Diagnostic cluster analysis of mathematics skills

Yoon Soo Park and Young-Sun Lee
Teachers College, Columbia University, New York, USA

Clustering and similarity trees are effective techniques for grouping and visualizing related objects; they can be implemented to assess how individuals think of psychological concepts. This study examined a method of clustering attributes required to solve mathematics problems by mapping item responses to an attribute matrix and from there conducting *K*-means clustering and hierarchical agglomerative cluster analysis (HACA). The analysis was broadened to examine how the extended similarity tree (EXTREE) algorithm (Corter & Tversky, 1986) can be used to illustrate the hierarchical and overlapping nature of the fine-grained attributes required to solve mathematics test items. Twenty-five items from the TIMSS 2007 Grade 4 mathematics test were used to generate a list of skills or attributes that together constituted a Q-matrix (Embretson, 1984; Tatsuoka, 1985). High-performing countries (Hong Kong SAR and Chinese Taipei), average-performing countries (Denmark, Sweden, and the United States), and low-performing countries (Colombia, Kuwait, Qatar, and Yemen) were selected to examine attribute-structure differences across countries, while two high-performing benchmark participants—the states of Massachusetts and Minnesota in the United States—were selected to examine attribute-structure differences within a country. Results showed that the structure of attributes in the higher-performing countries had a clearer, more hierarchical structure than the structure of attributes evident in the lower-performing countries. Examining cluster structures of attributes can thus serve as a useful method for exploring the structure of attributes and providing diagnostic feedback to policymakers and educational researchers on areas where students may need further instruction.

IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments Volume 4

INTRODUCTION

Scholars and policymakers have for a long time used broad domain-based scores from international assessments to modify and attempt to improve their countries' education systems. Using these indicators, researchers have examined measures governing the curricula and textbooks used in their countries (Hook, Bishop, & Hook, 2007; McNeely, 1997) as well as the quality of their teachers and teacher education (Rautalin & Alasuutari, 2007; Simola, 2005). Researchers have also used these indicators to conduct cross-national explorations of teaching practices and patterns of teaching (Givvin, Hiebert, Jacobs, Hollingsworth, & Gallimore, 2005; Hiebert et al., 2005), instructional methods (House, 2005), curricula and education systems (Menon, 2000), and attitudinal or instructional components (Papanastasiou, 2002). There has, however, been relatively little cross-national comparative research into how students in different countries or regions within countries respond to curricular materials.

More specifically, the outcomes of analyses designed to explore the relationship between the specific skills (or, to use a broader term, attributes) that students need to solve a particular test item and how that relationship can be confounded by other related skills provide important understandings for instructors. Often, students answer a mathematics test item incorrectly not only because of their arithmetic miscalculation, but also because they confuse the skill they need to apply to a specific learning area with the skill or skills they need for another learning area (Cai, 2007; Kuhs & Ball, 1986; Lubienski, 2000). For students, the ability to identify whether they need to apply a specific skill or set of skills with respect to a specific learning area, or whether they think they can apply the same skills or sets of skills to two or more learning areas, can serve as an important learning aid.

The Trends in Mathematics and Science Study (TIMSS) provides an opportunity to explore not only how students within a country and across countries perceive the utility of certain skills relative to a learning area, but also whether those skills link with other skills. TIMSS, which has been conducted since 1995 and had 43 participating countries in the 2007 administration at Grade 4, provides data that allows each participating country to determine the relative standing of its students' mathematics performance within an international context. The released data also offer information that policymakers and educators can use when developing measures designed to improve their education systems (Cai & Silver, 1995).

To date, most researchers have relied solely on using the overall mean proficiency scores of each country or the overall scores of students on the content and cognitive domains assessed by TIMSS to analyze the performance of the students within their own country or relative to the other participating countries. Despite the complexity and the richness of the data collected, TIMSS has been criticized for lacking studies that can be directed toward improving student performance at the attribute level. As a response to this criticism, this study extends beyond merely using overall proficiency scores to analyze student performance. Instead, we use students' response patterns on TIMSS test items to gain an understanding of which attributes, whether single or in association with others, students seem to see as necessary to solve the problem embedded within a specific item.

The skills that TIMSS assesses relate to three main content domains: number, geometric shapes and measurement, and data and display. Each domain, in turn, contains several "fine-grained" topic areas. The number domain covers, for example, whole numbers, fractions, and decimals, number sentences with whole numbers, and patterns and relationships. The geometric shapes and measurement domain focuses on lines and angles, two- and three-dimensional shapes, and location and movement. The data and display domain includes reading and interpreting, and organizing and representing. These topic areas were developed from the TIMSS framework (Mullis, Martin, Ruddock, O'Sullivan, Arora, & Erberber, 2005), such that the content areas associated with an item can be mapped onto the specific skills or attributes students need to apply when solving a particular problem.

The analytic framework of TIMSS yields an ideal platform from which to conduct multivariate methods of cluster analysis. These methods generate visualizations useful for exploring similarities and dissimilarities in data. Clustering and similarity trees are particularly effective techniques for grouping related objects, and their use can help researchers determine how individuals think with respect to psychological concepts. Because multiple attributes are required to solve a particular item, investigating how students perceive these specific attributes—whether they view them as distinct objects or cluster them with other attributes—can provide additional information useful for instructors. If subsets of attributes emerge as being grouped together, the next step is to identify which are held in common and to investigate why they are related. Additionally, if their grouping is hierarchical in nature, determining the likely reason for such a structure should also prove useful.

Attributes can be grouped into clusters using the *K*-means method (MacQueen, 1967) or formulated to explore their hierarchical nature via the hierarchical agglomerative cluster analysis (HACA; Hartigan, 1975). They can also be analyzed through use of the extended similarity tree method (Corter & Tversky, 1986). This last approach, which is also known as EXTREE, allows simultaneous examination of attributes' hierarchical and overlapping features. However, in order to incorporate attributes into the response data, the aforementioned methods of cluster analysis, unlike traditional methods of cluster analysis, require the attributes to be specified in an incidence matrix, that is, a Q-matrix (Embretson, 1984; Tatsuoka, 1985). This matrix maps the fine-grained attributes that a test-taker needs to solve a particular item correctly. Associations between the attributes can then be transformed into measures of distance and, from there, examined via the clustering and similarity trees analyses, with respect to group-related attributes and with respect to how students from a particular country utilize these attributes when problem-solving.

In this study, we analyzed 25 items from the TIMSS 2007 Grade 4 mathematics assessment in order to examine the clustering of attributes required to solve these items. We then used the *K*-means, HACA, and EXTREE methods to examine the clusters. We framed the results from the cluster analyses to answer these three questions:

1. What types of attribute clusters emerge during examination of how Grade 4 students solve mathematics problems?
2. How do these clusters of attributes differ among high-, average-, and low-performing countries and regions within the same country?
3. What are the differences and similarities with respect to how students studying within different education systems, under different curriculum configurations, and from different textbooks perceive and process the fine-grained attributes needed to answer mathematics test items?

For the comparative purposes of the study, we used the overall average test scale score for each of the participating TIMSS 2007 countries to select nine countries. These were Hong Kong SAR (henceforth, Hong Kong), which was ranked first on the international achievement scale, Chinese Taipei (Taiwan), which ranked third, the United States (13th), Denmark (15th), Sweden (24th), Colombia (37th), Kuwait (41st), Qatar (42nd), and Yemen (43rd). The first two countries represent countries where student performance was, on average, high. The next three countries represent average performance, and the final four, low performance. We then examined how the students in these countries perceived and processed the attributes prescribed in the Q-matrix. In order to compare how these clusters differed within the United States, we also drew on data from two American states that elected to participate in TIMSS 2007 as benchmarking participants.[1] The two states were Massachusetts, which ranked fourth on the international achievement scale, and Minnesota, which ranked sixth. They followed the same procedure to administer the TIMSS test to their students that the 43 countries used.[2]

CLUSTER ANALYSIS FOR COGNITIVE DIAGNOSIS

Cluster analysis models are based on measures of proximity, such as similarities or dissimilarities, which represent the degree of correspondence among objects across all others used in the analysis (Hair, Black, Babin, Anderson, & Tatham, 2006). Previous studies have examined clustering of items. Beller (1990), for example, used a multidimensional scaling (MDS) model (smallest space analysis) and a hierarchical clustering method (additive tree model) to study the interrelationships among items. The two methods differ in that the former represents objects in a continuous multidimensional space, whereas the latter classifies objects into discrete clusters (Shepard, 1980; Shepard & Arabie, 1979).

1 Regions or countries that elect to participate in TIMSS as benchmarking participants do so because they want "to assess the comparative international standing of their students' achievement and to view their curriculum and instruction in an international context" (IES National Center for Education Statistics, n.d., state/district participation section).

2 In the rest of this paper, we use the term "regions" when referring to both the countries and the two states.

Beller's (1990) study showed that the hierarchical clustering method demonstrated more interpretable and meaningful results than the MDS with respect to identifying the structure of tests and their items. Sireci and Geisinger (1992) used a combination of MDS analysis and hierarchical clustering analysis to evaluate the content representation of a test. They found that this approach was effective in showing the correspondence between item similarity ratings from judges and the item groupings prescribed in the test blueprint. Corter (1995) examined clusters using subtraction-fraction items to verify the attributes required to solve the items. He used cluster methods that included the EXTREE (Corter & Tversky, 1986) approach for investigating hierarchical and overlapping features in order to calculate and analyze measures of similarity. The results of this study confirmed EXTREE as a useful tool for validating matrices of attribute specification.

