Cluster analysis for cognitive diagnosis: An application to the 2001 PIRLS reading assessment

Chia-Yi Chiu
Department of Educational Psychology, Rutgers University, New Brunswick, New Jersey, United States of America

Minhee Seo
Department of Educational Research Methodology, University of North Carolina at Greensboro, Greensboro NC, United States of America

Demand for large-scale assessments that report more diagnostically informative results about examinees' cognitive profiles is increasing. Traditionally, classification based on examinees' attribute patterns has been carried out by fitting data with cognitive diagnosis models. A recently proposed method of reaching the same classification goal uses classical cluster analysis without utilizing an item response model. The only requirements are a valid sample statistic, usually obtained by summarizing data, and the assumed item-by-attribute matrix used in most cognitive diagnosis modeling. After constructing 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. An application to the 2001 Progress in International Reading and Literacy Study (PIRLS) (Gonzalez & Kennedy, 2001) reading data is conducted to illustrate how the methods can be implemented in practice.

INTRODUCTION

The significance of large-scale assessments in public education has grown tremendously in recent years. Along with the increasing demand for this type of assessment, there is increasing pressure to make the assessments more diagnostically informative about students' cognitive strengths and weaknesses (Leighton & Gierl, 2007). However, most existing large-scale assessments, including PIRLS, report only students' overall performances on the test, a practice that provides limited diagnostic information about students' cognitive capacities. One factor limiting application of cognitive diagnosis analysis to large-scale data is that these assessments are usually not designed for the purpose of diagnosis. To adequately extract information on examinees' cognitive abilities, items have to be written in a way that ensures the item responses identify if the examinee possesses the required skills for answering a particular item correctly. Issues such as these associated with test construction and diagnostic testing have been widely discussed. Gorin (2007), for example, having assembled several popular methods of diagnostic testing, provides a thorough discussion on how to develop and construct and evaluate them.

Among the various types of analysis in the cognitive diagnosis context, classification based on students' mastery or non-mastery of each attribute in a set of attributes is considered necessary because it can lead to more efficient remediation for examinees' learning. Progress in International Reading and Literacy Study (PIRLS) data have been analyzed using Item Response Theory (IRT)-based models, in which continuous latent traits are assumed. However, most models developed for the purpose of cognitive diagnosis are from the family of latent class models. Models with continuous latent variables may achieve the goal of classifying examinees by partially ordering them according to their general latent traits, but those with discrete latent variables classify examinees by directly assigning them to the most likely group. By modeling many ordered latent classes, we can see that the IRT model can be approximated by the latent class model.

However, several questions arise during exploration of differences between the conclusions that might be reached from these models. Certainly, there is some convenience in sorting examinees into a few small bins, but we must question whether this level of classification is too coarse. Also, if the item response probabilities across the latent classes cannot be ordered, we must question if there is also some sort of multidimensionality when the latent variable is seen as continuous. This consideration raises issues about the definition of dimensionality and its dependence on whether the viewpoint taken is a latent trait or latent class one. We prefer these latent class models in applications in which the number of latent classes is not too large. Therefore, with the purpose of identifying examinees' cognitive profiles, the latent variable is assumed to be discrete, and the underlying model is taken from the cognitive diagnosis context.

When a cognitive diagnostic model is specified, classification can be done by fitting the model and by estimating parameters through likelihood functions. However, applying complex cognitive diagnostic models requires sophisticated software along with the expectation–maximization (EM) algorithm, or Markov chain Monte Carlo (MCMC). However, most software for estimating cognitive diagnostic models is not available in the public domain, and the implementation of either EM or MCMC requires advanced computational skills.

As an alternative, Chiu, Douglas, and Li (in press), propose a new classification method for achieving the same goal of classifying examinees based on their attribute profiles—a method that utilizes exploratory cluster analysis. The classification requires clustering on a properly chosen summary score of the data, which may be constructed by incorporating the pre-determined item-by-skill information; no further model assumptions are needed. Unlike the model-based method, users can run familiar and widely available software to conduct classifications; depending on the method being used, computer running time can be very short. We elaborate details of the new method in a later section.

As mentioned, most large-scale assessments are not designed to measure students' cognitive capacities. Another concern additional to the issue of test construction is that of missing the pre-established Q-matrix (Tatsuoka, 1985) during classification. A Q-matrix is an array of entries of either 1 or 0 indicating the item-by-skill information, and is usually established according to experts' opinions. Through the Q-matrix, information about whether an examinee possesses particular cognitive attributes can be assessed by analyzing his or her responses to items requiring those skills. The Q-matrix for the data of interest to us, PIRLS 2001, is not yet available. However, the fact that the items in PIRLS were written with reference to specific reading purposes and procedures provides a possible framework from which to relate items to their potential required skills.

It is this feature that made this assessment stand out for us over the other large-scale assessments and gave us an opportunity to conduct this preliminary classification study. Despite the possibility of extra noise in the data due to rough Q-matrix specifications and instability of the classification results, this item-by-skill structure provided us with the information that we needed not only to make use of the new classification method but also to inform future research based on more elaborate Q-matrix specifications. Our primary interest therefore in conducting the study was to investigate if the way the PIRLS items are written provides an applicable means of obtaining information about examinees' skill profiles. We also wanted to examine the effect of utilizing cluster analysis to classify PIRLS data based on the method developed by Chiu and colleagues (in press).

RESTRICTED LATENT CLASS MODELS FOR COGNITIVE DIAGNOSIS

Most specialized latent class models for cognitive diagnosis are formulated under the assumption of a pre-determined Q-matrix. A Q-matrix is a $J \times K$ array $\mathbf{Q}$, in which the (j, k) entry q_{jk} denotes whether or not the j^{th} item requires the k^{th} attribute. The latent ability variable is assumed to be discrete and multidimensional, meaning that, for each examinee, his or her profile is a composite of multiple discrete attributes. An attribute could refer to a skill required to solve items or an unobservable psychological construct. More specifically, let $\boldsymbol{\alpha}$ be a K-dimensional vector for which the k^{th} component, α_k, indicates whether or not an examinee possesses the k^{th} attribute or skill, for $k = 1,2,...,K$. The vector $\boldsymbol{\alpha}$ can take 2^K distinct values, each indicating one of the 2^K latent classes in the model. The feature that distinguishes the models from one another is the assumptions that dictate how attributes are utilized to construct responses. In this study, we used DINA (deterministic input, noisy output "AND" gate) (Junker & Sijtsma, 2001) as a model for comparison to cluster analysis. (For more information about cognitive diagnostic models, see Rupp & Templin, 2007.)

The DINA Model

The DINA model is a member of restricted latent class models. Its item response function is expressed as,

$$P(Y_{ij} = 1 \mid \alpha_i) = (1 - s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})},$$

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 j^{th} item, and $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ is the ideal response equaling 1 if the i^{th} examinee possesses all skill(s) required for answering the j^{th} item correctly, and 0 if any required skill(s) is missing. As we can see, the η_{ij} maps the examinee's skill possession and the item requirements into the set $\{0,1\}$.

