldsymbol{\alpha}_1) = \left(\sum_{j=1}^{J} g_j q_{j1}, \sum_{j=1}^{J} g_j q_{j2} \right),$$

$$\mathbf{T}(\boldsymbol{\alpha}_2) = \left(\sum_{j=1}^{J} (1 - s_j) q_{j1} (1 - q_{j2}) + \sum_{j=1}^{J} g_j q_{j1} q_{j2}, \sum_{j=1}^{J} g_j q_{j2} \right),$$

$$\mathbf{T}(\boldsymbol{\alpha}_3) = \left(\sum_{j=1}^{J} g_j q_{j1}, \sum_{j=1}^{J} (1 - s_j) q_{j2} (1 - q_{j1}) + \sum_{j=1}^{J} g_j q_{j1} q_{j2} \right),$$

$$\mathbf{T}(\boldsymbol{\alpha}_4) = \left(\sum_{j=1}^{J} (1 - s_j) q_{j1}, \sum_{j=1}^{J} (1 - s_j) q_{j2} \right).$$

If we only focus on the first attribute, it turns out there are 3 different values:

$$T_1(\boldsymbol{\alpha}_1) = T_1(\boldsymbol{\alpha}_3) = \sum_{j=1}^{J} g_j q_{j1},$$

$$T_1(\boldsymbol{\alpha}_2) = \sum_{j=1}^{J} (1 - s_j) q_{j1} (1 - q_{j2}) + \sum_{j=1}^{J} g_j q_{j1} q_{j2},$$

$$T_1(\boldsymbol{\alpha}_4) = \sum_{j=1}^{J} (1 - s_j) q_{j1}.$$

With the assumption of $1 - s_j > g_j$, $T_1(\boldsymbol{\alpha}_1) = T_1(\boldsymbol{\alpha}_3) < T_1(\boldsymbol{\alpha}_2) < T_1(\boldsymbol{\alpha}_4)$. Assume that $P[\boldsymbol{\alpha} = \boldsymbol{\alpha}_1] = P[\boldsymbol{\alpha} = \boldsymbol{\alpha}_2] = P[\boldsymbol{\alpha} = \boldsymbol{\alpha}_3] = P[\boldsymbol{\alpha} = \boldsymbol{\alpha}_4] = 1/4$. Then if we duplicate $\boldsymbol{Q}$ many times to construct a longer and longer exam, performing cluster analysis on $\boldsymbol{W}$ becomes equivalent to

TABLE 1.
K-means classification for the first dimension of $\boldsymbol{W}$.

	$T_1(\boldsymbol{\alpha})$			
	9	9	17	81
Cluster	1	1	1	2
Center		11.667		81

doing so on $\boldsymbol{T}$, which was essentially the basis for the proof. As the test grows longer, suppose items with $\boldsymbol{q} = (1, 0)$ and $\boldsymbol{q} = (0, 1)$ each occur 10% of the time, and items with $\boldsymbol{q} = (1, 1)$ comprise 80% of the items. Also, assume $s_j = g_j = 0.1$ for all $j = 1, \ldots, J$. Then when $J = 100$, for example, the expectations of $\boldsymbol{W}$ on the first dimension become:

$$T_1(\boldsymbol{\alpha}_1) = T_1(\boldsymbol{\alpha}_3) = 9,$$
$$T_1(\boldsymbol{\alpha}_2) = 17,$$
$$T_1(\boldsymbol{\alpha}_4) = 81.$$

These 4 expectations are then grouped by K-means with the number of groups being 2. The result is in Table 1.

In addition, when HACA with complete linkage is applied to the same set of expectations, the same incorrect clustering as when using K-means is returned. Table 1 shows that the cluster with centers $T_1((0, 0))$, $T_1((1, 0))$, and $T_1((0, 1))$ formed a group, and $T_1((1, 1))$ formed the other one, in which $T_1((1, 0))$ was expected to be grouped with $T_1((1, 1))$. This implies that even when $\boldsymbol{\alpha}$ is uniformly distributed over its sample space, but the possible choice of elements in $\boldsymbol{Q}$ are imbalanced in some particular way, classification only based on a single attribute can be inconsistent, no matter whether K-means or HACA is used.

Analyzing sum-scores one at a time can be even more problematic with compensatory models. Consider the following example in which the expected sum-score for the first attribute is actually higher for a subject with $\boldsymbol{\alpha} = (1, 0)$ than a subject with $\boldsymbol{\alpha}^* = (0, 1)$. Recall the GDM, which was introduced as a compensatory model that features linking an additive function of the attribute indicators to the probability of a correct response. The conditional probability of the compensatory GDM is as follows:

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \frac{\exp[\beta_{1j} + \sum_{k=1}^{K} \gamma_{jk} q_{jk} \alpha_{ik}]}{1 + \exp[\beta_{1j} + \sum_{k=1}^{K} \gamma_{jk} q_{jk} \alpha_{ik}]}.$$

Suppose we know $\boldsymbol{\alpha} = (1, 0)$ and $\boldsymbol{\alpha}^* = (0, 1)$, and also a Q-matrix

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

Then we have

$$\mathbf{T}(\boldsymbol{\alpha}) = \left(\frac{\exp(\beta_{11} + \gamma_{11})}{1 + \exp(\beta_{11} + \gamma_{11})}, \frac{\exp(\beta_{12})}{1 + \exp(\beta_{12})}, \frac{\exp(\beta_{13} + \gamma_{31})}{1 + \exp(\beta_{13} + \gamma_{31})} \right) \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$$

and

$$\mathbf{T}(\boldsymbol{\alpha}^*) = \left(\frac{\exp(\beta_{11})}{1+\exp(\beta_{11})}, \frac{\exp(\beta_{12}+\gamma_{22})}{1+\exp(\beta_{12}+\gamma_{22})}, \frac{\exp(\beta_{13}+\gamma_{32})}{1+\exp(\beta_{13}+\gamma_{32})} \right) \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Let $\beta_{11} = 1$, $\gamma_{11} = 0.1$, $\beta_{13} = 0.2$, $\gamma_{31} = 0.2$ and $\gamma_{32} = 2$, $T_1(\boldsymbol{\alpha}) = 1.35$ and $T_1(\boldsymbol{\alpha}^*) = 1.63$ even though $\alpha_1 = 1$ but $\alpha_1^* = 0$.

5. Simulation of Competing Classification Methods

The aim of this section is to examine through simulation the performances of K-means and HACA with several common linkages in classification, and additionally contrast these with classification using maximum likelihood classification when using the correct model with item parameters obtained by maximum marginal likelihood estimation. Because the K-means solution can depend on the starting value, we used the true expected values of $\boldsymbol{W}$ within each latent class as a best possible case, and used the observed cluster centers obtained by HACA with Ward's linkage as a more realistic case. Because the clusters are unlabeled, defining the "correct" cluster becomes problematic, so results were summarized by measures of within-cluster homogeneity and a measurement of agreement between partitions.

For each condition, 25 data sets were simulated using the DINA model. In each data set, N examinees were drawn from a particular distribution to take a test of J items with K required attributes, where $N = 100$ or 500, $J = 20$, 40, or 80, and $K = 3$ or 4. Conditions were formed by crossing these values of N, J, and K.