Other researchers have examined other methods of clustering examinees. These methods require the specification of a matrix that indicates the attributes required for solving an item (i.e., Q-matrix). Using this matrix and the item responses of examinees, Chiu, Douglas, and Li (2009) conducted a cluster analysis designed to group individuals who possessed the same skills. Chiu et al. (2009) showed that this technique was nearly as effective a method as latent class models for the purposes of cognitive diagnosis. Chiu and Seo (2009) applied this method to the 2001 Progress in International Reading and Literacy Study (PIRLS) in order to demonstrate its implementation in practice.

Another type of clustering methodology is the rule space methodology or RSM (Tatsuoka, 1985), which has been used to classify students into a dichotomous pattern of attribute mastery and non-mastery (i.e., knowledge states). The RSM uses the Q-matrix to generate a probability that a given student belongs to a probable knowledge state, based on his or her response patterns. Gierl (2007) proposed a version of the RSM pertaining to the attribute hierarchy method (AHM), which is based on the hierarchies of skills evident in a performance task.

Applications of RSM have been widely studied, including within the context of analyses of data from international comparative assessments such as TIMSS. Dogan and Tatsuoka (2008) used the RSM to analyze the mastery levels of Turkish students and of American students who participated in TIMSS 1999. They found that the Turkish students had weaker algebra and probability/statistics skills than the American students. Um, Dogan, Im, Tatsuoka, and Corter (2003) conducted a similar study. They compared student attribute mastery in Korea, the Czech Republic, and the United States. Using data from 20 countries that participated in the TIMSS 1999-Repeat assessment, Tatsuoka, Corter, and Tatsuoka (2004) examined students' mastery of 23 attributes by comparing their mean mastery levels. They found a high association between mastery of TIMSS geometry items and mathematical thinking skills—skills that were lacking among the United States students.

Birenbaum, Tatsuoka, and Yamada (2004) also used the TIMSS 1999-Repeat data to compare the attribute mastery of students in the United States, Japan, and Israel.

This study additionally examined the performance of Jewish and Arab students in Israel who were studying the same curriculum. Results showed that Japanese students outperformed the students in the other two countries and that the attribute patterns of the Jewish students were significantly more effective than the attribute patterns of the Arab students in terms of correct answers on the TIMSS items. Chen, Gorin, Thompson, and Tatsuoka (2008) studied the performance data of culturally diverse groups on the TIMSS 1999-Repeat assessment. They compared the Taiwanese and American students and used the fit of the RSM as a measure to validate the equivalence of the achievement scale scores of these two student cohorts. Based on the results of classification rates and the prediction of scores, they concluded that a cognitive-psychometric modeling approach such as the RSM is useful for exploring issues related to score validity.

In summary, various researchers have examined clustering of items and examinees in an effort to identify the cognitive diagnostic properties of large-scale achievement tests and the performance of students in relation to those attributes. There have also been many applications of RSM using the TIMSS data. However, in both cases, only a few studies exist in which the researchers involved specifically investigated the clustering of attributes.

The Sum-Score Matrix

As described in Chiu et al. (2009), cluster analysis models used to diagnose cognitive performance require a measure of examinee scores for each attribute. However, this measure needs to be created through the use of two matrices: one that consists of the item responses (correct or incorrect) of test-takers, and one that maps the relationships between the items and attributes required to solve the item (i.e., Q-matrix). Therefore, the primary consideration resides in constructing the Q-matrix (Embretson, 1984; Tatsuoka, 1985), while the secondary consideration focuses on combining the two matrices. The two matrices are combined in order to create a matrix that represents a sum score of a particular attribute; this matrix thus represents both examinee response patterns as well as the attribute specification for each item. Finally, the ensuing matrix should be transformed into a measure of distance that is subsequently used to conduct the cluster analysis.

In the remainder of this section of our paper, we discuss the theory underpinning the following steps: (a) combining the response data to the Q-matrix using the sum-score matrix, (b) transforming the combined matrix into a measure of distance (here we also explain the different measures and their implications), and (c) conducting the *K*-means analysis, the HACA, and the extended similarity tree (EXTREE) analysis. As we noted earlier, the *K*-means and the HACA are methods commonly used to group and examine the hierarchical structure of the clusters, respectively, whereas the EXTREE method handles both overlapping and hierarchical features of the clusters simultaneously, which is an advantage of its use.

A Q-matrix can be constructed by defining q_{jk} to be an incidence matrix, with value "1" signaling the requirement of the attribute and "0" representing otherwise for item j={1,2,...,J} and attribute k={1,2,...,K}. Therefore, this framework allows the formation of a J x K binary matrix; the element in the jth row and kth column of the matrix, q_{jk}, corresponds to whether the kth attribute is required to solve the jth item correctly. Validating the Q-matrix requires multiple coders to independently assign binary values to the matrix and to generate a finalized Q-matrix through discussion and consensus. However, depending on the type of problem embedded in a test item, the Q-matrix can vary by coder, because test-takers can use different approaches to solve an item. However, when specifying the final Q-matrix used for a study, researchers should use the most dominant method, as validated by domain experts.

The finalized Q-matrix is combined with the item responses of the test-takers to generate examinee scores for each attribute, which are then used in the subsequent cluster analysis. If we let Y_{ij} be examinee i's response for item j, such that i={1,2,...,I} and j={1,2,...,J}, then this step becomes tantamount to combining the I x J matrix with the J x K matrix. Although various formulations can be used to join the two matrices, we used the sum-scores matrix that Chiu et al. (2009) employed. Thus:

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

Here, the vector $\underline{W}_i = (W_{i1}, W_{i2},... W_{iK})'$ is the score profile of the attributes for examinee i that is derived from their item responses in Y_{ij}. In other words, the vector is similar to a score for each attribute for examinee i, weighted by the number of times attribute k was required across items. By using the matrix W_{ik}, researchers can create a similarity or dissimilarity matrix that represents the distances between the attributes and which can then be used to conduct the cluster analysis.

Measures of Similarity and Dissimilarity

Although measures of correlation can be used to associate multivariate measures, distance measures are most commonly applied. These represent similarity as the proximity of observations to one another across variables in the cluster. In fact, distance measures are measures of dissimilarity, because larger distance measures represent less similarity. Therefore, to create measures of similarity, an inverse relationship is often used. A proximity distance, which represents the nearness of two objects, r and s, must satisfy the following three conditions:

1. $d_{(\underline{w}_i, \underline{w}_{i'})} \geq 0$ for all $\underline{w}_i$ and $\underline{w}_{i'}$,
2. $d_{(\underline{w}_i, \underline{w}_{i'})} = 0$ if, and only if, $\underline{w}_i = \underline{w}_{i'}$, and
3. $d_{(\underline{w}_i, \underline{w}_{i'})} = d_{(\underline{w}_{i'}, \underline{w}_i)}$.

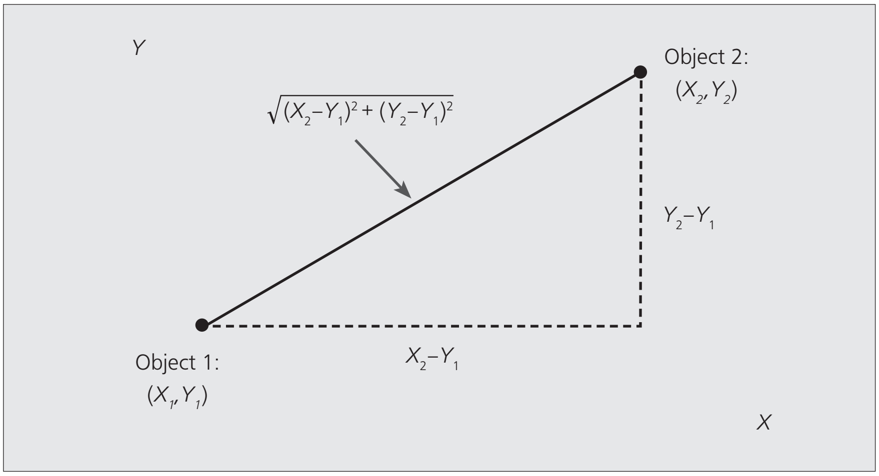
Various measures that satisfy these conditions have been developed. These include the Euclidean distance, the squared Euclidean distance, the city-block distance, the Chebychev distance, and the Mahalanobis distance. A general metric used for distances is the Minkowski p-metric (Hair et al., 2006), which can be generalized to other forms of distances by varying p. For two K-dimensional datapoints $\underline{w}_i$ and $\underline{w}_{i'}$, the following equation defines the Minkowski distance:

$$d_{L_p}(\underline{w}_i, \underline{w}_{i'}) = [\sum_{k=1}^{K}(|w_{ik} - w_{i'k}|)^p]^{\frac{1}{p}} \; .$$

This equation requires the triangle inequality $d_{(\underline{w}_i,\underline{w}_{i'})} \le d_{(\underline{w}_i,\underline{w}_{i''})} + d_{(\underline{w}_{i''},\underline{w}_{i'})}$ to be satisfied. The Minkowski metric can be simplified to form the Euclidean distance when $p = 2$. The Mahalonabis distance is another popular distance measure. It takes into account the covariance between the variables $d_M(\underline{w}_i, \underline{w}_{i'}) = (\underline{w}_i - \underline{w}_{i'})^T \Sigma^{-1} (\underline{w}_i - \underline{w}_{i'})$, where Σ is the covariance matrix of W_{ik}, which is inversed and used as a weight.