The DINA model is characterized by its conjunctive structure, where the probability of answering an item correctly will substantially drop if any of the required attributes are not mastered or possessed. The estimation of the DINA model has been successfully carried out by employing the EM algorithm (Haertel, 1989) or MCMC (de la Torre & Douglas, 2004; Tatsuoka, 2002).

Cluster Analysis

The DINA model and other latent variable models for cognitive diagnosis all require sophisticated software for fitting, either with the EM algorithm or by MCMC. Cluster analysis can serve as an alternative method of classifying examinees despite not having complete knowledge of the underlying cognitive diagnostic model. This section briefly describes two commonly used cluster analysis methods—K-means and hierarchical agglomerative cluster analysis (HACA). The rationale for using cluster analysis in the cognitive diagnosis setting is outlined in the next section.

K-means

K-means cluster analysis is a widely used partitional clustering technique for clustering subjects based on a vector of data. The *K*-means algorithm requires estimating the cluster centers, with the number of clusters being pre-determined. Once the centers are decided, data are sent to the closest cluster. Specifically, consider a data matrix with *N* subjects and *K* observed variables, where the row entries are $\boldsymbol{w}_1, \boldsymbol{w}_2,\ldots,\boldsymbol{w}_N$, and the goal is to cluster the *N* subjects into *M* clusters, where $M=2^K$. Then, taking Euclidean distance, for instance, the *K*-means assigns data point $\boldsymbol{w}_1$ to the $\boldsymbol{m^{th}}$ cluster if $||w_1 - \hat{c}_m||^2$ is minimal over all **m**. Here, $\hat{c}_m$ is the estimated center of the $\boldsymbol{m^{th}}$ cluster, and is simply the average of the data in the cluster. The iteration to carry out the final solution is as follows:

1. Choose *M* initial *K*-dimensional cluster centers;
2. Assign data points to the closest cluster;
3. Obtain the updated cluster center by averaging the assigned observations;
4. Repeat 2 and 3 until the assignment is not changed.

It is known that the *K*-means outcome is affected by the initial values. Convergence to local optima may occur if poor starting values are used. Many methods of initializing starting values for the *K*-means algorithm have been proposed (Bradley & Fayyad, 1998; Forgy, 1965; Kaufman & Rousseuw, 1990; MacQueen, 1967; Steinley, 2006). Random selection is a commonly used criterion, but in order to make it more efficient and adequate, several variations are available. For example, a variation proposed by Forgy (1965) begins with the analyst randomly selecting a set of data points as seeds of cluster centers, assigning the rest of the data to the closest cluster, updating the centers of the clusters by averaging the data within the clusters, and then taking these averages as starting values for the *K*-means algorithm. In addition to the series of random selection methods, Kaufman and Rousseuw (1990) proposed a sophisticated method that determines one center at a time by maximizing a criterion index.

To learn which of the initialization methods worked more adequately than others, Pena, Lozano, and Larranaga (1999) conducted an empirical comparison study for a variety of methods, and suggested that the random method and the Kaufman and Rousseuw (1990) methods outperformed the others in terms of some criteria. However, it is worth noting that while one method may be more appropriate than the other under certain conditions, there is no global solution for this issue. In our study, we used the software R to carry out the analysis. If the user specifies only the number of desired clusters—2^K in our case—but not a starting point, R selects a random set of 2^K distinct rows as the initial center. The function for multiple sampling is also available by furthermore specifying the number of such random sets being used in the algorithm to create a better initial point.

Hierarchical agglomerative cluster analysis (HACA)

HACA differs from *K*-means in a few respects. First, with *K*-means, data are partitioned into exclusive clusters; with HACA, hierarchical clustering forms a dynamic tree structure. Second, HACA is much simpler than *K*-means in computational terms, and

does not require the selection of initial values. There are different variations of HACA's clustering algorithm, depending on how we define the distance between data points and the distance between clusters. Once the distance measures are decided, the HACA algorithm starts by defining one cluster for each subject. The next step is to cluster the two subjects for whom the distance between them is smallest. (Note that the distance of two clusters is basically the distance of the two data points in the clusters.) At each step thereafter, a new cluster is formed by fusing the two closest clusters, and the distance between clusters here is a function of the distances between their data points. Defining these distances between clusters is what distinguishes the different linkage methods for HACA. Clusters are combined by using one of the following linkages to minimize the distance between the clusters that are fused in each step until the process is stopped at a fixed number of clusters or until only one cluster, containing all of the objects, remains. In our application, the process is stopped at the point where there are 2^K clusters.

Following are some common linkages for HACA. The first is complete linkage, in which the distance between clusters is the maximum distance between two data points from the two target clusters. The two clusters that have the smallest such maximum distance are merged to form a new cluster. Specification of this linkage means that the data in each cluster at each stage are enclosed within a certain known range; complete linkage clustering therefore 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. Single linkage tends to produce long, stringy clusters and non-convex shapes, a tendency that is known as the chaining effect. Findings (Chiu & Douglas, 2008; Chiu et al., in press) show that this method performs poorly in similar applications, so we did not employ this linkage in our study.

In contrast to the two extreme distance measures, average linkage clustering defines the mean of distances between data points in two different clusters as the distance between the two clusters. Average linkage tends to produce ball-shaped clusters and is a quite robust method. Considering cluster homogeneity from the within-cluster variability point of view, Ward (1963) proposed a hierarchical clustering method in which clusters are chosen to merge so that the updated within-cluster sum of squared errors is minimized. The classification rationale behind this linkage is similar to that of *K*-means. Therefore, Ward's (1963) linkage possesses properties similar to those of *K*-means because it tends to produce nearly equal-sized clusters that are convex and compact. However, it suffers from sensitivity to outliers (Milligan, 1980).

RATIONALE BEHIND THE STUDY METHOD

In this section we introduce the classification method developed by Chiu and colleagues (in press). We do this by presenting a description of how to construct an appropriate sample statistic as an input for the clustering technique selected to classify examinees into the correct latent classes based on their cognitive patterns.

The Selected Sample Statistic

The first step in this method is to construct a sum-score vector, with entries indicating the sub-sum scores corresponding to the required skills. Begin by defining the vector as $\boldsymbol{W}_i$ for the i^{th} examinee; the k^{th} entry of $\boldsymbol{W}_i=(W_{i1},W_{i2},...,W_{iK})'$ is then expressed as

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

where Y_{ij} is the response of the i^{th} examinee on the j^{th} item, and q_{jk} is the Q-matrix entry corresponding to the j^{th} item and k^{th} skill. As we can see, each element in $\boldsymbol{W}$ corresponds to a sum score on a certain skill. In the next step, the vector $\boldsymbol{W}$ is taken as the input to a user-chosen method of cluster analysis, with a fixed number of 2^K clusters. The methods of cluster analysis we have investigated for this application are *K*-means and HACA with a variety of linkages. Supported by the asymptotic classification theory, which we introduce later, HACA with complete linkage and other linkages can classify data accurately and consistently to correct groups under some convergence assumptions.

In their study, Chiu and colleagues (in press) state three lemmas and develop a formal theory that justifies the application of cluster analysis with the system-scores as input. For details of the proofs, please refer to Chiu et al. (in press).