The attribute pattern $\boldsymbol{\alpha}$ was generated for each simulated subject in two different ways. First, the distribution of $\boldsymbol{\alpha}$ was assumed to be uniform of the 2^K possible values, so each value was assigned the probability $1/2^K$. A second method was to force a realistic situation in which the attributes were correlated and of unequal prevalence. This is achieved by assuming a multivariate normal distribution underlying the discrete $\boldsymbol{\alpha}$'s. Subjects were drawn from a multivariate normal distribution $MVN(\mathbf{0}_K, \Sigma)$, where the covariance matrix Σ had the structure

$$\begin{pmatrix} 1 & & \rho \\ & \ddots & \\ \rho & & 1 \end{pmatrix}$$

with ρ taking values of 0.25 and 0.5. Assuming that the underlying continuous ability for the ith examinee was $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iK})$, the attribute pattern $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})$ was determined by

$$\alpha_{ik} = \begin{cases} 1 & \text{if } \theta_{ik} \geq \Phi^{-1}(\frac{k}{K+1}), \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

As mentioned previously, to identify all possible attribute patterns, all $\boldsymbol{e}_k$'s have to be included in the Q-matrix. Hence, the Q-matrices for tests of 20 items with $K = 3$ and 4 were designed as in Table 2.

The Q-matrices for tests of 40 items and 80 items were obtained by duplicating this matrix two times and four times, respectively. The item parameters s and g were generated for each item from a Uniform(0, 0.15) distribution for a case that allowed little deviation from ideal responses, and from a Uniform(0, 0.3) distribution in case that allowed for more noise in the data. Across the 25 replications, the classes of examinees and parameters s and g were fixed, whereas the

TABLE 2.
Q-matrices for test of 20 items.

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

responses were sampled from Bernoulli$((1-s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})})$, the item response function of the DINA, in each replication.

To evaluate quality of classifications, correct classification rate is often used as the measure. However, it is applicable only when objects are grouped into labeled sets, which is not the case with cluster analysis. Therefore, instead of applying classification rate, an indicator of agreement between partitions, the Adjusted Rand Index (ARI), and a measure of within-cluster homogeneity were used.

The ARI (Hubert & Arabie, 1985), which originated from the Rand index, is a widely used index for comparing partitions. Assume that $\{g_1, \ldots, g_G\}$ and $\{h_1, \ldots, h_H\}$ are two partitions of N objects $\{o_1, \ldots, o_N\}$. Furthermore, assume that N_{ij} is the number of objects classified into both group g_i and group h_j. It is possible for an object pair to locate in either the same group or different groups under a particular partition. With two partitions involved, there are thus four possible locating relationships between pairs. The Rand index is constructed based on how object pairs are classified into the four categories, in which agreement is defined as the counts of pairs which are placed in the same group and distinct groups under both partitions, and disagreement is the counts of pairs which are placed in the same group under one partition, but in different groups under the other partition. The Rand index is then the ratio of the agreement and the total pair counts, as shown in the following:

$$Rand = \frac{C_2^N + 2\sum_{i=1}^{G}\sum_{j=1}^{H} C_2^{N_{ij}} - \sum_{i=1}^{G} C_2^{N_{i\cdot}} - \sum_{j=1}^{H} C_2^{N_{\cdot j}}}{C_2^N}, \tag{26}$$

where a binomial coefficient $C_2^{(\cdot)}$ is defined as 0 when the number of classified objects is 0 or 1.

The ARI differs from Rand index by incorporating a factor that corrects for chance. With the hypergeometric assumption for the entries of a contingency table, (26) is modified to the

following:

$$ARI = \frac{\sum_{i=1}^{G}\sum_{j=1}^{H} C_2^{N_{ij}} - \sum_{i=1}^{G} C_2^{N_{i\cdot}} \sum_{j=1}^{H} C_2^{N_{\cdot j}} / C_2^{N}}{\frac{1}{2}[\sum_{i=1}^{G} C_2^{N_{i\cdot}} + \sum_{j=1}^{H} C_2^{N_{\cdot j}}] - \sum_{i=1}^{G} C_2^{N_{i\cdot}} \sum_{j=1}^{H} C_2^{N_{\cdot j}} / C_2^{N}},$$

which is also bounded between 0 and 1. In addition to the feature of correcting for chance, the ARI does not require equal numbers of clusters. Though we always use the same number of clusters, 2^K, as there are attribute values, this feature could be valuable in studying cases of misfit where the number of clusters may not correspond with the number of latent classes. In this simulation study, one partition used by the ARI is always the true latent class, and the other is the partition generated by the clustering or modeling procedure.

Our other measure of quality, which will be referred to as ω, assesses the within-cluster homogeneity with respect to the true values of $\boldsymbol{\alpha}$. One application of cognitive diagnosis is to identify groups of subjects that have the same attributes or skills, because this then determines placement or tailored remediation strategies. The index ω measures how similar subjects from the same cluster are to one another, and sums this over the clusters. Unlike the ARI, ω gives some credit for being close, much like a weighted κ index, but without correcting for chance. For instance, assuming that one subject has $\boldsymbol{\alpha} = (1, 1, 0)$ as the true pattern, then more credit should be given for grouping this subject with one having pattern $\boldsymbol{\alpha} = (1, 0, 0)$ than one with $\boldsymbol{\alpha} = (0, 0, 1)$. Due to this concern, ω is used to measure the homogeneity of a cluster taking degree of similarity into account.

$$\omega = 1 - \frac{\sum_{i=2}^{N}\sum_{i'=1}^{i}\sum_{k=1}^{K} |\alpha_{ik} - \alpha_{i'k}| I[\hat{c}_i = \hat{c}_{i'}]}{\sum_{i=2}^{N}\sum_{i'=1}^{i} K \times I[\hat{c}_i = \hat{c}_{i'}]},$$

where $\hat{c}_i$ refers to the estimated class or cluster for the ith of N subjects, and $I[\hat{c}_i = \hat{c}_{i'}]$ indicates whether or not subjects i and i' are placed in the same class or cluster. This index is bounded below by 0 and above by 1, and takes value 1 if the true $\boldsymbol{\alpha}$'s are the same for all pairs of examinees classified together.

5.1. Results

In this simulation, means and standard errors of ARI's and ω's for K-means, HACA with several linkages, and the DINA model were calculated over 25 replications for each condition. The standard errors were controlled with sizes less than 0.03. Due to length limitation, only the means were reported.

This first part was regarded to the conditions where attribute vector $\boldsymbol{\alpha}$ was generated to take each of the 2^K possible values with equal probability. Tables 3 and 4 give means of ARI's and ω's with s and g both being drawn from Uniform(0, 0.15) and Uniform(0, 0.3), respectively.

From the tables, we can see that Ward's linkage performs best among the four linkages, average and complete linkages perform similarly well, but single linkage does poorly. Table 3 shows that for $K = 3$ with s and g smaller than 0.15, K-means with either the best case or Ward's starting values and the DINA-EM perform well across all values of J, whereas HACA with Ward's linkage is not quite as efficient when the test length is short, but can do comparably well when test length is over 40. When $K = 4$, the DINA-EM remained good overall, and K-means and HACA with Ward's linkage required as many as 40 items to reach the standard of ARI > 0.8 and $\omega > 0.9$. In addition, Table 4 indicates that J has to go up at least for one level to obtain similar results to those from Table 3, showing the anticipated result that accurate classification is more difficult when more latent classes are present.