Figure 1 shows a representation of a distance measure between two objects for variables X and Y using Euclidean distances. We have provided a simplified illustration of the Euclidean distance given that this distance measure can be generalized to other forms of distances. Figure 1 thus demonstrates how distance measures are *generally* calculated.

Figure 1: An example of Euclidean distance between two objects on variables X and Y



K-means, HACA, and Extended Similarity Tree

The key task with respect to the *K*-means algorithm is that of estimating the cluster centers based on the data, with the number of clusters being predetermined. In *K*-means clustering, the object $\underline{w}_i$ is assigned to cluster m, using, for example, the Euclidean distance:

$$m = \arg \min_{u \in \{1,2\ldots,M\}} || \underline{w}_i - \underline{\hat{c}}_u ||^2 \; .$$

Here, $\underline{\hat{c}}_u$ is the estimated center of the u^{th} cluster derived from the average of the observations within the cluster. The use of M initial K-dimensional clusters requires datapoints to be assigned to clusters by way of the above constraint. The cluster centers are then reset by calculating the average of assigned observations. This process is looped until relocation of observations is exhausted.

While the *K*-means algorithm groups clusters to exclusive clusters, the HACA involves a quite different approach. The hierarchical structure of the data is explored by taking into account distances between clusters in addition to distances between the data. Computationally, HACA is much simpler than the *K*-means. Again using the Euclidean distance as an example, we can define the initial distance matrix for two objects as $d_{ii'} = \sqrt{\sum_{k=1}^{K} (|w_{ik} - w_{i'k}|)^2}$. Each object thus initially begins with its own cluster. Having defined the distance between two clusters C_l and $C_{l'}$ as $d^*_{ll'}$, our next step is to cluster the two objects *i* and *i′* according to which $d^*_{ll'} = d_{ll'}$ is the smallest. For each additional step, two clusters are grouped to achieve the minimum distance by adjoining two of the existing clusters. From here on, the cluster distances are modified after each new join, and the algorithm will vary according to the linkage chosen. This iterative process is continued until all clusters are exhausted. (For further details, refer to Chiu et al., 2009.)

The extended similarity tree (EXTREE) algorithm can be divided into a three-step process (Corter & Tversky, 1986):

1. Obtaining a best-fit additive tree;
2. Estimating a measure of each possible marked feature and selecting the optimal set; and
3. Using a least-squares method to simultaneously estimate all model parameters.

The first step transforms data to satisfy the metric axioms. The neighbor score matrix is then calculated and joined to elements that are mutual nearest-neighbors (MNN), a process that satisfies the following equation:

If $c_{ij} = \max(c_{ik}, \forall\, k \neq i)$ and $c_{ji} = \max(c_{jk}, \forall\, k \neq i)$, then *i, j* is MNN.

This process is looped until all possible combinations are exhausted. The next process eliminates measure of the possible marked feature and picks the best group via the following equation:

$$W_A = \frac{1}{2N} \sum_{Q} [d(x, v) = d(y,u) - d(x,u) - d(y,u)].$$

This step allows both the elimination of redundant features and clique selection of pair-wise features. Finally, the least-squares method is used to estimate the parameters.

The advantage of using the EXTREE model over hierarchical or addtree models is that it allows the graphical representation of overlapping or non-nested features. It also allows the illustration of both common features as well as unique features. These strengths make the EXTREE model ideal for our study because it investigates attributes used to solve mathematics problems and that tend to have overlapping features.

In contrast to models for cognitive diagnosis, which classify test respondents (see, for example, Rupp & Templin, 2008; Tatsuoka, 1983; von Davier, 2005), the methods that we present in this study cluster attributes. We used all three of the above methods for this process. We based our definition of proximity across attributes on distance measures, which allowed us to examine, using the vector of student responses as a basis, which attributes were more similar and which were more distinct in terms of their relative distance.

METHOD

Data

Forty-three countries participated in the TIMSS 2007 Grade 4 mathematics assessment, which generated data from over 360,000 students (Olson, Martin, & Mullis, 2009). Raw data for the analyses were obtained from Boston College's TIMSS website and were converted to SAS data files via use of a modified SAS macro provided by the website; the file was scored and merged according to guidelines supplied by Foy and Olson (2009).

The TIMSS released dataset contains selected items and groups of examinees. We selected data from Blocks 4 and 5 for analysis because they encompass the greatest number of dichotomous items (23 dichotomous items out of 25 total items), a situation that eliminated the need to consider scores with partial credit. Out of the total 25 items used for this study, only two items (Items 12 and 21) were originally scored polytomously, with a maximum score of 2, rather than 1. We dichotomized these polytomous responses by treating responses with partial credit as incorrect and responses with full credit as correct. We also scored as incorrect omitted or unreached items.

As we stated above, we selected for the purposes of our study 11 regions based on their overall TIMSS ranking. Table 1 shows the rank, sample size, and the mean proficiency scores for these nine countries as well as for the two American states that took part in TIMSS as benchmarking participants.

Table 1: Mean proficiency statistics for the TIMSS 2007 Grade 4 participants

Performance	Rank	Country	Sample size	Mean proficiency	Standard error
High	1	Hong Kong SAR	3,791	606.80	3.58
	3	Chinese Taipei	4,131	575.82	1.73
	4	Massachusetts, USA*	1,747	572.48	3.51
	6	Minnesota, USA*	1,846	554.12	5.86
Average	13	United States	7,896	529.01	2.45
	15	Denmark	3,519	523.11	2.40
	24	Sweden	4,676	502.57	2.53
Low	37	Colombia	4,801	355.45	4.97
	41	Kuwait	3,803	315.54	3.65
	42	Qatar	7,019	296.27	1.04
	43	Yemen	5,811	223.68	5.97

Notes:
Based on TIMSS 2007 technical report (Olson et al., 2009).
* Regional entities.

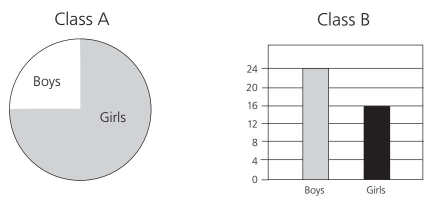
Development of the Q-matrix

As described earlier in this paper, a Q-matrix is a table of skills that indicates whether an attribute is required for an item. When conducting our study, we used the TIMSS framework (Mullis et al., 2005) to identify 15 attributes derived from the 25 items. The TIMSS framework identifies 38 objectives. Using as our reference the list of attributes required to correctly solve an item, we modified and simplified the framework to complement the 25 items so that it represented two blocks of the TIMSS 2007 Grade 4 mathematics assessment. Figure 2 demonstrates how we created the Q-matrix for one of the items that we used in this study. Table 2 presents the list of attributes that we used to develop the Q-matrix. The table also presents the number of times each attribute was specified.

Although we combined and/or modified some objectives, the topic areas were preserved to an extent that enabled us to create 15 attributes that were not only fine-grained enough to allow us to make meaningful statements about the specific skills, but also small enough to prevent the measurement errors for each attribute from becoming too large. The average number of attributes required by all 25 items was 2.80. Three items required a single attribute, eight required two attributes, another eight required three attributes, and four required four attributes. The remaining two items required five and six attributes respectively.

Figure 2: Illustration of the *Q*-matrix specification for TIMSS Item 14

Class A and B each have 40 students.



There are more girls in Class A than in Class B. How many more?

A 14
B 16
C 24
D 30

Item		Number								Geometric shapes and measures				Data and display		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
17	M041336	1	1	0	0	1	1	0	0	0	0	0	0	1	1	0

Table 2: Attributes developed from the 2007 TIMSS framework for Grade 4 mathematics

Content domain	Attributes		Times specified
Number	**Whole Numbers**		
	1. Represent, compare, and order whole numbers as well as demonstrating knowledge of place value. 2. Recognize multiples, computing with whole numbers using the four operations, and estimating computations.	Whole Number (1)	6
	3. Solve problems, including those set in real-life contexts (for example, measurement and money problems).	Whole Number (2)	16
		Whole Number (3)	11
	4. Solve problems involving proportions.	Whole Number (4)	3
	Fractions and Decimals		
	5. Recognize, represent, and understand fractions and decimals as parts of a whole and their equivalents.	Fractions & Decimals (1)	3
	6. Solve problems involving simple fractions and decimals including their addition and subtraction.	Fractions & Decimals (2)	2
	Number Sentences with Whole Numbers		
	7. Find the missing number or operation and model simple situations involving unknowns in number sentences or expressions.	Number Sentences	2
	Patterns and Relationships		
	8. Describe relationships in patterns and their extensions; generate pairs of whole numbers by a given rule and identify a rule for every relationship given pairs of whole numbers.	Patterns & Relationships	3
Geometric Shapes & Measurement	**Lines and Angles**		
	9. Measure, estimate, and understand properties of lines and angles and be able to draw them.	Lines & Angles	3
	Two- and Three-dimensional Shapes		
	10. Classify, compare, and recognize geometric figures and shapes and their relationships and elementary properties.	Two- & Three-dimensional Shapes (1)	7
	11. Calculate and estimate perimeters, area, and volume.	Two- & Three-dimensional Shapes (2)	2
	Location and Movement		
	12. Locate points in an informal coordinate to recognize and draw figures and their movement.	Location & Movement	3
Data & Display	**Reading and Interpreting**		
	13. Read data from tables, pictographs, bar graphs, and pie charts.	Reading & Interpreting (1)	4
	14. Compare and understand how to use information from data.	Reading & Interpreting (2)	3
	Organizing and Representing		
	15. Understand different representations and organize data using tables, pictographs, and bar graphs.	Organizing & Representing	2

Note: The bold headings in the attributes column designate the topic areas within the content domains as indicated by the TIMSS framework (Mullis et al., 2005).