ANALYSIS OF PIRLS DATA

Data

We used the PIRLS 2001 reading assessment data for our empirical application. The PIRLS reading assessment aimed to measure the progress of 9- to 10-year-old children on reading literacy. The reading comprehension test consisted of eight sets of blocks, based on different reading passages (Table 1). Each block included 11 to 14 items, and a booklet was formed by including two blocks. In each booklet, two item formats, multiple-choice and constructed-response, were used, and each format contained both dichotomous and polytomous responses. Because the classification theory focused on analyzing dichotomous responses, our analysis included only items of dichotomous responses. Table 1 summarizes the booklet contents and item characteristics.

Table 1: Characteristics of PIRLS 2001 data by test booklet

Booklet	Topic	Required skill	Number of items			
			MC [a]	CR [b]	Dichotomous[c]	Total
1	Antarctica	Acquire and use information	4	7	7	11
2	Leonardo	Acquire and use information	6	6	8	12
3	Pufflings	Acquire and use information	8	5	10	13
4	River	Acquire and use information	3	8	6	11
Subtotal			*21*	*26*	*31*	*47*
5	Clay	Literary experience	6	7	10	13
6	Flower	Literary experience	7	6	10	13
7	Hare	Literary experience	5	6	7	11
8	Mice	Literary experience	7	7	12	14
Subtotal			*25*	*26*	*39*	*51*
Total			*46*	*52*	*70*	*98*

Note: [a] MC = multiple-choice items, [b] CR = constructed-response items; [c] dichotomous: dichotomously scored items included both multiple-choice and constructed-response items.

In 2001, a total of 146,490 students from 35 countries took the test; each booklet was administered to about 25% of the examinees. To apply the classification method, examinees had to take common items. Therefore, examinees were grouped according to which booklets they took, and were classified within the group. Table 2 presents the distribution of examinees based on blocks/booklets.

Table 2: Distribution of examinees according to blocks taken

		Block								Total	Total
Group ID	*Size*	*1 (*L1)*	*2 (L2)*	*3 (L4)*	*4 (L3)*	*5 (**I2)*	*6 (I1)*	*7 (I4)*	*8 (I3)*	*(Items used)*	*(Items)*
A	10,697	x	x							*15*	*23*
***B	11,090	x					x			*17*	*24*
***C	11,446	x							x	*19*	*25*
D	10,943		x		x					*14*	*23*
***E	10,787		x			x				*18*	*25*
***F	29,551			x				x		*17*	*24*
***G	11,343				x		x			*16*	*24*
***H	11,409				x				x	*18*	*25*
I	10,561					x	x			*20*	*26*
J	10,875					x			x	*22*	*27*
Total	*128,702*	*33,233*	*32,427*	*29,551*	*32,379*	*32,223*	*32,994*	*29,551*	*33,730*		

Note: * L = the skill of literary experience, ** = the skill of acquire and use information, *** = the groups taking booklets where combined items required both skills (i.e., acquire and use information and literary experience).

We can see from Table 2 that there were about 10,000 examinees from across the 35 countries in each group, with the exception of Group F, where about 30,000 examinees were identified. We acknowledge that examinees of different backgrounds may interpret or answer items differently. If item bias does not exist, the above differentiations can be seen as due solely to the differences in examinees' cognitive capability based on the assumption of local independence.

Although differential item functioning (DIF) studies on PIRLS data are currently unavailable, we looked at the effect of two possible factors of item bias evident in the literature of large-scale assessments, namely gender and language. As Mullis, Martin, Gonzalez, and Kennedy (2003) and Twist, Sainsbury, Woodthorpe, and Whetton (2003) record, girls outperformed boys in all participating countries in PIRLS. However, in terms of item response functions, we can again consider this as a difference in cognitive ability that will not produce influential noise in the clustering algorithm. In reporting their study on test equality, Whetton and Twist (2003) specifically pointed out that English is regarded as a deep orthography, more inconsistent and complex than some other European languages. Although we were still uncertain as to whether the differences would cause item bias, we decided to eliminate the possibility by conducting our analysis on data from the English- and European-speaking countries so that the samples would be more homogeneous in terms of language spoken by the examinees.

Because every student took items from only one booklet, the dataset was very sparse. At the current developmental stage of our new classification method, we do not yet have a fully available mechanism for dealing with missing data due to administration design. But because these missing data are not missing at random, imputation techniques are inappropriate for filling up the incomplete spots. Suspecting that responses were missing due to reasons other than administration method, we excluded the affected cases from subsequent analysis. Table 3 indicates the number of cases that we used for our analysis and the deletion rate for each dataset.

Table 3: Sample sizes of datasets used for analysis, and missing data deletion rates

		Sample size		
Group	*Language*	*Before deletion*	*After deletion*	*Deletion rate (%)*
Group B	English	2,569	2,419	5.8
	European	6,434	6,091	5.3
Group C	English	2,546	2,420	4.9
	European	6,549	6,286	4.0
Group E	English	2,552	2,400	6.0
	European	6,489	5,770	11.1
Group F	English	7,060	6,638	6.0
	European	18,749	16,734 (6,000 were used)	10.7
Group G	English	2,602	2,475	4.9
	European	6,518	6,220	4.6
Group H	English	2,544	2,423	4.8
	European	6,509	6,262	3.8

Preparation of the Q-matrices

The items of the PIRLS assessment were written on the basis of two purposes for reading, across four processes of comprehension. Although this specific structure can be used to construct the Q-matrices by taking the two purposes, the four processes, or (possibly) the eight crossed blocks composed by the two purposes and the four processes as required skills, there are limitations on Q-matrix construction for the PIRLS data. Let us take an assessment of four skills as an example.

A test comprising four required skills means $2^4 = 16$ possible examinees' attribute patterns in the data. Although some clusters are likely to be empty, a more reasonable assumption to maintain when dealing with large-scale data of large sample size is that all possible clusters are non-empty so that the number of misclassified examinees can be minimized. However, under the PIRLS' design, every examinee took, on average, 24.6 items, as shown in Table 2. Of these items, only 17.6 items were dichotomous. If we form 16 clusters with 17.6 items, using cluster analysis, it is likely that we will obtain unreliable and inconsistent results (Chiu et al., in press). Furthermore, as shown in the study by Chiu and colleagues, if we want to identify all possible examinees' attribute patterns, the test needs to include all possible single-skill items. In other words, items of patterns (1 0 0 0), (0 1 0 0), (0 0 1 0), and (0 0 0 1) are needed. Because, on average, there were only 17.6 items in a booklet, it is possible that the single-skill items for particular skills would be missing. As a result, the test would not be able to identify examinees of certain attribute patterns, and it is likely that these examinees would be misclassified to a wrong cluster. This is an especially likely occurrence with reading assessments because more than one skill is usually required to answer an item correctly. Given this situation, we decided, when applying the new classification method to the PIRLS data, to take only the "purpose" information to form two skills. The reading purpose of each block was indicated in Table 1 above; the Q-matrix for the dataset with examinees taking Blocks 1 and 6 is listed in Table 4.