Inspection of the two tables reveals that fitting the correct model is superior in all cases, though the clustering procedures can perform well without assuming a model, provided the test

TABLE 3.
Mean (se) of ARI and ω for DINA data by fitting DINA-EM, K-means and HACA: $s_j \sim$ Uniform$(0, 0.15)$; $g_j \sim$ Uniform$(0, 0.15)$; and $\boldsymbol{\alpha} \sim$ Discrete Uniform.

J	K	Index	$N = 100$						
			DINA-EM	K-means		HACA			
				Best	Ward's	Complete	Single	Average	Ward's
20	3	ARI	0.9355	0.8298	0.7711	0.6760	0.3598	0.6134	0.7600
		ω	0.9760	0.9376	0.9249	0.8647	0.6677	0.8183	0.9130
	4	ARI	0.8176	0.6411	0.5685	0.4697	0.1717	0.4424	0.5595
		ω	0.9383	0.8979	0.8629	0.8050	0.6088	0.7764	0.8566
40	3	ARI	0.9448	0.9675	0.9515	0.8926	0.5602	0.8747	0.9389
		ω	0.9772	0.9887	0.9844	0.9578	0.7536	0.9434	0.9773
	4	ARI	0.9518	0.8741	0.8386	0.7564	0.4072	0.7782	0.8272
		ω	0.9827	0.9663	0.9556	0.9147	0.6916	0.9117	0.9496
80	3	ARI	0.9900	0.9944	0.9912	0.9880	0.8913	0.9875	0.9960
		ω	0.9968	0.9987	0.9980	0.9970	0.9477	0.9950	0.9987
	4	ARI	0.9929	0.9447	0.9133	0.8617	0.5272	0.8672	0.9140
		ω	0.9971	0.9874	0.9807	0.9583	0.7798	0.9536	0.9804
J	K	Index	$N = 500$						
			DINA-EM	K-means		HACA			
				Best	Ward's	Complete	Single	Average	Ward's
20	3	ARI	0.9653	0.8761	0.8366	0.6028	0.1184	0.6413	0.8066
		ω	0.9894	0.9519	0.9459	0.8109	0.5385	0.8148	0.9301
	4	ARI	0.8863	0.6952	0.6508	0.4270	0.0404	0.4897	0.6220
		ω	0.9676	0.9912	0.8938	0.7573	0.5166	0.7770	0.8760
40	3	ARI	0.9949	0.9797	0.9559	0.8738	0.4122	0.8512	0.9581
		ω	0.9981	0.9935	0.9867	0.9524	0.6757	0.9290	0.9865
	4	ARI	0.9698	0.8591	0.8547	0.6507	0.1064	0.6456	0.8444
		ω	0.9904	0.9634	0.9619	0.8567	0.5501	0.8360	0.9555
80	3	ARI	0.9998	0.9970	0.9951	0.9877	0.6374	0.9874	0.9941
		ω	0.9999	0.9992	0.9986	0.9963	0.8016	0.9950	0.9982
	4	ARI	0.9965	0.9735	0.9735	0.8968	0.4696	0.9038	0.9765
		ω	0.9981	0.9937	0.9937	0.9605	0.7273	0.9591	0.9936

length is sufficiently large. As J increases, the classification abilities from the three methods tend to converge close to 1, which is consistent with the theorem of the previous chapter. Additionally, as expected, the more clusters there are, the less accurate the classification is. Regarding to the impact of sample size, we can see that basically, the case of $N = 500$ classifies slightly better than the case of $N = 100$ for the DINA estimated by MMLE and K-means, though this trend does not hold for HACA when $J = 20$ and 40. Because HACA does not involve fitting either item parameters or cluster centers, it can be thought of as more independent of sample size, and its behavior has more to do with test length. From the aspect of model specification, we can see that larger s and g parameters produce less accurate classification for all the three methods. This is because large slipping and guessing parameters create more noise and greater deviations from ideal response patterns, causing the distributions of $\boldsymbol{W}$ to overlap for different latent classes.

It can be found from the tables that some ARI values under some HACA procedures are much smaller than those under the two K-means methods and DINA-EM, even though the cor-

TABLE 4.
Mean (se) of ARI and ω for DINA data by fitting DINA-EM, K-means and HACA: $s_j \sim$ Uniform$(0, 0.3)$; $g_j \sim$ Uniform$(0, 0.3)$; and $\boldsymbol{\alpha} \sim$ Discrete Uniform.

J	K	Index	$N = 100$						
			DINA-EM	K-means		HACA			
				Best	Ward	Complete	Single	Average	Ward's
20	3	ARI	0.8153	0.6458	0.5797	0.4668	0.1518	0.4441	0.5628
		ω	0.9323	0.8742	0.8424	0.7657	0.5624	0.7393	0.8328
	4	ARI	0.6262	0.3044	0.2828	0.2506	0.0351	0.2322	0.2730
		ω	0.8745	0.7593	0.7474	0.7122	0.5193	0.6783	0.7352
40	3	ARI	0.9531	0.8150	0.7852	0.6879	0.3497	0.6873	0.7515
		ω	0.9812	0.9373	0.9261	0.8661	0.6444	0.8472	0.9055
	4	ARI	0.7817	0.5200	0.4692	0.3987	0.1000	0.3849	0.4679
		ω	0.9264	0.8526	0.8316	0.7768	0.5538	0.7480	0.8252
80	3	ARI	0.9905	0.9093	0.9044	0.8094	0.4930	0.7774	0.8725
		ω	0.9965	0.9726	0.9709	0.9237	0.7278	0.8967	0.9588
	4	ARI	0.9859	0.8239	0.7808	0.7127	0.3112	0.7135	0.7779
		ω	0.9953	0.9521	0.9370	0.8986	0.6478	0.8869	0.9326
J	K	Index	$N = 500$						
			DINA-EM	K-means		HACA			
				Best	Ward's	Complete	Single	Average	Ward's
20	3	ARI	0.8787	0.7073	0.6953	0.4764	0.0826	0.4528	0.6622
		ω	0.9588	0.8938	0.8879	0.7514	0.5231	0.7114	0.8670
	4	ARI	0.6186	0.3292	0.2893	0.2219	0.0184	0.2151	0.2775
		ω	0.8732	0.7543	0.7334	0.6743	0.5057	0.6528	0.7203
40	3	ARI	0.9584	0.8129	0.8094	0.6201	0.1778	0.6311	0.7801
		ω	0.9868	0.9360	0.9348	0.8201	0.5601	0.8034	0.9179
	4	ARI	0.8840	0.5803	0.5438	0.4198	0.0235	0.4560	0.5294
		ω	0.9666	0.8742	0.8560	0.7722	0.5070	0.7693	0.8401
80	3	ARI	0.9951	0.9303	0.9291	0.8281	0.4435	0.8281	0.9237
		ω	0.9977	0.9790	0.9786	0.9290	0.6913	0.9188	0.9754
	4	ARI	0.9786	0.7618	0.7501	0.5811	0.1079	0.5757	0.7445
		ω	0.9928	0.9365	0.9324	0.8496	0.5573	0.8144	0.9243

responding ω values for the three methods are similar. For instance, in Table 3, for the case $(N, J, K) = (100, 20, 4)$, ARI under HACA with complete linkage is 0.4697, which is far off 0.5685 and 0.8176 from K-means with Ward's starting values and DINA-EM, respectively. This is not too surprising because there is no component in ARI to differentiate degrees of similarity of attribute patterns, whereas the ω index gives different weights for different "wrong" patterns. The results imply that HACA may not be as effective as fitting the correct cognitive diagnosis model or using K-means in exactly the correct groups when those groups are uniformly distributed, but behaves similarly when the degree of error is considered.