The item in Figure 2 asked students how many more girls are in Class A than in Class B. Although students would have used various methods to solve this item, the dominant strategy validated by experts was coded into the Q-matrix. To solve this item, a student would need to be able to read and compare the proportions between boys and girls in Class A to infer that three-quarters of the students are girls. Because there are 40 students in Class A, students should also be able to deduce that there are 30 girls. For Class B, students would need to be able to read the bar graph and to understand that there are 16 girls. Finally, by determining the difference between 30 and 16, students should arrive at the correct answer, 14.

Correctly solving the problem associated with this item required mastery of six important attributes:

- *Attribute 1:* representing, comparing, and ordering whole numbers as well as demonstrating knowledge of place value;
- *Attribute 2:* recognizing multiples, computing with whole numbers by using the four operations, and estimating computations;
- *Attribute 5:* recognizing, representing, and understanding fractions and decimals as parts of a whole and their equivalents;
- *Attribute 6:* solving problems involving simple fractions and decimals, including their addition and subtraction;
- *Attribute 13:* reading data from tables, pictographs, bar graphs, and pie charts; and
- *Attribute 14:* comparing and understanding how to use information from data.

Table 3 shows the Q-matrix used for the current analysis. Because the validity of the Q-matrix rests on the content mastery and experience of the coders who develop it, we had three mathematics educators independently score our initial Q-matrix. We also asked two professionals with college-level mathematics training and experience in the field to complete the same exercise. We then, through a discussion and consensus process centered on the dominant method used to solve the item, combined the coding of the three Q-matrices to finalize the matrix that we used in this study.

The finalized Q-matrix formed the building block of our analysis, and its accuracy, as just implied, holds the validity of the findings of this study. Although, with respect to educational measurement, the implementation and application of a Q-matrix within a cognitive diagnostic modeling framework relates, to some degree, to what we discuss here, it is different in that we measured the clustering and association of attributes and not their prevalence or the classification of student mastery of a specific attribute.

In cognitive diagnostic models, use of a Q-matrix with 15 attributes and 25 items may require additional complexities in estimation (see Haberman & von Davier, 2006; von Davier 2005) because the latent trait space needed to classify individual students for a model with 15 attributes would contain 2^{15} different attribute mastery types. It is important to understand the difference between the cluster-based approach that essentially "explains" the correlations between the 15 sum scores used in our analysis

Table 3: Q-matrix for TIMSS 2007 Grade 4 mathematics test items

Item	Number (N)								Geometric shapes and measures				Data and display (DD)		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
3	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
4	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
5	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
7	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0
8	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0
9	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
10	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
11	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0
12	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1
13	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0
14	1	1	0	0	1	1	0	0	0	0	0	0	1	1	0
15	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
16	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
17	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
18	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
19	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0
20	0	1	1	0	0	0	0	1	0	0	0	0	0	1	0
21	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0
23	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
25	1	1	0	0	0	0	0	0	0	0	0	0	1	0	1

with a low number of (two to three) clusters and the model for individual cognitive diagnosis utilized to classify individual respondents into multiple binary mastery or ordered attributes.

Analysis

To conduct a cluster analysis of attributes based on examinee performance, we combined the examinees' item response data (Y_{ij}) with the Q-matrix (q_{jk}) developed for the study (see Table 3) via the sum-score vector method. The resulting examinee-by-attribute sum-score matrix, W_{ik}, represents a combination of the summed item responses to the 15 attributes specified in the Q-matrix; in other words, it gives examinees a score for each of the 15 attributes. As such, we could assume that test-takers with a higher score in the W_{ik} matrix would be the test-takers most likely to have the attribute specified in the Q-matrix.

Using the sum-score matrix W_{ik}, we then used Euclidian distances to create a matrix of dissimilarities. (The Euclidian measure tends to be the default choice of distance measure for most cluster analyses.) We used the matrix dissimilarity command in Stata 10 (StataCorp, 2007) to do this. This command offers a flexible syntax by which to create and modify different types of similarity and dissimilarity matrices.

Our next step required us to use the procedure described above to create a separate dissimilarities matrix for the 11 regions (i.e., the nine countries and the two American states). We then used the dissimilarities matrices to conduct the cluster analysis. Because the *K*-means cluster method requires a prespecified number of clusters, we calculated the fusion coefficient for each cluster and its confidence intervals (Aldenderfer & Blashfield, 1984; Wishart, 2005) for each region and examined these in order to determine the largest gap (i.e., the elbow in the scree plots depicted in the results section below). Although this measure can be subjective, it provided us with an arbitrary starting point for the number of clusters. We again used Stata 10 (StataCorp, 2007) to conduct the *K*-means and HACA analyses and a PASCAL-based EXTREE program (Corter & Tversky, 1986) to run the EXTREE analysis.

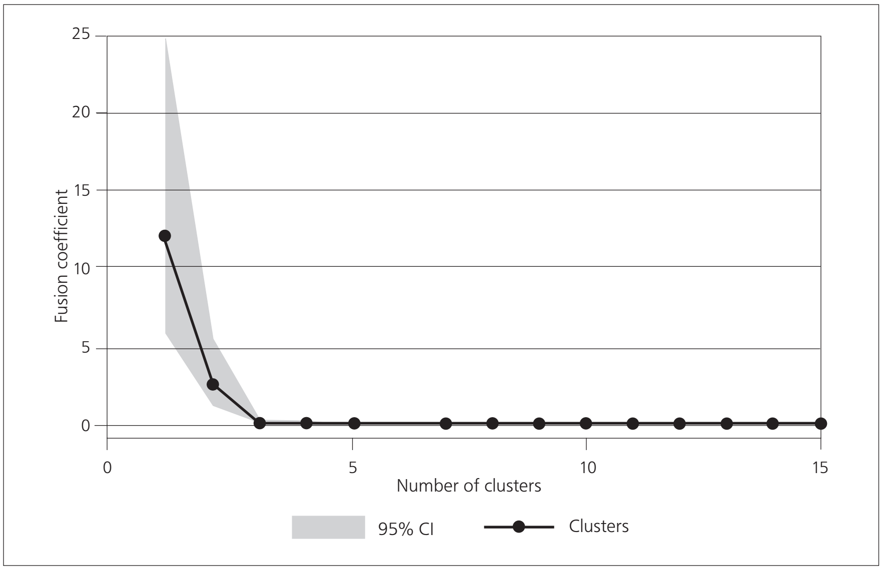
RESULTS

Because we selected the 11 regions on the basis of their students' overall performance (i.e., each region's mean proficiency score), the magnitude of the derived distance measures reflected attribute dissimilarities: better-performing countries had larger dissimilarity indices. At the same time, the values in the dissimilarity matrix may have been affected by the variation in the attribute specification of the Q-matrix; thus, test-takers needed to draw on some attributes more frequently than others, as shown in Table 3. As such, the number of times an attribute was required, which ranged from 2 to 16 times, could also have changed the dissimilarity matrix.

Hong Kong (ranked first on the international TIMMS achievement scale) had the highest values for each cell of the matrix because we used the sum-score vector to calculate each value. Likewise, Yemen, which ranked 43rd on the international scale, had the poorest-performing students and the lowest dissimilarity indices. As we noted earlier, the dissimilarities represent further distance in space and so are less related to one another. Although the overall performance implied the variability of the distances, they did not necessarily infer that the distances would be strictly greater for a better performing country over another country of lower performance. This is because the presence or the absence of a required attribute in the Q-matrix influences the outcome of the dissimilarity matrix.

To determine the number of clusters assigned for the *K*-means analysis, we used the fusion coefficient to generate a plot that would allow us to examine where the greatest difference occurred. Figure 3 shows this result for one region, the United States. Here we can see that the elbow of the coefficient is formed of three clusters. Although we created fusion coefficients for each country, the elbow of the coefficient for most of the selected regions was made up of three clusters. Because the number of clusters formed by using the fusion coefficient can be the product of subjectivity, we selected three clusters that we could use uniformly for all countries. This meant that we could summarize the differences between test-takers on the 15 attribute scores in terms of their membership with respect to one of these three clusters. We then used these in the *K*-means analyses to explain the dependencies between the 15 scores developed from the Q-matrix generated by the mathematics educators and experts.