Taking as our basis the booklets that the examinees took, we divided the data into 10 groups, as shown in Table 2 above. However, as we mentioned in the previous section, if a test is to identify all possible examinees' attribute patterns, all possible single-skill items have to be included. Only six sets of items covered both required skills of *literary experience* and *acquire and use information*, and these, as indicated in Table 2, were the ones we accordingly used in our analysis.

Table 4: The Q-matrix for the dataset containing examinees taking items in Blocks 1 and 6, by assessment purpose

Block	Item	Purpose	Q-matrix	
1	1	Acquire and use information	1	0
1	2	Acquire and use information	1	0
1	3	Acquire and use information	1	0
1	5	Acquire and use information	1	0
1	6	Acquire and use information	1	0
1	10	Acquire and use information	1	0
1	11	Acquire and use information	1	0
6	61	Acquire and use information	0	1
6	62	Literary experience	0	1
6	63	Literary experience	0	1
6	64	Literary experience	0	1
6	65	Literary experience	0	1
6	66	Literary experience	0	1
6	68	Literary experience	0	1
6	70	Literary experience	0	1
6	71	Literary experience	0	1
6	73	Literary experience	0	1

Procedures

We carried out the classification analysis according to the following procedures. First, we constructed sum-scores vectors, ***W***, from the data and used these as an input to the clustering algorithms. Next, we used *K*-means and HACA with complete, average, and Ward's linkages to classify examinees. We then fitted the DINA model using EM algorithm for parameter estimation. Note that we did not take the DINA model as the true underlying model of the data in the analyses, but rather used it to compare the outcomes of the cluster analysis and model-based method.

Evaluation

Cluster size, within-cluster mean of ***W***, and within-cluster sum of squares (WCSS) of ***W*** were the indices we used to evaluate the quality of classification for each method. The within-cluster mean of ***W*** indicates how well the examinees' patterns within a cluster have been identified, in the sense that the means, when taken as vectors, should be quite distinct across the possible clusters. If examinees in a particular cluster have the same attribute pattern, mean ***W*** should have a pattern of relatively large value(s) on the dimension(s) of 1's, and much smaller value(s) on 0's. If there are misclassified examinees in a cluster, the expected pattern of ***W*** is likely to become unclear, thus allowing examination of the classification quality. WCSS provides us with a sense of the extent to which grouping the data in a certain way explains the variability in the dataset. An adequate classification should yield separate clusters,

each of which is formed compactly. This requirement implies that WCSS should be small for a well-classified cluster.

In order to study the interrelationships between methods, we applied the adjusted Rand index (ARI) to indicate the agreement between classifications. Let $\{g_i\}_{i=1}^{G}$ and $\{h_j\}_{j=1}^{H}$ be two partitions of N objects. Denote $N_{i\bullet}$ and N_j as the numbers of objects classified into clusters g_i and h_j, respectively, and N_{ij} as the number of objects classified into both cluster g_i and cluster h_j. The ARI is then defined as

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

Note that the *ARI* ranges from 0 to 1 and does not require equal numbers of clusters. We conducted, for each group, the analysis with full data. The only exception was the group taking Blocks L4 and I4. Data-mining left 16,734 examinees across the European countries, but this process ran out of the memory limit set up by R (the software we used for analyses). We thus reduced the sample size by randomly drawing 6,000 examinees out of the total 16,734 so that the sample size was about the same as the sample sizes of the other groups. In regard to the organization of the outputs, we labeled clusters formed by running DINA-EM with the attribute pattern that maximized the posterior likelihood. Although *K*-means and HACA algorithms did not directly provide labels for clusters, we sorted the results along with the sum of the raw scores, taking advantage of the feature whereby patterns were partially ordered among the four attribute patterns.

RESULTS

The data were stratified into English-speaking countries, including Belize, Canada (Ontario and Quebec), England, New Zealand, Scotland, Singapore, and the United States, and European-speaking countries, including Bulgaria, Cyprus, the Czech Republic, France, Germany, Greece, Hungary, Iceland, Italy, Latvia, Lithuania, Macedonia, the Netherlands, Norway, Romania, the Russian Federation, the Slovak Republic, Slovenia, Sweden, and Turkey. Both language sets were fitted by the DINA model and analyzed by HACA with various linkages and *K*-means through the statistics ***W***. In this design, we did not take DINA as the assumed model, but as a contrast in order to show what results we could obtain, and whether it is beneficial to analyze the data using cluster analysis, in a situation where we have little information about the true model and where we choose DINA as the fitted model.

Table 5 displays the classification results for the data from Group B, which took items from Blocks 1 and 6. In this set of items, seven items required the skill of acquiring and using information, and the other 10 items required the skill of literacy experience. According to the results under the mean ***W*** category, HACA with complete and average linkages provided clear and interpretable patterns for both languages. DINA-EM, however, produced larger mean ***W*** values for the cluster of pattern (0 0) than did the other clustering methods, a finding which demonstrates that DINA tended

Table 5: Classification results for Group B (took Blocks 1 and 6)

(A) Cluster analysis									
	English-speaking					*European-speaking*			
Size	*Mean W*		*WCSS (total)*	*MSum Y*	*Size*	*Mean W*		*WCSS (total)*	*MSumY*
HACA with complete linkage					HACA with complete linkage				
286	2.00	3.00	1,089.00	5.00	149	1.23	2.15	207.69	3.38
223	4.75	4.23	387.27	8.98	201	1.56	5.44	309.18	7.00
481	3.99	8.20	1,220.42	12.20	1641	4.59	5.14	5650.57	9.72
1,429	6.28	8.50	2,922.59	14.78	4100	5.90	8.60	8,979.12	14.49
			(5,619.28)					*(15,146.56)*	
HACA with average linkage					HACA with average linkage				
136	1.46	1.68	239.50	3.13	393	2.34	3.15	1,388.94	5.49
540	3.69	4.96	2,034.61	8.65	25	5.04	1.60	20.96	6.64
2	2.00	9.50	0.50	11.5	1,353	4.58	5.51	4,672.18	10.09
1,741	5.94	8.60	4,259.74	14.53	4,220	5.86	8.54	9,363.41	14.40
			(6,534.35)					*(15,445.49)*	
HACA with Ward linkage					HACA with Ward linkage				
564	3.08	3.91	2777.29	6.99	621	3.17	3.02	2169.31	6.19
615	5.11	6.99	1576.03	12.09	1654	4.23	6.24	6076.90	10.47
545	5.37	9.42	529.27	14.79	2148	5.72	7.99	3266.31	13.71
695	6.81	9.00	579.78	15.81	1668	6.56	9.44	820.89	16.00
			(5,462.37)					*(12,333.41)*	
K-means					*K*-means				
Size	*Mean W*		*WCSS (total)*	*MSum Y*	*Size*	*Mean W*		*WCSS (total)*	*MSumY*
354	2.40	3.00	1,297.63	5.40	830	3.13	3.53	2,962.96	6.66
508	5.50	6.04	948.36	11.53	905	3.36	7.34	1,737.15	10.70
408	3.87	7.89	732.84	11.76	1284	5.87	6.26	1,742.01	12.13
1,149	6.36	9.18	1,189.44	15.54	3072	6.19	8.95	3,794.63	15.14
			(4,168.27)					*(10,236.75)*	