In this second part, we consider classes with associated and unidentically distributed attributes generated according to the underlying multivariate normal model discussed above. In particular, $\boldsymbol{\alpha}$'s were generated from $MVN(\mathbf{0}_K, \Sigma)$, with variances equal to 1 and correlations of 0.25 and 0.5. Due to the structure of nonequal cutoff points for different attributes (as shown in (25)), some of the 2^K possible clusters were missing in some samples of the simulation. Nev-

TABLE 5.
Mean (se) of ARI and ω for DINA data by fitting DINA-EM, K-means and HACA: $s_j \sim$ Uniform$(0, 0.15)$; $g_j \sim$ Uniform$(0, 0.15)$; and the underlying $\boldsymbol{\alpha} \sim MVN(\mathbf{0}, \Sigma)$ with $\rho = 0.25$.

J	K	Index	$N = 100$						
			DINA-EM	K-means		HACA			
				Best	Ward's	Complete	Single	Average	Ward's
20	3	ARI	0.9542	0.8308	0.6610	0.6876	0.4393	0.7461	0.6572
		ω	0.9870	0.9535	0.9504	0.9235	0.7562	0.9163	0.9463
	4	ARI	0.8571	0.6530	0.4669	0.5472	0.2335	0.6413	0.4476
		ω	0.9678	0.9432	0.9222	0.9135	0.7100	0.9142	0.9109
40	3	ARI	0.9851	0.9277	0.8325	0.8773	0.5915	0.8847	0.8293
		ω	0.9947	0.9896	0.9785	0.9721	0.8195	0.9631	0.9756
	4	ARI	0.9495	0.8487	0.5725	0.6740	0.4786	0.8028	0.5675
		ω	0.9889	0.9801	0.9727	0.9690	0.8085	0.9722	0.9691
80	3	ARI	0.9868	0.9953	0.9323	0.9791	0.8741	0.9910	0.9329
		ω	0.9950	0.9985	0.9940	0.9956	0.9469	0.9974	0.9937
	4	ARI	0.9894	0.9617	0.7409	0.8274	0.8021	0.9143	0.7385
		ω	0.9966	0.9942	0.9903	0.9839	0.9245	0.9859	0.9882
J	K	Index	$N = 500$						
			DINA-EM	K-means		HACA			
				Best	Ward's	Complete	Single	Average	Ward's
20	3	ARI	0.9757	0.8804	0.7937	0.6967	0.2517	0.8341	0.7715
		ω	0.9943	0.9709	0.9681	0.8949	0.6726	0.9244	0.9561
	4	ARI	0.9140	0.6534	0.5156	0.4862	0.0606	0.5723	0.4855
		ω	0.9789	0.9404	0.9158	0.8495	0.6243	0.8489	0.8962
40	3	ARI	0.9982	0.9720	0.8778	0.8097	0.6215	0.8526	0.8779
		ω	0.9996	0.9923	0.9820	0.9398	0.8027	0.9362	0.9801
	4	ARI	0.9842	0.8438	0.6438	0.6788	0.2816	0.8019	0.6402
		ω	0.9965	0.9774	0.9724	0.9308	0.7183	0.9312	0.9673
80	3	ARI	0.9999	0.9960	0.9618	0.9716	0.758	0.9952	0.9653
		ω	1	0.9991	0.9964	0.9941	0.8756	0.9986	0.9968
	4	ARI	0.9900	0.9729	0.7178	0.8156	0.6180	0.9584	0.7375
		ω	0.9965	0.9960	0.9912	0.9792	0.8290	0.9886	0.9917

ertheless, we took 2^K as the number of clusters, which is consistent with the procedure as it has been described, but does raise the issue of searching for a more appropriate number of clusters in some cases. This made the resulting clusters more homogeneous, but could lower the classification indices. Therefore, the agreement indices shown under HACA in the following tables are lower bounds for the true ones. Except for varying attribute distribution, the specifications on s and g remained the same. Tables 5 and 6 show the means and standard errors of ARI and ω under the conditions of small s and g, with correlations between attributes being 0.25 and 0.5, respectively.

These tables showed that among the four linkages of HACA, average linkage outperforms the others in terms of ARI, and complete linkage generally performs better than Ward's linkage. Single linkage does not have comparable performance. K-means with Ward's starting values did not perform as well as HACA with average linkage, but did perform better than the Ward's

TABLE 6.
Mean (se) of ARI and ω for DINA data by fitting DINA-EM, K-means and HACA: $s_j \sim \text{Uniform}(0, 0.15)$; $g_j \sim \text{Uniform}(0, 0.15)$; and the underlying $\boldsymbol{\alpha} \sim MVN(\mathbf{0}, \Sigma)$ with $\rho = 0.5$.

J	K	Index	$N = 100$						
			DINA-EM	K-means		HACA			
				Best	Ward's	Complete	Single	Average	Ward's
20	3	ARI	0.8862	0.7288	0.5945	0.6548	0.5190	0.6678	0.5701
		ω	0.9695	0.9439	0.9464	0.9140	0.8074	0.8871	0.9349
	4	ARI	0.8649	0.6834	0.4900	0.4992	0.3508	0.5861	0.4689
		ω	0.9655	0.9416	0.9368	0.9063	0.7571	0.9024	0.9261
40	3	ARI	0.9852	0.9498	0.7464	0.8127	0.7351	0.9066	0.7396
		ω	0.9943	0.9887	0.9910	0.9819	0.8961	0.9868	0.9870
	4	ARI	0.9369	0.8228	0.5400	0.6089	0.4943	0.7202	0.5445
		ω	0.9759	0.9699	0.9665	0.9670	0.8185	0.9676	0.9664
80	3	ARI	0.9955	0.9883	0.8791	0.9843	0.8756	0.9915	0.8804
		ω	0.9984	0.9992	0.9922	0.9978	0.9477	0.9978	0.9923
	4	ARI	0.9707	0.9555	0.6431	0.7235	0.7818	0.8311	0.6530
		ω	0.9877	0.9934	0.9959	0.9975	0.9346	0.9973	0.9952
J	K	Index	$N = 500$						
			DINA-EM	K-means		HACA			
				Best	Ward's	Complete	Single	Average	Ward's
20	3	ARI	0.9650	0.8175	0.6796	0.6520	0.2339	0.6952	0.6526
		ω	0.9909	0.9682	0.9594	0.8922	0.6881	0.8753	0.9464
	4	ARI	0.8921	0.5751	0.4470	0.4759	0.1185	0.5609	0.4340
		ω	0.9755	0.9378	0.9204	0.8667	0.6575	0.8631	0.9078
40	3	ARI	0.9962	0.9064	0.7252	0.8314	0.5932	0.8624	0.7281
		ω	0.9990	0.9900	0.9821	0.9571	0.8209	0.9465	0.9789
	4	ARI	0.9889	0.7920	0.5923	0.6441	0.2983	0.7375	0.5879
		ω	0.9973	0.9803	0.9697	0.9253	0.7362	0.9146	0.9639
80	3	ARI	1	0.9975	0.8726	0.9945	0.8286	0.9961	0.8780
		ω	1	0.9994	0.9992	0.9984	0.9236	0.9985	0.9990
	4	ARI	0.9997	0.9434	0.6821	0.8300	0.5302	0.9282	0.6939
		ω	0.9999	0.9963	0.9916	0.9838	0.8120	0.9783	0.9910