Figure 3: Plot of fusion coefficients



Tables 4, 5, and 6 present the results of the *K*-means clusters for the high-, average-, and low-performing regions, respectively. Table 4 shows the clusters for Hong Kong, Chinese Taipei, Massachusetts, and Minnesota.

The three clusters for *Hong Kong* were as follows:

- *Cluster One:* Attributes 1, 2, and 3 (whole numbers 1, 2, and 3, number domain) and Attribute 10 (two- and three-dimensional shapes 1, geometric shapes and measurement domain) grouped within Cluster 1;
- *Cluster Two:* Attributes 5 (fractions and decimals 1, number domain), 8 (patterns and relationships, number domain), 12 (location and movement, geometric shapes and measurement domain), 13 (reading and interpreting 1, data and display domain), 14 (reading and interpreting 2, data and display domain), and 15 (organizing and representing, data and display domain);
- *Cluster Three:* Attributes 4 (whole number 4, number domain), 6 (fractions and decimals 2, number domain), 7 (number sentences, number domain), 9 (lines and angles, geometric shapes and measurement domain), and 11 (two- and three-dimensional shapes 2, geometric shapes and measurement domain).

Table 4: *K*-means clustering for high-performing regions

	Cluster 1	Cluster 2	Cluster 3
Hong Kong (Rank #1)	Whole Number (1) Whole Number (2) Whole Number (3) 2- & 3-D Shapes (1)	Fraction & Decimal (1) Patterns & Relations Location & Movement Reading & Interpreting (1) Reading & Interpreting (2) Organizing & Representing	Whole Number (4) Fraction & Decimal (2) Number Sentences Lines & Angles 2- & 3-D Shapes (2)
Chinese Taipei (Rank #3)	Whole Number (1) 2- & 3-D Shapes (1) Reading & Interpreting (1)	Whole Number (2) Whole Number (3)	Whole Number (4) Fraction & Decimal (1) Fraction & Decimal (2) Number Sentences Patterns & Relations Lines & Angles 2- & 3-D Shapes (2) Location & Movement Reading & Interpreting (2) Organizing & Representing
Massachusetts (Rank #4)	Whole Number (1) 2- & 3-D Shapes (1) Reading & Interpreting (1)	Whole Number (2) Whole Number (3)	Whole Number (4) Fraction & Decimal (1) Fraction & Decimal (2) Number Sentences Patterns & Relations Lines & Angles 2- & 3-D Shapes (2) Location & Movement Reading & Interpreting (2) Organizing & Representing
Minnesota (Rank #6)	Whole Number (1) 2- & 3-D Shapes (1) Reading & Interpreting (1)	Whole Number (2) Whole Number (3)	Whole Number (4) Fraction & Decimal (1) Fraction & Decimal (2) Number Sentences Patterns & Relations Lines & Angles 2- & 3-D Shapes (2) Location & Movement Reading & Interpreting (2) Organizing & Representing

Note: Chinese Taipei, Massachusetts, and Minnesota have the same attributes.

Chinese Taipei and the participating regional entities of *Massachusetts* and *Minnesota* (see Table 5) shared the same clusters:

- *Cluster 1:* Attributes 1 (whole number 1, number domain), 10 (two- and three-dimensional shapes 1, geometric shapes and measurement domain), and 13 (reading and interpreting 1, data and display domain);
- *Cluster 2:* Attributes 2 and 3 (whole number 2 and 3, number domain);
- *Cluster 3:* all remaining attributes. This cluster thus contained all attributes from the geometric shapes and measurement domain and the two remaining attributes from the data and display domain.

With respect to the average-performing countries (Table 5), the structure of the attribute clusters for the *United States* and *Sweden* (Table 5) replicated the structure for *Chinese Taipei*, *Massachusetts*, and *Minnesota*. However, for *Denmark*, Attributes

1, 2, and 3 (whole numbers 1, 2, and 3, number domain) and 10 (two- and three-dimensional shapes 1, geometric shapes and measurement domain) grouped into Cluster One. This cluster contained the same attributes that comprised Cluster One for *Hong Kong*.

The attributes in Cluster Two—Attributes 12 (location and movement, geometric shapes and measurement domain), 13 (reading and interpreting 1, data and display domain), and 15 (organizing and representing, data and display domain)—were also found in *Hong Kong's* Cluster Two. Although the remaining attributes fell into Cluster Three, the classifications of attributes in *Denmark* and in *Hong Kong* were similar.

Table 5: *K*-means clustering for average-performing regions

	Cluster 1	Cluster 2	Cluster 3
United States (Rank #13)	Whole Number (1) 2- & 3-D Shapes (1) Reading & Interpreting (1)	Whole Number (2) Whole Number (3)	Whole Number (4) Fraction & Decimal (1) Fraction & Decimal (2) Number Sentences Patterns & Relations Lines & Angles 2- & 3-D Shapes (2) Location & Movement Reading & Interpreting (2) Organizing & Representing
Denmark (Rank #15)	Whole Number (1) Whole Number (2) Whole Number (3) 2- & 3-D Shapes (1)	Location & Movement Reading & Interpreting (1) Organizing & Representing	Whole Number (4) Fractions & Decimals (1) Fractions & Decimals (2) Number Sentences Patterns & Relationships Lines & Angles 2- & 3-D Shapes (2) Reading & Interpreting (2)
Sweden (Rank #24)	Whole Number (1) 2- & 3-D Shapes (1) Reading & Interpreting (1)	Whole Number (2) Whole Number (3)	Whole Number (4) Fraction & Decimal (1) Fraction & Decimal (2) Number Sentences Patterns & Relations Lines & Angles 2- & 3-D Shapes (2) Location & Movement Reading & Interpreting (2) Organizing & Representing

Note: The United States and Sweden have the same attributes in the clusters.

Table 6 shows the *K*-means clusters for the four low-performing countries (Colombia, Kuwait, Qatar, and Yemen). Note that the attribute classifications for these countries differ from the ones presented earlier.

The three clusters for *Colombia* were as follows:

- *Cluster One:* Attributes 1 (whole number 1, number domain), 10 (two- and three-dimensional shapes, geometric shapes and measurement domain), 13 (reading and interpreting, data and display domain), and 15 (organizing and representing, data and display domain);

Table 6: *K*-means clustering for low-performing regions

	Cluster 1	Cluster 2	Cluster 3
Colombia (Rank #37)	Whole Number (1) 2- & 3-D Shapes (1) Reading & Interpreting (1) Organizing & Representing	Whole Number (2) Whole Number (3)	Whole Number (4) Fractions & Decimals (1) Fractions & Decimals (2) Number Sentences Patterns & Relationships Lines & Angles 2- & 3-D Shapes (2) Location & Movement Reading & Interpreting (2)
Kuwait (Rank #41)	Whole Number (1) Whole Number (2) Whole Number (3) 2- & 3-D Shapes (1)	Fractions & Decimals (1) Fractions & Decimals (2) Number Sentences	Whole Number (4) Patterns & Relationships Lines & Angles 2- & 3-D Shapes (2) Location & Movement Reading & Interpreting (1) Reading & Interpreting (2) Organizing & Representing
Qatar (Rank #42)	Whole Number (1) Whole Number (2) Whole Number (3) 2- & 3-D Shapes (1)	Patterns & Relationships 2- & 3-D Shapes (2) Location & Movement Reading & Interpreting (1)	Whole Number (4) Fractions & Decimals (1) Fractions & Decimals (2) Number Sentences Lines & Angles 2- & 3-D Shapes (2) Reading & Interpreting (2)
Yemen (Rank #43)	Whole Number (1) Whole Number (3) 2- & 3-D Shapes (1)	Whole Number (2)	Whole Number (4) Fraction & Decimal (1) Fraction & Decimal (2) Number Sentences Patterns & Relations Lines & Angles 2- & 3-D Shapes (2) Location & Movement Reading & Interpreting (1) Reading & Interpreting (2) Organizing & Representing

- *Cluster Two:* Attributes 2 and 3 (whole number 2 and 3, number domain);
- *Cluster Three:* all remaining attributes.

The first cluster for *Kuwait* and for *Qatar* were similar in that they both contained Attributes 1, 2, and 3 (whole numbers 1, 2, and 3, number domain) and Attribute 10 (two- and three-dimensional shapes, geometric shapes and measurement domain). However, Clusters Two and Three for these two regions differed. The Cluster Two attributes for Kuwait were from the number domain, namely Attributes 5 and 6 (fractions and decimals 1 and 2), and 7 (number sentences). For *Qatar*, the Cluster Two attributes contained attributes from the number, geometric shapes and measurement, and data and display domains.

The Cluster One attributes for the final country—*Yemen*—comprised Attributes 1 and 3 (whole number 1 and 3, number domain) and 10 (two- and three-dimensional shapes, geometric shapes and measurement domain). The second cluster for *Yemen* contained only one attribute—Attribute 2 (whole number 2, number domain).

In general, we found differences in the classification of attributes across the 11 regions. However, the clusters across the higher-performing countries were more similar in structure than the clusters across the lower-performing countries.

Figures 4, 5, and 6 show the hierarchical clusters—derived from the HACA method via complete linkage—across the high-, average-, and low-performing countries, respectively. All of the dendrograms for the 11 regions show three main clusters, the hierarchical structures of which were similar to one another and resembled the results from the *K*-means analysis.