(B) DINA-EM									
	English-speaking					*European-speaking*			
Pattern	*Size*	*Mean W*		*WCSS (total)*	*Pattern*	*Size*	*Mean W*		*WCSS (total)*
(0 0)	583	2.97	4.07	3,014.90	(0 0)	1,521	3.36	4.83	6,985.22
(1 0)	46	6.26	4.50	26.37	(1 0)	198	6.23	4.55	218.40
(0 1)	47	2.70	8.45	45.45	(0 1)	161	2.94	8.53	125.44
(1 1)	1,743	5.96	8.56	4,214.29	(1 1)	4,211	6.03	8.42	9,268.02
				(7,301.01)					*(16,597.1)*

(C) ARI across all selected methods										
	English-speaking					*European-speaking*				
	DINA	*Comp*	*Ave*	*Ward*	*K-means*	*DINA*	*Comp*	*Ave*	*Ward*	*K-means*
DINA	1	0.27	0.81	0.28	0.37	1	0.67	0.64	0.31	0.47
Comp	*	1	0.59	0.24	0.41	*	1	0.82	0.38	0.46
Ave	*	*	1	0.22	0.36	*	*	1	0.37	0.45
Ward	*	*	*	1	0.48	*	*	*	1	0.39
K-means	*	*	*	*	1	*	*	*	*	1

to misclassify data to the lowest cluster, for both languages. This finding implies that cluster analysis is more robust than the model-based method, even with such a short test. However, HACA with Ward linkage and *K*-means do not yield recognizable ***W*** patterns for some clusters, a happenstance which signals that examinees of a certain attribute pattern were misclassified to a wrong cluster.

In terms of within-cluster variability, DINA-EM produced clusters of larger variability than did the other methods for both the English and the European data, indicating that the clusters formed by using DINA-EM were not as homongenous as those formed through use of the other methods. Note that the WCSS should be interpreted with caution. As mentioned, HACA with Ward linkage and *K*-means tend to minimize the sum of within-cluster variances, and therefore tend to produce tight clusters of small WCSS.

The ARI portion of Table 5 indicates that HACA with complete and average linkages were in high agreement for both datasets. However, HACA with complete linkage had a high agreement with DINA-EM for the European data but a low agreement for the English data. Note that a large ARI does not necessarily imply good classification quality. Good quality can only be assumed when one of the compared partitions is known to perform well. A large ARI may simply reflect that both compared partitions behave similarly with some particular data structure.

Table 6 concludes the results for Group C, which took Blocks 1 and 8. Seven of the items required the skill of acquiring and using information; 12 required the skill of literacy experience. For the English data, HACA with average linkage performed better than the other methods. For the European data, HACA with complete and average linkages both performed well, based on their mean ***W*** patterns. DINA not only again clustered most examinees to the two ends of the classes with large mean ***W*** values for the lowest class, but also yielded much larger WCSSs than the other methods. The ARI portion of the table shows that, for the European data, HACA with average and complete linkages classified data in high agreement. What is evident, in addition to these two methods generating a well-recognized mean ***W*** pattern (see the information provided above), is that these two methods can produce consistent classifications with this group of European data.

Table 7 shows the results for Group E, which took Blocks 2 and 5, where eight of the 18 items required the skill of acquiring and using information and 10 required the skill of literacy experience. For both the English and the European data, DINA-EM performed well, producing recognizable mean ***W*** patterns. HACA with average linkage and *K*-means also performed quite well. However, DINA-EM still had the problem of large WCSS. The ARI portion of the table does not show any method providing high agreement.

Table 8 displays the results for Group F, which took Blocks 3 and 7. Ten of the items required the skill of acquiring and using information, and seven required the skill of literacy experience. As discussed, only 6,000 European data were randomly drawn and analyzed. In this case, HACA with complete linkage performed better than

Table 6: Classification results for Group C (took Blocks 1 and 8)

(A) Cluster analysis									
	English-speaking					*European-speaking*			
Size	*Mean W*		*WCSS (total)*	*MSum Y*	*Size*	*Mean W*		*WCSS (total)*	*MSumY*
HACA with complete linkage					HACA with complete linkage				
208	1.60	3.11	677.75	4.07	365	1.90	3.56	960.43	5.46
93	3.60	1.96	206.11	5.56	558	2.87	7.10	1,089.68	9.97
574	5.28	6.60	1,438.75	11.88	1248	5.17	5.62	4,162.54	15.68
1,545	5.59	10.05	6,050.10	15.64	4115	5.93	9.75	10,420.02	10.79
			(8,372.71)					*(16,632.67)*	
HACA with average linkage					HACA with average linkage				
235	2.15	2.19	818.17	4.34	470	2.17	3.33	1,373.60	5.51
542	4.29	5.87	1,527.84	10.15	228	2.34	8.45	555.48	10.79
11	1.82	9.36	6.18	11.18	1516	4.83	6.09	4,236.05	10.91
1,632	5.81	10.01	4,769.02	15.83	4072	5.97	9.74	9,976.01	15.71
			(7,121.21)					*(16,141.14)*	
HACA with Ward linkage					HACA with Ward linkage				
229	1.93	2.29	701.41	4.23	1,077	3.06	4.42	5,083.81	7.49
700	4.64	6.25	2,515.52	10.89	1,312	2.87	7.07	2,120.32	11.95
985	5.41	9.61	2,377.17	15.02	1,955	5.60	8.92	4,337.04	14.52
506	6.55	11.42	248.44	17.97	1,947	6.45	10.76	1,814.54	17.21
			(5,842.54)					*(13,355.71)*	
K-means					*K*-means				
Size	*Mean W*		*WCSS (total)*	*MSum Y*	*Size*	*Mean W*		*WCSS (total)*	*MSumY*
275	2.16	2.53	1,031.45	4.69	810	2.75	3.98	3,276.68	6.73
510	4.41	6.00	1,325.34	10.41	1,260	4.04	8.33	2,251.59	12.37
738	5.24	8.85	1,439.00	14.09	1,358	5.76	6.79	2,505.92	12.56
897	6.27	10.99	1,079.63	17.25	2,858	6.30	10.30	4,232.76	16.61
			(4,875.42)					*(12,266.95)*	

(B) DINA-EM									
	English-speaking					*European-speaking*			
Pattern	*Size*	*Mean W*		*WCSS (total)*	*Pattern*	*Size*	*Mean W*		*WCSS (total)*
(0 0)	635	3.03	4.49	4,009.23	(0 0)	1,712	3.46	5.47	9,039.31
(1 0)	51	6.14	5.14	54.08	(1 0)	148	6.26	5.22	213.80
(0 1)	24	2.54	9.58	25.79	(0 1)	96	2.70	9.61	82.98
(1 1)	1710	5.87	9.82	5,484.36	(1 1)	4330	6.02	9.55	1,2215.95
				(9,573.46)					*(21,552.04)*