solutions from which starting values were obtained. Note that although average linkage could classify data with high agreement with the true latent class memberships, the ω index showed that resulting grouping from average linkage was not as homogeneous as Ward's linkage or as complete linkage. From Tables 5 and 6, we find that overall the DINA still performs the best among the three. K-means is slightly better than HACA, but if the cluster linkage is appropriately selected, the difference between K-means and HACA could be very small. If we take the fact that the ARI and ω's are lower bounds for the true performance of HACA into consideration, HACA can be seen as comparable to K-means. The influence of sample size on classification for the three methods is not significant, with larger ARI and ω with $N = 500$ for MMLE and K-means only when J is small, K is small, and the covariance is large. Consistent with the previous part, sample size had little impact on HACA. Similar to the first part of the simulation, the classifications of the three methods improve with test length, and finally converge to values close

to 1. The size of K is still a factor on classification, with worse classification when K is larger. For the three methods, the impact of the size of correlation on classification is noticeable only when the size of items is small, where the smaller the correlation, the better the classification. Similar results can be found for the cases with larger s and g (results are not shown in here), only differing in that K-means and HACA require more items to reach satisfactory agreement levels when $K = 4$.

6. Analysis of Language Testing Data

Next, we apply the methods of clustering, as well as fit cognitive diagnosis models, from data taken from the Examination for the Certificate of Proficiency in English (ECPE). The ECPE is a test developed, administered, and scored by the University of Michigan English Language Institute to test English Language skills. It examines extremely high-level English language skills to determine the English language proficiency of nonnative speakers of English around the world. It is given annually at approximately 125 test centers in 20 countries. The final examination of ECPE contains questions to assess grammar, vocabulary, and reading. Tests designed to evaluate listening, speaking, and writing skills are also available. The data analyzed in this study were collected from the 2003–2004 ECPE grammar section, with 40 multiple-choice questions on conversational American English grammar. A total of 2922 examinees, in which approximately 50% were Portuguese and 31% were Spanish, were used. The average age of examinees was around 23 years old. The same data set has been analyzed by Henson and Templin (2007) and Liu, Douglas, and Henson (2007). In the former work, after 10 trial items were removed, 30 out of 40 items were selected to fit cognitive diagnosis models. The Q-matrix was constructed based on three attributes: Lexical form, Morphosyntactic form, and Cohesive form. The Q-matrix was as well taken from their study, and is listed in Table 7. Note that Items 2 and 7 require none of the three attributes, so they were removed before the analysis.

Henson and Templin (2007) showed that the Reduced RUM fitted the data very well, but Liu et al. (2007) found that with the DINA model, some parameter estimates were unreasonably large. This suggested to adopt the Reduced RUM, rather than the DINA model, as our choice of a cognitive diagnosis model to compare with the clustering techniques. Also, in Henson and Templin's work, they estimated parameters in the Reduced RUM using an MCMC algorithm to sample from the posterior distribution of a Bayesian formulation of the Reduced RUM. In determining the prior of examinees' attribute patterns, they proposed to follow the empirical Bayesian method. Therefore, we fitted the reduced RUM, and took the output obtained by Henson and Templin (2007), as shown in Tables 8, 9, and 10, to calculate the prior of all possible attribute patterns. Specifically, the tetrachoric correlation matrix in Table 10 was used to describe the relationships between skill pairs, and a underlying multivariate normal distribution on examinees' attribute patterns was assumed. Based on the estimates of skill mastery proportions as shown in Table 9, a cutoff point with respect to each skill was obtained by taking $\Phi^{-1}(1 - \text{Prop of Mast})$. The prior of examinees' attribute patterns was computed by evaluating the probability determined by the cutoff points and the multivariate normal curve with mean $(0, 0, 0)$' and covariance matrix as Table 10. Examinee's classes were consequently decided by taking the pattern maximizing the posterior probability over the 8 possible values of $\boldsymbol{\alpha}$. In addition to classification according to the Reduced RUM, examinees were also classified by applying K-means and HACA with complete linkage, with statistic $\boldsymbol{W}$ as input.

Three indices, cluster size, within-cluster mean of $\boldsymbol{W}$, and square root of mean squared residual (MSR) of $\boldsymbol{W}$ were used to evaluate classification resulting from the Reduced RUM, K-means, and HACA with complete linkage. The mean of $\boldsymbol{W}$ is an indicator of how well the examinees' pattern within a cluster are identified, in the sense that the means when taken as

TABLE 7.
Q-matrix for ECPE data.

Item	Attribute		
	Mor	Coh	Lex
G1	1	1	0
G2	0	0	0
G3	0	1	0
G4	1	0	1
G5	0	0	1
G6	0	0	1
G7	0	0	0
G8	0	0	1
G9	1	0	1
G10	0	1	0
G11	0	0	1
G12	1	0	0
G13	1	0	1
G14	1	0	1
G15	1	0	0
G16	1	0	0
G17	0	0	1
G18	1	0	1
G19	0	1	1
G20	0	0	1
G21	0	0	1
G22	1	0	1
G23	1	0	1
G24	0	0	1
G25	0	1	0
G26	0	1	0
G27	1	0	0
G28	0	0	1
G29	1	0	0
G30	0	0	1

Mor: Morphosyntactic form
Coh: Cohesive form
Lex: Lexical form

vectors should be somewhat distinct across the 8 clusters. If examinees in a particular group have the same attribute pattern, mean $\boldsymbol{W}$ is expected to have the pattern of relatively large value(s) on the dimension(s) of 1's, and smaller value(s) on 0's. In addition, MSR of $\boldsymbol{W}$ is an index reflecting how homogeneous a cluster is. Note that the MSR of $\boldsymbol{W}$ for cluster m is calculated as follows:

$$MSR(m) = \frac{\sum_{i=1}^{N_m} \| \boldsymbol{W}_i^{(m)} - \overline{\boldsymbol{W}}^{(m)} \|^2}{N_m},$$

where N_m is the number of examinees classified into cluster m. For the Reduced RUM, clusters were labeled with the attribute pattern which maximized the posterior likelihood. Attribute labels are not directly available for K-means and HACA. However, considering the feature of partial ordering among the 8 attribute patterns, the results from K-means and HACA were sorted along with means of sum-scores, illustrating how one would infer attribute patterns from clusters in

TABLE 8.
Estimated item parameters for the Reduced RUM.