In 10 regions (the exception was *Kuwait*), one cluster comprised Attributes 2 and 3 (whole number 2 and 3, number domain); this is the right-most cluster on the dendrograms. The left-most cluster for the high- and average-performing countries (Hong Kong, *Chinese Taipei*, *Massachusetts*, *Minnesota*, the *United States*, *Denmark*, and *Sweden*) contained Attribute 1 (whole numbers 1, number domain), 10 (two- and three-dimensional shapes, geometric shapes and measurement domain), and 13 (reading and interpreting, data and display domain). The remaining attributes clustered into two subgroups comprising attributes from the number, geometric shapes and measures, and data and display domains. The hierarchical clusters that emerged from the low-performing countries showed differences for the far-left cluster (Figure 6). *Colombia's* HACA, for example, had five attributes, while *Kuwait's* had four.

Figure 4: Hierarchical agglomerative cluster dendrogram for high-performing regions

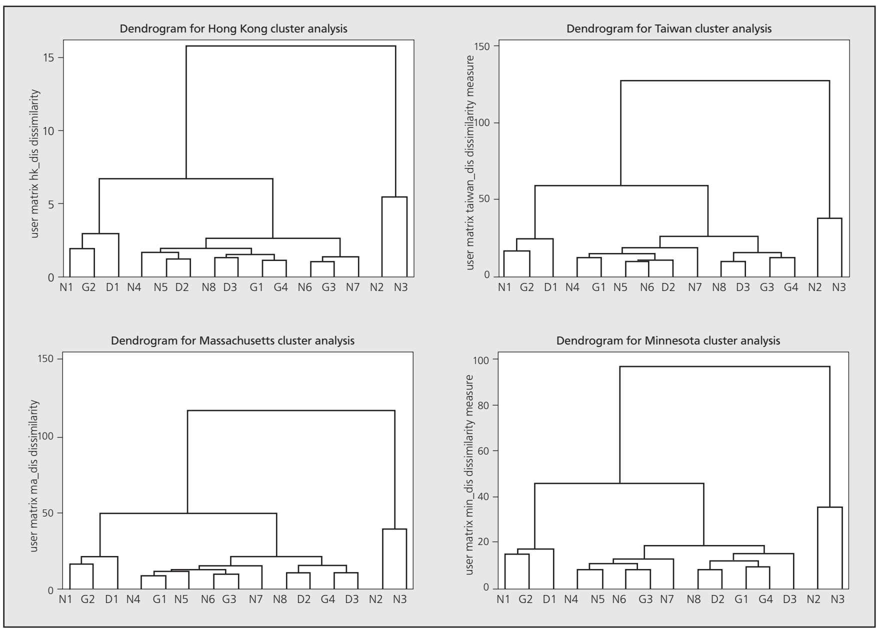


Figure 5: Hierarchical agglomerative cluster dendrogram for average-performing regions

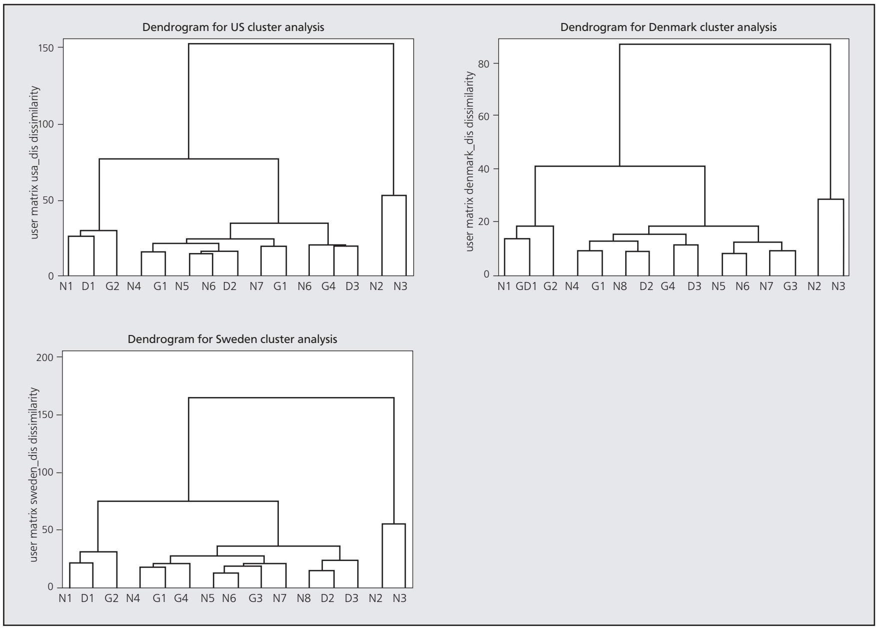
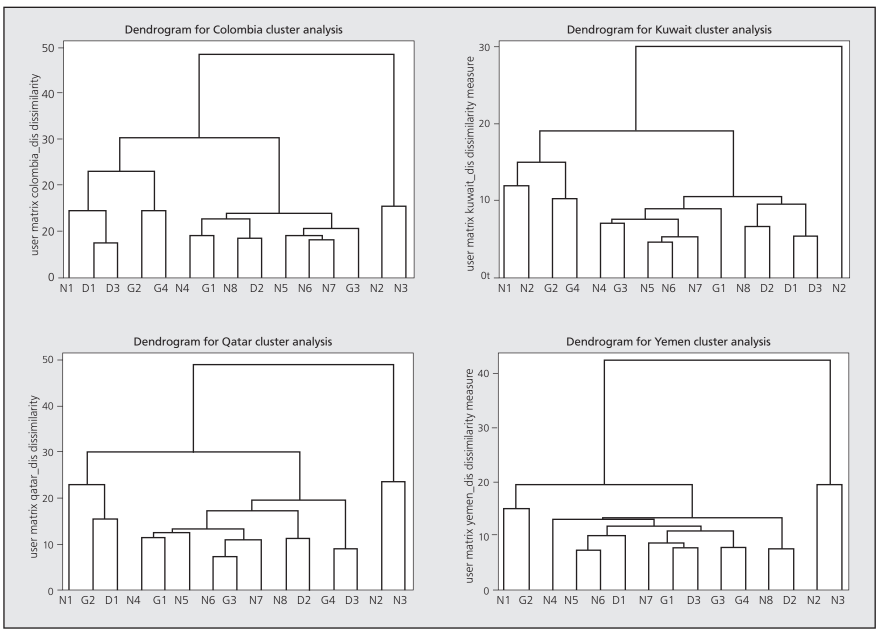


Figure 6: Hierarchical agglomerative cluster dendrogram for low-performing regions



Note: N = number domain, G = geometric shapes and measures domain, D = data and display domain.

We used the results from both the *K*-means and the HACA clusters to conduct the EXTREE analysis. Table 7 shows the fit statistics derived from running the EXTREE procedure. The proportion of variance accounted for by each dataset was approximately or above 99%, and the stress statistics were between good to excellent, with these ratings based on Borg and Groenen's (2005) recommendations. These indicators thus showed that the EXTREE model had a good fit with the data.

Table 7: Fit statistics

Rank	Fit statistic	Stress formula 1	Stress formula 2	*R*(monotonic) squared	*R*-squared (p.v.a.f.)*
1	Hong Kong	0.035	0.053	0.997	0.994
3	Chinese Taipei	0.021	0.032	0.999	0.999
4	Massachusetts	0.025	0.039	0.999	0.997
6	Minnesota	0.026	0.042	0.998	0.997
13	USA	0.030	0.050	0.998	0.995
15	Denmark	0.034	0.058	0.997	0.997
24	Sweden	0.031	0.050	0.998	0.997
37	Colombia	0.031	0.064	0.996	0.991
41	Kuwait	0.039	0.092	0.992	0.978
42	Qatar	0.038	0.090	0.992	0.983
43	Yemen	0.032	0.064	0.996	0.987

Notes:
Stress values less than 0.050 are considered good fit, and values less than 0.025 are considered excellent fit (Borg & Groenen, 2005).
*Proportion of variance accounted for.

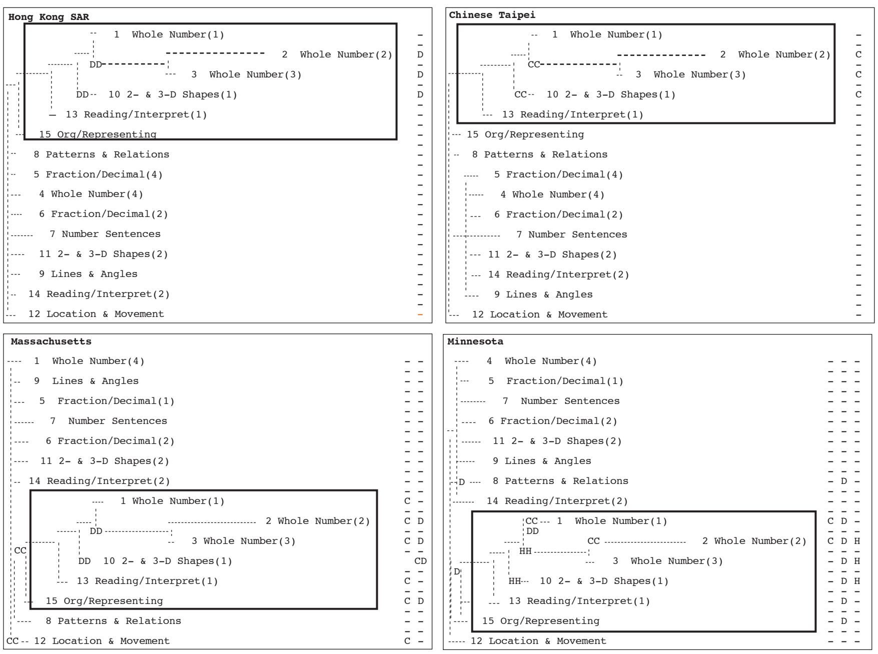
Figures 7, 8, and 9 show the hierarchical clusters produced by the EXTREE model for the high-, average-, and low-performing regions, respectively. The cluster structures that emerged from the EXTREE analysis were similar to those that emerged from the *K*-means and HACA analyses. The following presents the attributes grouped within the EXTREE clusters for the three groups of countries.