(C) ARI across all selected methods										
	English-speaking					*European-speaking*				
	DINA	*Comp*	*Ave*	*Ward*	*K-means*	*DINA*	*Comp*	*Ave*	*Ward*	*K-means*
DINA	1	0.40	0.69	0.24	0.34	1	0.65	0.67	0.31	0.35
Comp	*	1	0.60	0.45	0.38	*	1	0.92	0.32	0.46
Ave	*	*	1	0.43	0.52	*	*	1	0.33	0.45
Ward	*	*	*	1	0.47	*	*	*	1	0.43
K-means	*	*	*	*	1	*	*	*	*	1

Table 7: Classification results for Group E (took Blocks 2 and 5)

(A) Cluster analysis									
	English-speaking					*European-speaking*			
Size	*Mean W*		*WCSS (total)*	*MSum Y*	*Size*	*Mean W*		*WCSS (total)*	*MSumY*
HACA with complete linkage					HACA with complete linkage				
351	1.83	1.93	820.76	3.76	882	2.73	2.73	2,211.83	5.47
390	4.48	2.89	938.41	7.37	1,624	4.98	4.80	6,706.74	9.77
744	4.66	7.28	2,537.81	11.94	1,100	3.88	7.74	3,647.03	11.62
915	7.01	7.94	2,994.28	17.94	2,164	6.70	8.12	4,402.68	14.82
			(7,291.26)					*(16,968.28)*	
HACA with average linkage					HACA with average linkage				
609	2.77	2.20	2,004.71	4.97	610	2.54	1.81	1593.25	4.35
372	4.04	5.97	690.15	10.01	4	0.75	8.50	1.75	9.25
214	6.01	4.01	334.94	10.02	2,496	4.59	5.08	9,490.90	9.67
1,205	6.47	8.36	3,385.94	14.82	2,660	6.11	8.45	6,571.69	14.56
			(6,415.74)					*(17,657.59)*	
HACA with Ward linkage					HACA with Ward linkage				
506	2.42	2.02	1,641.36	4.45	1789	3.93	3.15	7,625.00	7.08
555	5.05	4.45	1,466.67	9.50	1290	3.85	6.79	2,958.54	10.64
737	5.56	7.32	2,105.11	12.88	706	6.55	6.43	548.29	12.98
602	6.88	9.30	805.08	16.18	1985	6.38	8.74	3202.37	15.11
			(6,018.22)					*(14,334.2)*	
K-means					*K*-means				
Size	*Mean W*		*WCSS (total)*	*MSum Y*	*Size*	*Mean W*		*WCSS (total)*	*MSumY*
585	2.58	2.20	1,890.62	4.79	1317	3.19	2.74	4,449.50	5.93
536	5.53	4.76	1,455.84	10.30	1114	5.64	5.11	2,064.73	10.75
436	4.49	7.47	903.53	11.96	1210	4.01	7.32	2,245.62	11.33
843	6.92	8.76	1,327.98	15.68	2129	6.54	8.51	3,700.89	15.05
			(5,577.97)					*(12,460.74)*	

(B) DINA-EM									
	English-speaking					*European-speaking*			
Pattern	*Size*	*Mean W*		*WCSS (total)*	*Pattern*	*Size*	*Mean W*		*WCSS (total)*
(0 0)	855	3.17	2.90	3,957.36	(0 0)	1,905	3.57	3.42	8,349.24
(1 0)	44	6.70	3.09	88.80	(1 0)	48	7.13	3.33	45.92
(0 1)	33	2.61	7.64	31.52	(0 1)	192	2.69	7.52	255.17
(1 1)	1468	6.25	7.91	5,341.40	(1 1)	3,625	5.96	7.77	11,899.17
				(9,419.08)					*(20,549.5)*

(C) ARI across all selected methods										
	English-speaking					*European-speaking*				
	DINA	*Comp*	*Ave*	*Ward*	*K-means*	*DINA*	*Comp*	*Ave*	*Ward*	*K-means*
DINA	1	0.39	0.59	0.33	0.38	1	0.30	0.35	0.37	0.37
Comp	*	1	0.47	0.29	0.36	*	1	0.41	0.41	0.54
Ave	*	*	1	0.41	0.55	*	*	1	0.42	0.48
Ward	*	*	*	1	0.58	*	*	*	1	0.56
K-means	*	*	*	*	1	*	*	*	*	1

Table 8: Classification results for Group F (took Blocks 3 and 7)

(A) Cluster analysis									
	English-speaking					*European-speaking*			
Size	*Mean W*		*WCSS (total)*	*MSum Y*	*Size*	*Mean W*		*WCSS (total)*	*MSumY*
HACA with complete linkage					HACA with complete linkage				
496	2.07	1.89	1,116.48	3.95	471	2.70	2.61	1,064.28	5.31
732	2.27	5.22	1,246.84	7.49	462	2.20	5.24	989.04	7.44
2,163	4.81	5.85	5,267.41	10.66	203	6.47	3.17	415.51	9.64
3,247	7.81	6.35	6,926.22	14.15	4864	6.74	6.14	18,995.98	12.88
			(14,556.95)					*(21,464.81)*	
HACA with average linkage					HACA with average linkage				
481	1.96	1.92	927.40	3.89	281	1.63	2.43	1,052.52	4.05
230	5.39	2.70	578.68	8.10	1	7.00	0.00	0.00	7.00
1,325	3.17	5.13	3,402.09	8.30	2,931	4.59	5.47	9,648.06	10.05
4,602	6.97	6.45	15,136.72	13.41	2,787	8.07	6.26	4,864.93	14.33
			(20,044.89)					*(15,565.51)*	
HACA with Ward linkage					HACA with Ward linkage				
933	2.37	2.78	3,143.42	5.15	1,122	2.84	3.80	4,912.55	6.65
1,973	4.05	5.91	3,908.66	9.95	2,096	5.13	5.94	3,371.26	11.07
1,931	6.52	6.14	2,673.32	12.66	1,901	7.50	6.16	2,169.12	13.65
1,801	8.70	6.61	1,660.29	15.31	881	9.31	6.51	551.48	15.81
			(11,385.69)					*(11,004.41)*	
K-means					*K*-means				
Size	*Mean W*		*WCSS (total)*	*MSum Y*	*Size*	*Mean W*		*WCSS (total)*	*MSumY*
855	2.47	2.54	2,731.26	5.02	824	3.27	2.95	2,860.88	6.22
1,222	3.23	5.83	2,127.72	9.06	895	3.25	5.90	1,230.97	9.15
2,741	6.04	6.11	4,394.48	12.15	2,474	6.13	6.06	3,560.68	12.19
1,820	8.70	6.58	1,838.26	15.28	1,807	8.65	6.33	2,213.76	17.98
			(11,091.72)					*(9,866.29)*	

(B) DINA-EM									
	English-speaking					*European-speaking*			
Pattern	*Size*	*Mean W*		*WCSS (total)*	*Pattern*	*Size*	*Mean W*		*WCSS (total)*
(0 0)	1602	3.26	3.49	7,128.80	(0 0)	1401	3.69	3.72	5,722.63
(1 0)	7	8.57	2.86	2.57	(1 0)	12	8.25	2.83	3.92
(0 1)	976	3.55	6.44	1,507.73	(0 1)	842	3.81	6.39	1,167.00
(1 1)	4,053	7.33	6.44	10,513.01	(1 1)	3745	7.45	6.28	9,124.10
				(19,152.11)					*(16,017.65)*