Item	π^*	r^*		
		Mor	Coh	Lex
G1	0.93	0.89	0.84	–
G3	0.90	–	0.81	–
G4	0.78	0.63	–	0.83
G5	0.82	–	–	0.56
G6	0.96	–	–	0.78
G8	0.92	–	–	0.76
G9	0.94	0.73	–	0.70
G10	0.97	–	0.84	–
G11	0.79	–	–	0.67
G12	0.89	0.58	–	–
G13	0.92	0.77	–	0.69
G14	0.73	0.51	–	0.38
G15	0.90	0.73	–	–
G16	0.82	0.66	–	–
G17	0.96	–	–	0.76
G18	0.91	0.75	–	0.72
G19	0.94	–	0.93	0.91
G20	0.91	–	–	0.78
G21	0.84	–	–	0.53
G22	0.76	0.49	–	0.52
G23	0.92	0.85	–	0.69
G24	0.79	–	–	0.37
G25	0.94	–	0.70	–
G26	0.70	–	0.47	–
G27	0.77	0.68	–	–
G28	0.78	–	–	0.69
G29	0.69	0.43	–	–
G30	0.91	–	–	0.69

TABLE 9.
Estimates of Skill Mastery.

Skill	Prop of Mast
Mor	0.406
Coh	0.571
Lex	0.663

practice. Although this sorting criteria does not give an absolutely correct order, it does provide much information about examinees' mastery or nonmastery on a particular attribute.

The data were fitted by the Reduced RUM, and also analyzed by HACA and K-means through the statistic $\boldsymbol{W}$, with results shown in Tables 11, 12, and 13, respectively. As we can see from Table 11, the reduced RUM classified most examinees into the two polar ends of classes, which could be due to high correlations between skill pairs. The mean of $\boldsymbol{W}$ shows that the Reduced RUM gave well-separated cluster means, which is helpful with identifying examinees' overall patterns. However, clusters formed by this method tended to have relatively large MSRs. This indicates the given data are not grouped compactly based on examinees' patterns so that separate cluster means might lead to more heterogeneous clustering.

TABLE 10.
Estimates of Skill Association.

	Skill 1	Skill 2	Skill 3
Skill 1	1	0.748	0.607
Skill 2		1	0.761
Skill 3			1

TABLE 11.
Classification with the Reduced RUM model.

Class	Pattern	Size	Mean $\boldsymbol{W}$			$\sqrt{MSR\boldsymbol{W}}$	Mean Sum
			W_1	W_2	W_3		
1	(0 0 0)	889	5.70	4.00	9.11	3.13	14.65
2	(1 0 0)	16	10.00	3.75	10.50	1.15	19.13
3	(0 1 0)	0	–	–	–	–	–
4	(0 0 1)	320	7.37	3.72	13.54	2.09	19.09
5	(1 1 0)	37	9.24	5.54	10.76	1.34	19.49
6	(1 0 1)	12	10.58	2.58	15.75	1.58	21.50
7	(0 1 1)	515	7.28	5.42	13.37	1.98	20.36
8	(1 1 1)	1133	11.01	5.37	15.80	2.12	24.33

TABLE 12.
Classification by HACA with complete linkage.

Size	Mean $\boldsymbol{W}$			$\sqrt{MSR\boldsymbol{W}}$	Mean Sum
	W_1	W_2	W_3		
72	2.75	3.33	5.01	2.23	9.08
88	4.51	2.45	6.72	1.71	10.86
31	7.13	4.77	6.77	1.79	14.23
449	5.00	4.42	9.35	2.12	14.93
201	7.66	3.34	12.7	1.66	18.21
309	8.15	4.74	10.93	1.92	18.31
471	6.79	4.87	13.51	1.73	19.63
1301	10.73	5.3	15.63	2.26	23.94

In contrast, Table 12 shows that HACA classified most data to the highest class of pattern (1, 1, 1), with large sum-scores across all dimensions. Some values of mean $\boldsymbol{W}$ are ambiguous in their patterns, but clusters are most homogeneous among the three methods. This is consistent with the fact that the HACA with complete linkage tends to form tight and homogeneous clusters, but in this case, the trade-off is confounding attribute patterns among examinees within a cluster. Regarding the results of K-means, Table 13 shows that examinees were more evenly distributed into classes, which might not be consistent with the estimated structure of highly correlated skills. Also, each column of mean $\boldsymbol{W}$ increases in size along with sum-score. This implies that the within-cluster mean values of $\boldsymbol{W}$ obtained from K-means do not have recognizable patterns to reflect examinees' attribute patterns.

Taking the Reduced RUM as the standard, we further examined how well HACA and K-means classify examinees by investigating the classification agreement of the two methods with the Reduced RUM, as well as the agreement between themselves, as shown in Table 14. The agreement between the Reduced RUM and HACA is much higher than the other pairs with an ARI of 0.53 compared to an ARI of 0.29 for the agreement between K-means and the Reduced

TABLE 13.
Classification by K-means.

Size	Mean $\boldsymbol{W}$			$\sqrt{MSR\boldsymbol{W}}$	Mean Sum
	W_1	W_2	W_3		
192	3.83	3.45	5.77	2.27	10.58
221	4.27	3.92	9.31	1.63	14.07
243	6.87	4.49	9.18	1.86	15.87
467	6.17	4.51	12.19	1.65	17.96
385	8.76	4.71	12.19	1.55	19.67
430	8.25	4.94	14.77	1.42	21.42
534	10.58	5.19	15.19	1.51	23.40
450	12.17	5.52	17.2	1.40	26.22

TABLE 14.
ARI table for Reduced RUM, K-means, and HACA with a variety of linkages.

	R-RUM	K-means	HACA			
			Complete	Single	Average	Ward's
R-RUM	–	0.292	0.531	<0.001	0.388	0.292
K-means		–	0.326	<0.001	0.342	0.503
HACA						
Complete			–	0.003	0.493	0.347
Single				–	0.003	0.001
Average					–	0.400
Ward's						–

RUM. Combining this result to the previous analysis, one can argue that HACA with complete linkage outperformed K-means for this data set, perhaps because of its highly varied cluster sizes that the skill associations, if they are valid, would imply.

Next, we consider how to implement this procedure without fitting all possible 2^K clusters. The number of clusters will be chosen according to a scree plot of the fusion coefficients, which are the distances between clusters that are joined in each step of a HACA analysis. Once the number of clusters has been determined, the remaining step is to solve the labeling problem, which is to associate each cluster with a specific attribute pattern. Though there is much work to do on this problem, particularly when the number of clusters is large, we present a method based on an exhaustive search of all possible ways of labeling that utilizes an objective function measuring the consistency of the clusters with respect to underlying attribute patterns.

A scree plot shows the relationship between the number of clusters and the fusion coefficient that resulted from joining two clusters to result in the current number as the HACA routine goes from the maximum possible 2^K clusters to any smaller number of clusters, which essentially means that some attribute patterns might be ignored. It can be helpful to read a scree plot from right to left and notice that reducing the number of clusters is quite reasonable as long as the distance between the joined clusters grows very slowly. However, when at a stage where joining two clusters would result is a dramatic change in distance, the elbow of the scree is revealed and the HACA process should be terminated.

Once the number of clusters has been decided, the next issue is associating each cluster with an attribute pattern. A benefit of cognitive diagnosis is that it allows instructors and examinees to finely understand examinees' cognitive strengths and weaknesses so that appropriate remediation can be developed to suit examinees' learning needs. This, however, relies heavily on correctly assigning examinees to a homogeneous cluster and properly labeling the cluster based

on the attribute pattern that the examinees possess. The theory has shown that given a particular cognitive diagnosis model, K-means and HACA can perform nearly as well as model-based methods, though labeling the estimated clusters based on examinees' attribute patterns remains an unsolved issue. When classifying data using cluster analysis in the cognitive diagnosis setting, there is no known or estimated parametric model to utilize. Therefore, a useful labeling method should be designed to be somewhat robust to the true model. This is a challenging problem that will require much more research, but we present one possible method below and apply it to the language testing data.