- *High-performing regions* (Figure 7): a single hierarchical cluster emerged that contained Attributes 1, 2, and 3 (whole number 1, 2, and 3, number domain), 10 (two- and three-dimensional shapes 1, geometric shapes and measurement domain), and 13 (reading and interpreting 1, data and display domain). However, there were variations in this cluster. For example, in *Hong Kong*, *Massachusetts*, and *Minnesota*, Attribute 15 (organizing and representing, data and display domain) was included in the hierarchical cluster. A secondary cluster also seems to have formed in *Chinese Taipei* and *Minnesota*.
- *Average-performing regions* (Figure 8): here we observed a distinct second hierarchical cluster that included Attributes 1, 2, 3, 10, 13, and 15. We also noted a secondary cluster that included Attributes 4 (whole number 4, number domain), 5 and 6 (fractions and decimals 1 and 2, number domain), 7 (number sentences, number domain), and 11 (two- and three-dimensional shapes 2, geometric shapes

and measurement domain). The hierarchical clusters for *Denmark* and the *United States* also included Attribute 14 (reading and interpreting, data and display domain), and for *Sweden*, included Attribute 9 (lines and angles, geometric shapes and measurement domain).

- *Low-performing regions* (Figure (9): the cluster pattern for these regions was similar to the patterns for both the high- and average-performing regions; the cluster formed from Attributes 1, 2, 3, and 10 was also present for this last grouping but with additional attributes. We also observed a hierarchical cluster formed by Attributes 5, 6, 7, and 11 for *Colombia*, *Kuwait*, and *Qatar*. Although we noted variations in the attribute clusters across the low-performing regions, the hierarchical structure of these clusters varied little from the structure of the average-performing regions.

Figure 7: Extended similarities tree dendrograms for high-performing regions



Tables 8, 9, and 10 present the overlapping clusters that were also generated by the EXTREE model for the high-, average-, and low-performing regions, respectively. The most prominent outcome of this analysis was the number of clusters associated with the regions' respective rankings on the TIMSS international achievement scale. The tables show that the regions of lower performance had more overlapping clusters. *Hong Kong* and *Chinese Taipei* had one, *Massachusetts* two, *Minnesota* three, and the *United States* four. Both *Denmark* and *Sweden* had five attributes.

Figure 8: Extended similarities tree dendrograms for average-performing regions

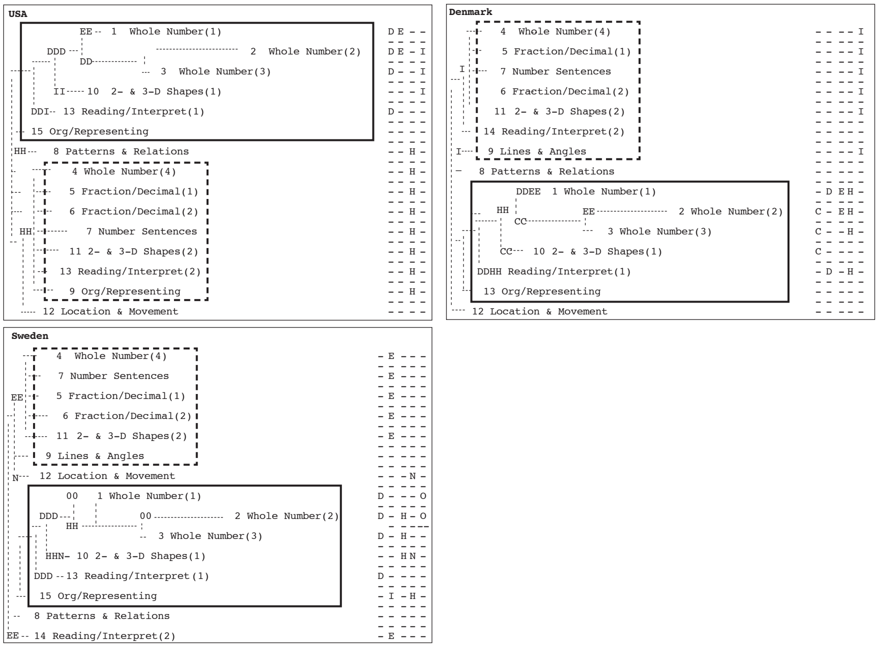


Figure 9: Extended similarities tree dendrograms for low-performing regions

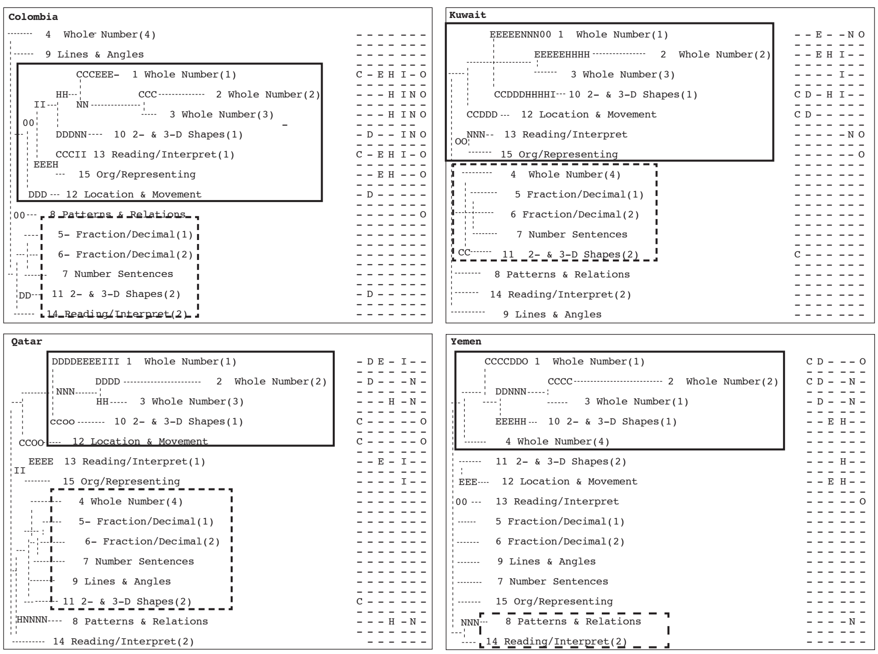


Table 8: Overlapping features for high-performing regions

Region	Overlapping features						
	1	2	3	4	5	6	7
Hong Kong (Rank #1)	Whole Number (2) Whole Number (3) 2-&3-D Shapes (1)						
Chinese Taipei (Rank #3)	Whole Number (2) Whole Number (3) 2-&3-D Shapes (1)						
Massachusetts (Rank #4)	Whole Number (2) Whole Number (2) Whole Number (3) 2-&3-D Shapes (1) 2- & 3- D Shapes (1)	Whole Number (1) Whole Number (1) Whole Number (2) Whole Number (3) Location & Movement Reading & Interpreting (1) Organizing & Representing					
Minnesota (Rank #6)	Whole Number (2) Whole Number (3) 2- & 3-D Shapes (1) 2- & 3- D Shapes (1)	Whole Number (1) Whole Number (2) Whole Number (3) Patterns & Relations Reading & Interpreting (1) Organizing & Representing	Whole Number (1) Whole Number (2)				

Note: Hong Kong and Chinese Taipei have the same overlapping clusters, as do Massachusetts and Minnesota.

Table 9: Overlapping features for average-performing regions

Region	Overlapping features						
	1	2	3	4	5	6	7
United States	Whole Number (1)						
(Rank #13)	Whole Number (1)	Whole Number (4)					
Whole Number (2)	Whole Number (2)	Whole Number (3)	Whole Number (2)	Number Sentences			
		2- & 3- D Shapes (1)	Whole Number (3)	Fraction & Decimal (1)			
		Reading & Interpreting (1)		Fraction & Decimal (2)			
				Lines & Angles			
				2- & 3- D Shapes (2)			
				Patterns & Relations			
				Reading & Interpreting (2)			
Denmak	Whole Number (1)	Whole Number (2)	Whole Number (1)	Whole Number (1)	Whole Number (4)		
(Rank #15)	Whole Number (2)	Whole Number (3)	Whole Number (2)	Reading & Interpreting (1)	Fraction & Decimal (1)		
		2- & 3-D Shapes (1)	Whole Number (3)		Fraction & Decimal (2)		
			Reading & Interpreting (1)		Number Sentences		
					2- & 3-D Shapes (2)		
					Lines & Angles		
Sweden	Whole Number (1)	Whole Number (2)	Whole Number (1)	2- & 3-D Shapes (1)	Whole Number (4)		
(Rank #24)	Whole Number (2)	Whole Number (3)	Whole Number (2)	Location & Movement	Fraction & Decimal (1)		
		2- & 3-D Shapes (1)	Whole Number (3)		Fraction & Decimal (2)		
			Reading & Interpreting (1)		Number Sentences		
					2- & 3-D Shapes (2)		
					Reading & Interpreting (2)		
					Organizing & Representing		

Note: Clusters with the same overlapping attributes across the different regions are marked to emphasize the prevalence of similar features.