(C) ARI across all selected methods										
	English-speaking					*European-speaking*				
	DINA	*Comp*	*Ave*	*Ward*	*K-means*	*DINA*	*Comp*	*Ave*	*Ward*	*K-means*
DINA	1	0.48	0.61	0.34	0.38	1	0.37	0.32	0.24	0.37
Comp	*	1	0.43	0.44	0.35	*	1	0.13	0.21	0.23
Ave	*	*	1	0.24	0.31	*	*	1	0.57	0.30
Ward	*	*	*	1	0.68	*	*	*	1	0.37
K-means	*	*	*	*	1	*	*	*	*	1

the other cluster methods for the European data, and HACA with average linkage performed best for the English data. The above notion was based on whether the mean ***W*** patterns were more recognizable than the other mean ***W*** patterns. However, the clusters formed by these methods contained large WCSSs, implying that misclassification occurs with the two methods when clustering data. If we take a closer look, it is not hard to find that the mean ***W*** values of the highest class produced by the two methods were smaller than those produced by the other methods. This outcome means that many data with patterns different from (1 1) were classified into the (1 1) cluster. Although the mean ***W*** values provided useful information for identifying the underlying pattern for each cluster, that information was not sufficient to allow us to determine if a particular method outperformed the others.

Table 9 shows the results of Group G, which took Blocks 4 and 6. In this set, the first six items required the skill of acquiring and using information, and the last 10 items required the skill of literacy experience. The results imply that, for both datasets, DINA-EM is a better choice than the others. HACA with average linkage produced acceptable classifications, although the mean ***W*** patterns were not very clear. An interesting finding is that HACA with average linkage formed clusters with unusually large WCSSs for both datasets, implying that HACA with average linkage was unable to classify examinees into the correct cluster with this particular data structure, rendering the results untrustworthy. As the ARI portion of the table shows, DINA-EM had good agreement with HACA, with average linkage for the English data, and a good agreement with HACA, but with complete linkage for the European data.

In Table 10, examinees took Blocks 4 and 8. Here, six items required acquiring and using information, while the last 12 items required literacy experience. Among the cluster analysis methods, HACA with complete linkage produced tighter and better separated means of ***W*** than the other methods for the European data. DINA-EM produced recognizable mean ***W*** patterns, but the issue of misclassification remained. In addition, HACA with average and complete linkages behaved similarly with both the English and the European data, based on the ARI index.

CONCLUSIONS AND IMPLICATIONS

Given that PIRLS was not designed to detect examinees' cognitive capacity, use of the retrofitting approach to analyze data is usually a great concern. As Gierl (2007) points out, using retrofitting procedures to pursue cognitive analysis for existing testing invariably produces weak fit between the cognitive model and the data. The main reason why is that the assessments were not designed for the cognitive diagnostic purpose and so could not provide much useful information. This area of research is an important and developing one because, at the current stage, applying retrofitting procedures to analyze large-scale data seems unavoidable. Under this circumstance, where the parametric approach encounters the issue of model fit, a useful direction for overcoming the limitation is to develop robust non-parametric or quasi-parametric approaches as alternatives to minimize the bias due to model misfit. Motivated by this

Table 9: Classification results for Group G (took Booklets 4 and 6)

(A) Cluster analysis									
	English-speaking					*European-speaking*			
Size	*Mean W*		*WCSS (total)*	*MSum Y*	*Size*	*Mean W*		*WCSS (total)*	*MSumY*
HACA with complete linkage					HACA with complete linkage				
366	1.68	2.87	1,198.20	4.55	338	1.33	2.30	688.04	3.63
281	4.56	5.14	425.00	9.71	1,120	3.13	4.88	2,486.86	8.02
731	3.32	7.78	1,844.10	11.09	407	1.97	8.40	1,289.37	10.37
1,097	5.49	8.91	1,336.61	14.40	4,355	4.81	8.32	11,455.33	13.12
			(4,803.91)					*(15,919.6)*	
HACA with average linkage					HACA with average linkage				
192	1.38	1.72	380.29	3.10	12	3.83	0.58	6.58	4.42
1	6.00	2.00	0.00	8.00	827	2.23	3.34	2,683.04	5.57
644	3.21	5.24	2,356.18	8.44	155	0.77	6.85	197.92	7.61
1,638	4.89	8.70	3,409.21	13.59	5,226	4.53	8.05	18,651.39	12.58
			(6,145.68)					*(21,538.93)*	
HACA with Ward linkage					HACA with Ward linkage				
243	1.24	2.29	7,18.51	3.53	2,194	2.87	5.04	11,230.79	7.91
504	3.59	4.96	1,078.27	8.55	1,146	3.53	8.82	1,170.44	12.34
766	4.30	7.56	1,628.35	11.86	1,270	5.10	7.50	929.40	12.60
962	5.14	9.47	1,017.02	14.61	1,610	5.52	9.44	797.52	14.96
			(4,442.15)					*(14,128.15)*	
K-means					*K*-means				
Size	*Mean W*		*WCSS (total)*	*MSum Y*	*Size*	*Mean W*		*WCSS (total)*	*MSumY*
301	1.53	2.48	914.12	4.01	926	2.24	3.48	3,024.07	5.72
471	3.77	5.23	951.38	9.00	1354	2.74	7.49	2,615.41	10.23
621	3.58	7.88	1,061.11	11.46	1200	4.74	6.30	3,226.02	11.04
1,082	5.44	9.11	1,202.74	14.55	2740	5.19	9.11	1,533.00	14.30
			(4,129.35)					*(10,398.5)*	

(B) DINA-EM									
	English-speaking					*European-speaking*			
Pattern	*Size*	*Mean W*		*WCSS (total)*	*Pattern*	*Size*	*Mean W*		*WCSS (total)*
(0 0)	616	2.26	3.87	2,904.95	(0 0)	1665	2.46	4.73	7,244.48
(1 0)	40	5.18	4.40	33.38	(1 0)	121	5.30	4.30	140.58
(0 1)	53	1.70	8.13	41.25	(0 1)	255	2.11	8.57	288.55
(1 1)	1,766	4.90	8.48	4451.05	(1 1)	4,179	4.89	8.45	9,539.33
				(7,430.63)					*(17,212.94)*

(C) ARI across all selected methods										
	English-speaking					*European-speaking*				
	DINA	*Comp*	*Ave*	*Ward*	*K-means*	*DINA*	*Comp*	*Ave*	*Ward*	*K-means*
DINA	1	0.34	0.70	0.33	0.37	1	0.64	0.42	0.27	0.37
Comp	*	1	0.44	0.35	0.71	*	1	0.53	0.13	0.31
Ave	*	*	1	0.48	0.52	*	*	1	0.02	0.23
Ward	*	*	*	1	0.53	*	*	*	1	0.33
K-means	*	*	*	*	1	*	*	*	*	1

Table 10: Classification results for Group H (took Booklets 4 and 8)