An underlying pattern of a particular cluster, say $\boldsymbol{\alpha}$, is identified by comparing its resulting true score vector $\boldsymbol{T}(\boldsymbol{\alpha})$ with $\boldsymbol{T}(\boldsymbol{\alpha}^*)$ of another pattern $\boldsymbol{\alpha}^*$. However, because $\boldsymbol{T}(\boldsymbol{\alpha})$ cannot be observed, we used the within cluster means of $\boldsymbol{W}$, $\overline{\boldsymbol{W}}$, to represent them. The proposed labeling method starts from defining an inconsistency index, which is used to indicate the extent that a particular $\overline{\boldsymbol{W}}$ orders clusters differently than some simple assumptions on the underlying model would suggest. Of all possible ways of assigning attribute patterns to clusters, we choose the one that minimizes the inconsistency index.

In constructing the inconsistency index, we utilize the assumption that for patterns $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$, where $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$, if $\alpha_k = 1$ and $\alpha_k^* = 0$, $T_k(\boldsymbol{\alpha}) > T_k(\boldsymbol{\alpha}^*)$ follows. This assumption is then used to generate an inconsistency index, IC, as follows:

$$\begin{aligned} IC = {} & I\big[T_k(\boldsymbol{\alpha}) \leq T_k\big(\boldsymbol{\alpha}^*\big)\big] I\big[\alpha_k > \alpha_k^*\big] \\ & + I\big[T_k(\boldsymbol{\alpha}) < T_k\big(\boldsymbol{\alpha}^*\big)\big] I\big[\alpha_k = \alpha_k^* = 1, \|\boldsymbol{\alpha}\| > \big\|\boldsymbol{\alpha}^*\big\|\big]. \end{aligned} \tag{27}$$

Equation (27) indicates whether entries in $\boldsymbol{T}(\boldsymbol{\alpha})$ and $\boldsymbol{T}(\boldsymbol{\alpha}^*)$ are inconsistent based on the patterns of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$. More specifically, suppose that there is a test requiring K attributes, then there exist $2^K = M$ possible attribute patterns. For K-means and HACA, since there is no estimated cluster with known underlying pattern, the searching and matching will proceed by taking into account all C_2^M possible pairs of $\overline{\boldsymbol{W}}$ vectors from each of the $M!$ possible orders. Given an intended order of attribute patterns, say,

$$\mathbf{A}_0 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

the best solution is obtained by taking the permutation, among the $M!$ permutations, in which the corresponding $\overline{\boldsymbol{W}}$'s of the permutated clusters, which are used in place of the unobservable true scores, result in a minimal inconsistency index with respect to $\mathbf{A}_0$.

Denote the $M!$ possible permutations as $P_1, P_2, \ldots, P_{M!}$, and the vector $\overline{\boldsymbol{W}}_m^{(i)} = (\overline{W}_{m1}^{(i)}, \overline{W}_{m2}^{(i)}, \ldots, \overline{W}_{mK}^{(i)})$ as the mean of $\boldsymbol{W}$ vectors in the mth estimated cluster under the permutation P_i, where $i = 1, 2, \ldots, M!$. Additionally, matrix $\overline{\boldsymbol{W}}^{(i)} = (\overline{\boldsymbol{W}}_1^{(i)}, \overline{\boldsymbol{W}}_2^{(i)}, \ldots, \overline{\boldsymbol{W}}_M^{(i)})^T$ represents the collection of those $\overline{\boldsymbol{W}}$'s permutated by P_i. If we define the inconsistency function between mean vectors $\overline{\boldsymbol{W}}_s^{(i)}$ and $\overline{\boldsymbol{W}}_t^{(i)}$ based on attribute vectors $\mathbf{A}_{0s}$ and $\mathbf{A}_{0t}$ as $IC((\overline{\boldsymbol{W}}_s^{(i)}, \overline{\boldsymbol{W}}_t^{(i)}), (\mathbf{A}_{0s}, \mathbf{A}_{0t}))$, then it is straightforward that the inconsistency index for $\overline{\boldsymbol{W}}^{(i)}$ with respect to $\mathbf{A}_0$ can be com-

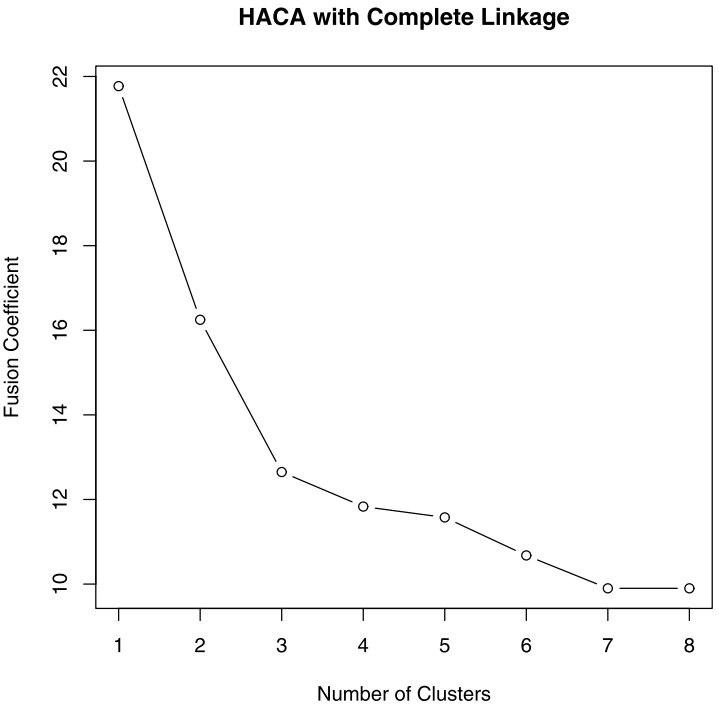

FIGURE 1.
The scree plot for HACA with complete linkage.

puted by

$$IC\big(\overline{\boldsymbol{W}}^{(i)}, \mathbf{A}_0\big) = \sum_{s=1}^{M-1} \sum_{t=s+1}^{M} IC\Big(\big(\overline{\boldsymbol{W}}_s^{(i)}, \overline{\boldsymbol{W}}_t^{(i)}\big), (\mathbf{A}_{0s}, \mathbf{A}_{0t})\Big).$$

Then the best solution P_i is taken by the following criterion:

$$P_i = \arg\min_{i=1}^{M!} IC\big(\overline{\boldsymbol{W}}^{(i)}, \mathbf{A}_0\big).$$

Note again that the assumption for this criterion is that examinees are well classified, meaning the error existing between $\boldsymbol{T}(\boldsymbol{\alpha})$ and $\overline{\boldsymbol{W}}$ should be small.

A concern is that the number of comparisons grows according to $M!$, which can become impossible to exhaustively compute as M grows. This problem can be partly alleviated by first partially ordering the clusters, and only calculating the IC's for the remaining permutations after the order of certain clusters are fixed. This can effectively reduce the number of required permutations. For example, for a test of K being 3, and using all 8 possible clusters, the number of total permutations is $2^3! = 40320$, while the number of remaining permutations, after taking partial order first, is $3! \times 3! = 36$.

By applying the stopping rule based on the scree plot, with an expanded version of the vector $\boldsymbol{W}$ that utilized additional sum scores for items measuring single attributes, a modification that works well with exams following either the DINA or the Reduced RUM; the ELI data using HACA with complete linkage required 3 clusters. The scree plot is shown in Figure 1. After determining the best number of clusters, the labeling algorithm was then applied. The results are in Table 15, which shows the $\overline{\boldsymbol{W}}$ for the first 3 sum scores of $\boldsymbol{W}$ for the three clusters. Note that

TABLE 15.
Mean of $\boldsymbol{W}$ by cluster labels.

$\overline{W}$			Label 1			Label 2		
4.232	3.442	6.408	0	0	0	0	0	0
7.125	4.620	11.986	0	1	1	1	1	0
10.949	5.282	16.014	1	1	1	1	1	1

two different possible sets of attribute labels tied for the best IC value. The first set of labels is most consistent with the estimated proportions of label mastery given by the Reduced RUM model. Though much work remains to be done on selection of the number of clusters and cluster labeling, the methods presented here represent initial steps in solving that difficult problem.

7. Discussion

Methods for utilizing cluster analysis for the purpose of cognitive diagnosis have been proposed as alternatives to fitting restricted latent class models. The value of these methods rests partly in convenience of reducing the response data to a vector of sum-scores, rather than specifying and fitting more elaborate latent class models. As was shown in the section concerning asymptotic theory, classification will be accurate if the statistic upon which clustering is based has expected values that are separated for each attribute pattern. In this manuscript, the statistic $\boldsymbol{W}$ was used and it was shown that this is appropriate for the DINA model. There are certainly unresolved theoretical issues, such as determining the list of models for which $\boldsymbol{W}$ is still appropriate, and finding better statistics for models in which it does not suffice. Ultimately, the utility of cluster analysis as an alternative to latent class modeling might rest in identifying a statistic that results in correct classification for a wide variety of underlying models, making robustness a selling point of the method.

The simulations presented here showed that using the latent class model is always more efficient, at least when it is known. However, the simplicity of cluster analysis is an attractive feature and sum-scores can always be easily computed. One drawback of cluster analysis for this application is that the output is a set of unlabeled clusters, unlike restricted latent class models that yield posterior probabilities for each of the 2^K attribute patterns. Though cluster analysis can place subjects in homogeneous groups, one must use post hoc techniques for discerning these attributes from the clusters. Some ideas to resolve this issue were presented in the analysis of the ELI data, though more work along this line is needed, perhaps defining partial orders in the mean vectors within-clusters and matching them with theoretical partial orders implied by differing attribute patterns, which was done in the analysis of language testing data.

More work is also needed to determine what method of clustering is most useful for this application, and perhaps even whether distance measures other than Euclidean are more appropriate. Hierarchical agglomerative cluster analysis with several different linkages performed quite well in simulation studies, and is relatively easy to study theoretically, in contrast with the more difficult to study K-means. Single linkage shares the rather tractable properties of other HACA methods, but we found that it breaks down too easily as one departs from the conditions used to prove the theorem on consistency. If much overlap between clusters exists, single linkage can suffer from chaining that degrades its performance. K-means tended to perform better, but more work on the theory to support it must be done, and reliable methods for determining good starting values must be found. The primary motivation for this work was to step back and see if standard techniques of multivariate analysis could be used on simple-sum scores, rather than engage in highly complicated and technical latent class modeling. The results indicate that such methods are promising, though the problem is large and many questions need to be answered.

References

Blashfield, P.K. (1976). Mixture model tests of cluster analysis: accuracy of four agglomerative hierachical methods. *Psychological Bulletin*, *83*, 377–385.

Bradley, P.S., & Fayyad, U.M. (1998). Refining initial points for *K*-means clustering. In J. Shavlik (Ed.), *Proceedings of the fifteenth international conference on machine learning* (pp. 91–99). Burlington: Morgan Kaufmann.

Bartholomew, D.J. (1987). *Latent variable models and factor analysis*. New York: Oxford University Press.

Cunnningham, K.M., & Ogilvie, J.C. (1972). Evaluation of hierachical grouping techniques: A preliminary study. *Computer Journal*, *15*, 209–213.

de la Torre, J., & Douglas, J.A. (2004). Higher order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353.

Embretson, S. (1997). Multicomponent response models. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). New York: Springer.

Everitt, B.S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). London: Arnold.

Forgy, E.W. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics*, *21*, 768–769.

Hartigan, J.A. (1978). Asymptotic distributions for clustering criteria. *The Annals of Statistics*, *6*, 117–131.

Haertel, E.H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 333–352.

Hands, S., & Everitt, B.S. (1987). A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techiniques. *Multivariate Behavioural Research*, *22*, 235–243.

Hartigan, J.A. (1975). *Clustering algorithms*. New York: Wiley.

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. (2007). Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Hoeffding, W. (1963). Probabilistic inequalities for sums of bounded random variables. *Annals of Mathematical Statistics*, *58*, 13–30.

Hubert, L.J., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

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.

Kaufman, J., & Rousseuw, P. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

Kuiper, F.K., & Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. *Biometrics*, *31*, 777–783.

Lattin, J., Carroll, J.D., & Green, P.E. (2003). *Analyzing multivariate data*. Pacific Grove: Brooks/Cole, Thomson Learning.

Liu, Y., Douglas, J., & Henson, R. (2007). *Testing person fit in cognitive diagnosis*. Unpublished manuscript.

MacQueen, J. (1967). Some methods of classification and analysis of multivariate observations. In L.M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth Bekeley Symposium on Mathematical Statistics and Probability* (pp. 281–207). Berkeley: 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*, *33*, 379–416.

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

Milligan, G.W. (1980). An examination of the effects of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45*, 325–342.

Muthén, L.K., & Muthén, B.O. (2006). *Mplus user's guide* (4th ed.). Los Angeles: Muthén & Muthén.

Pena, J., Lozano, J., & Larranaga, P. (1999). An empirical comparison of four initialization methods for the *K*-means algorithm. *Pattern Recognition Letters*, *20*, 1027–1040.

Pollard, D. (1981). Strong consistency of *K*-means clustering. *The Annals of Statistics*, *9*(1), 135–140.

Pollard, D. (1982). Quantization and the method of *K*-means. *IEEE Transactions on Information Theory*, *28*, 199–205.

Punj, G., & Stewart, D.W. (1983). Cluster analysis in marketing research: A review and suggestions for application. *Journal of Marketing Research*, *20*, 134–148.

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

Steinley, D. (2003). Local optima in *k*-means clustering: What you don't know may hurt you. *Psychological Methods*, *8*, 294–304.

Steinley, D. (2006). *K*-mean clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*, 1–34.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Applied Statistics (JRSS-C)*, *51*, 337–350.

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

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

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

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

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

Willse, J.T., Henson, R.A., & Templin, J.L. (2007). *Using sumscores or IRT in place of cognitive diagnostic models: Can more familiar models do the job*? Presented at the annual meeting of the National Council on Measurement in Education, Chicago, Illinois.

Manuscript Received: 27 SEP 2007
Final Version Received: 16 MAR 2009
Published Online Date: 5 MAY 2009