(A) Cluster analysis									
	English-speaking					*European-speaking*			
Size	*Mean W*		*WCSS (total)*	*MSum Y*	*Size*	*Mean W*		*WCSS (total)*	*MSumY*
HACA with complete linkage					HACA with complete linkage				
250	1.16	2.11	576.46	3.27	331	1.64	1.99	745.89	3.63
73	3.70	2.63	86.38	6.33	479	0.32	5.69	1,346.04	7.01
642	4.00	6.38	2,232.49	10.38	643	4.03	5.17	1,232.44	9.20
1,458	4.68	10.16	4,647.09	14.84	4809	4.60	9.27	17,133.88	13.87
			(7,542.42)					*(20,458.25)*	
HACA with average linkage					HACA with average linkage				
275	1.73	1.90	654.89	3.63	331	1.64	1.99	745.89	3.63
1	0.00	10.00	0.00	10.00	1,236	2.14	6.62	3,587.10	8.76
820	3.66	6.41	3,230.45	10.08	474	4.40	5.02	905.90	9.41
1,327	4.87	10.43	2,948.49	15.31	4,221	4.88	9.49	12,231.15	14.37
			(6,833.83)					*(17,470.04)*	
HACA with Ward linkage					HACA with Ward linkage				
454	1.91	2.96	1887.13	4.86	1679	2.60	4.95	8791.82	7.55
372	4.04	6.22	552.67	10.26	1619	3.32	9.13	2982.47	12.46
1,068	4.39	9.28	2,625.44	13.68	775	5.34	7.32	644.98	12.66
529	5.47	11.46	339.03	16.93	2189	5.48	10.28	2,775.63	15.76
			(5,404.27)					*(15,194.90)*	
K-means					*K*-means				
Size	*Mean W*		*WCSS (total)*	*MSum Y*	*Size*	*Mean W*		*WCSS (total)*	*MSumY*
329	1.65	2.26	909.35	3.91	849	2.25	3.42	3,258.60	5.67
512	3.52	5.90	1,299.72	9.42	1,103	2.35	7.61	2,106.76	9.96
731	4.27	8.86	1,401.50	13.13	2,184	4.87	7.93	3,846.41	12.79
851	5.27	11.03	1,045.37	16.30	2,126	5.06	10.66	3,043.66	15.72
			(4,655.94)					*(12,255.43)*	

(B) DINA-EM									
	English-speaking					*European-speaking*			
Pattern	*Size*	*Mean W*		*WCSS (total)*	*Pattern*	*Size*	*Mean W*		*WCSS (total)*
(0 0)	628	2.20	3.86	3,452.19	(0 0)	1,484	2.30	4.87	7,557.61
(1 0)	38	5.16	4.50	56.55	(1 0)	120	5.25	4.73	234.43
(0 1)	33	1.70	9.39	48.85	(0 1)	317	2.06	8.98	465.82
(1 1)	1,723	4.83	9.70	6,221.05	(1 1)	4,341	4.88	9.36	13,464.06
				(9,778.64)					*(21,721.92)*

(C) ARI across all selected methods										
	English-speaking					*European-speaking*				
	DINA	*Comp*	*Ave*	*Ward*	*K-means*	*DINA*	*Comp*	*Ave*	*Ward*	*K-means*
DINA	1	0.44	0.42	0.34	0.32	1	0.60	0.74	0.32	0.34
Comp	*	1	0.78	0.35	0.41	*	1	0.69	0.16	0.22
Ave	*	*	1	0.35	0.48	*	*	1	0.24	0.38
Ward	*	*	*	1	0.55	*	*	*	1	0.30
K-means	*	*	*	*	1	*	*	*	*	1

idea, we would like to learn whether the new classification method is reliable with large-scale assessment, and whether its robustness over the model-based method provides valuable benefit to remedy the weaknesses with retrofitting procedures.

The classification theory developed by Chiu and colleagues (in press) is built on the assumption of long tests. Application of this theory to the PIRLS data has provided a good opportunity for understanding how reliable the method is for short tests. This empirical application revealed that DINA-EM tends to classify data to the clusters at the two extreme ends. However, the cluster analysis seemed to have difficulty correctly assigning examinees to middle clusters. One possible explanation for the unreliable classifications is that the two procedures for reading and the four skills of comprehension were designed to nest within each other. Although we took only the two purposes as the required skills, we remain unclear as to whether examinees' wrong answers are caused purely by absence of a certain skill with respect to a purpose or by the absence of some more specific skills nested within the purpose. If the cause relates to missing skills of procedures, and not simply to missing the specific skill of the purpose, then we could assume that some measurement error is contributing to and thus inflating the systematic noise, and that the true absence or presence of the skill being measured will be difficult to detect. This consideration again reflects the importance of writing items appropriate to the cognitive diagnosis aim.

Previous studies concerning use of this cluster analysis method to classify examinees under the cognitive diagnosis setting indicate that this method, depending on the type of structure of the dataset, is more robust than is fitting a wrong model (see, for example, Chiu & Douglas, 2008). This finding is consistent with the results in the current study. When DINA-EM fits the data well (e.g., Group E or Group G), we have in place a cluster analysis that performs as well as the DINA-EM. When the DINA does not appear to be the true model (e.g., Group C and Group F), it is possible to find a cluster analysis that is more reliable. While more work is needed on labeling, given that this is critical and important for the *K*-means and HACA, the feature of easy access and the convincing results from simulations and real applications make the theory, at this point, an appreciable alternative for classification.

References

Bradley, P., & Fayyad, U. (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, MA: Morgan Kaufmann.

Chiu, C., & Douglas, J. (2008). *Cluster analysis for cognitive diagnosis: A robustness study in relation to model misspecification*. Paper presented at the annual meeting of the Psychometric Society, Durham, NH.

Chiu, C., Douglas, J., & Li, X. (in press). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*.

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

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

Gierl, M. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement, 44,* 325–340.

Gonzalez, E. I., & Kennedy, A. M. (2001). *PIRLS 2001 user guide for the international database*. Retrieved January 2007, from http://isc.bc.edu/pirls2001i/PIRLS2001_Pubs_UG.html.

Gorin, J. (2007). Test construction and diagnostic testing. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 173–201), Cambridge: Cambridge University Press.

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

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.

Leighton, J., & Gierl, M. (2007). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 3–18). Cambridge: Cambridge University Press.

MacQueen, J. (1967). Some methods of classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (pp. 281–207). Berkeley, CA: University of California Press.

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

Mullis, I., Martin, M., Gonzalez, E., & Kennedy, A. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools in 35 countries*. Chestnut Hill, MA: Boston College.

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.

Rupp, A. A., & Templin, J. L. (2007). *Unique characteristics of cognitive diagnosis models.* Paper presented at the annual meeting of the National Council for Measurement in Education, Chicago, IL.

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

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10*, 55–73.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Applied Statistics (JRSS-C), 5*, 337–350.

Twist, L., Sainsbury, M., Woodthorpe, A., & Whetton, C. (2003). *Reading all over the world: Progress in International Reading Literacy Study: National report for England*. Slough: National Foundation for Educational Research.

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

Whetton, C., & Twist, L. (2003). *What determines the range of reading attainment in a country?* Paper presented at the 29th International Association for Educational Assessment Conference, Manchester, United Kingdom.