Table 10: Overlapping features for low-performing regions

Region	Overlapping features						
	1	2	3	4	5	6	7
Colombia (Rank #37)	Whole Number (1) Whole Number (2) Reading & Interpreting (1)	Whole Number (2) Whole Number (3) 2- & 3-D Shapes (1)	2- & 3-D Shapes (1) 2- & 3-D Shapes (2) Location & Movement	Whole Number (1) Reading & Interpreting (1) Organizing & Representing	Whole Number (1) Whole Number (2) Whole Number (3) Reading & Interpreting (1) Organizing & Representing	Whole Number (1) Whole Number (2) Whole Number (3) Reading & Interpreting (1) 2- & 3-D Shapes (1)	Whole Number (1) Whole Number (2) Whole Number (3) 2- & 3-D Shapes (1) Reading & Interpreting (1) Organizing & Representing Patterns & Relations
Kuwait (Rank #41)	Whole Number (1) Whole Number (1) Whole Number (2)	Whole Number (2) Whole Number (2) Whole Number (3) 2- & 3-D Shapes (1)	2-&3-DShapes (1) 2-&3-DShapes (1) 2-&3-DShapes (2) Location & Movement	Whole Number (1) Whole Number (1) Reading & Interpreting (1) Organizing & Representing	Whole Number (1) Whole Number (1) Reading & Interpreting (1)	2- & 3-D Shapes (1) Location & Movement	Whole Number (2) 2- & 3-D Shapes (1)
Qatar (Rank #42)	Whole Number (1) Whole Number (2)	Whole Number (2) Whole Number (3) Patterns & Relations	2- & 3-D Shapes (1) 2- & 3-D Shapes (2) Location & Movement	Whole Number (1) Reading & Interpreting (1) Organizing & Representing	Whole Number (1) Reading & Interpreting (1)	2- & 3-D Shapes (1) Location & Movement	Whole Number (3) Patterns & Relations
Yemen (Rank #43)	Whole Number (1) Whole Number (2)	Whole Number (2) Whole Number (3) Patterns & Relations	2- & 3- D Shapes (1) 2- & 3- D Shapes (2) Location & Movement	Whole Number (1) Whole Number (2) Whole Number (3)	Whole Number (1) Reading & Interpreting (1)	2- & 3- D Shapes (1) Location & Movement	

Note: Clusters with the same overlapping attributes across the different regions are marked to emphasize the prevalence of similar features.

Seven overlapping clusters were evident among the low-performing regions (Table 10). *Colombia*, *Kuwait*, and *Qatar* had seven such clusters, while *Yemen*, the lowest-ranking region, had six. Unlike the patterns evident in the previous tables and figures, the attributes presented in Table 10 did not distinctly group in only one cluster. In other words, the same attribute appeared in more than one cluster, which was not surprising given the clusters represented overlapping attributes. Similar to the results shown previously, Attributes 2 (whole number 2, number domain), 3 (whole number 3, number domain), and 10 (two- and three-dimensional shapes 1, geometric shapes and measurement domain) formed one cluster for *Hong Kong*, *Chinese Taipei*, *Massachusetts*, *Minnesota*, the *United States*, *Denmark*, *Sweden*, *Colombia*, and *Kuwait*. Attributes 1 and 2 (whole number 1 and 2, number domain) also formed a cluster for *Minnesota*, the *United States*, *Denmark*, *Kuwait*, *Qatar*, and *Yemen*.

Of the two benchmarking participants, *Massachusetts* had two clusters. *Minnesota*, however, had these same two clusters and one other. The overlapping clusters for *Denmark* and *Sweden* were similar in that the attributes in four out of five of the clusters were the same. We also noted overlapping clusters that were present only among the low-performing regions. For example, in all four regions, Attributes 10 and 11 (two- and three-dimensional shapes 1 and 2, geometric shapes and measurement domain) and 12 (location and movement, geometric shapes and measurement domain) overlapped (see Cluster Three) and were from the geometric shapes and measurement domain. Again, Attributes 1 (whole number 1, number domain), 13 (reading and interpreting, data and display domain), and 15 (organizing and representing, data and display domain) were all present within the four regions. Furthermore, for *Kuwait*, *Qatar*, and *Yemen*, Attributes 1 and 13 overlapped, as did attributes 10 and 12 (see Clusters 5 and 6). Again, given the overlapping nature of the clusters, the presence of an attribute in more than one cluster shows that the attribute is not distinct and that it is less hierarchical.

DISCUSSION AND CONCLUSION

Our purpose in conducting this study was to identify, from the TIMSS Grade 4 mathematics test items, patterns of attribute clusters. We used the distances presented by the accuracy of student responses to generate a dissimilarity proximity matrix. We observed, with respect to the 25 items and 15 attributes that we selected for this study, a common pattern emerging from the hierarchical clusters and overlapping features.

When conducting our analyses, we used clustered attributes, a focus that differs from previous studies of clustering within the context of cognitive diagnosis, where the researchers concerned examined items (see, for example, Beller, 1990; Corter, 1995; Sireci & Geisinger, 1992) or examinees (e.g., Chiu et al., 2009; Chiu & Seo, 2009). The findings of these studies and ours nevertheless show that clusters tend to vary across items, examinees, and attributes, depending on the application. As such, this study provides a framework for examining structures of attribute clusters that may help researchers and policymakers not only view, from a macroscopic perspective, how students demonstrate their mastery of attributes, but also gain some idea of how students determine which attributes to use.

Our results also show that there is considerable value in examining the cluster structures produced by the *K*-means and the HACA, and that these analyses can be usefully extended to the EXTREE model, which, in our case, supported the cluster structures of the former two methods. We noted from the dissimilarity matrix that a region with higher performance also had greater measures of distance, meaning that an attribute was further apart in space. This was generally the case for the higher-performing regions both cross-nationally and within the United States. Furthermore, the findings from these cluster analyses suggest that Attributes 1 (whole numbers 1, number domain), 2 (whole numbers 2, number domain), 3 (whole numbers 3, number domain), 10 (two- and three-dimensional shapes 1, geometric shapes and measurement domain), and also 13 (reading and interpreting 1, data and display domain) would continue to cluster. The fact that this pattern emerged from all three methods both across and within all regions suggests a common skill derived from whole numbers, two- and three-dimensional shapes, and reading and interpreting data. We suspect this pattern was indeed the case with respect to our analyses, because the application of whole numbers constituted a wide range of joint usage with other attributes.

When examining the cluster structures across regions, we found a greater degree of similarity with respect to attribute classification across the higher- and average-performing regions than across the lower-performing regions. The *K*-means results revealed that Hong Kong had a greater distribution of attribute classification than any other country, while the United States and its two benchmarking states had the same classifications. We noted the same pattern within the results of the HACA, with attributes in the lower-performing regions being the most different in structure. However, the three retained clusters were quite similar to one another.

The outcomes of the EXTREE analysis, which simultaneously combined the hierarchical and overlapping attributes, showed that the poorer a region's performance, the higher the incidence of attribute clusters. Thus, higher numbers of clusters and classification of attributes into multiple clusters, as indicated by the overlapping clusters, were most evident among the lower-performing countries. Although the hierarchical structures within the United States and in Massachusetts and Minnesota were similar, the overlapping clusters showed that even within the United States differences could be observed intra-nationally.

We consider that the greatest value in conducting our cluster analysis resided in the opportunity it gave us to examine the overlapping clusters of attributes, especially in terms of whether students in a particular region perceived and processed a specific attribute with reference to or in the same way as another attribute. Although the analysis conducted in this study was exploratory, in that one cannot fully claim the clustering of attributes to imply a low mastery of a particular skill, the consistency in the patterns is notable. Furthermore, the clustering of attributes also indicates that attributes tend to be learned together, a happenstance that may be the product of various cognitive, developmental, and/or curriculum factors. Yet, given the assumption that the 15 attributes that we used in our study were distinct and separately used for solving problems, the results may lead to useful indicators for researchers and instructors. The clearer attribute structures and performance patterns of the higher-performing countries may thus be due to these countries having more standardized curricula, or their students having better basic skills.

In the case of the lower-performing regions (Colombia, Kuwait, Qatar, and Yemen)—the regions with the most overlapping clusters—the lack of clearly differentiated clusters makes it difficult to identify them as having distinct and unique attribute structures. Instructors can, however, take this information to identify the grouped clusters with the aim of separating out the overlapping structures of these clusters. In the two American states that we included in order to evaluate the cluster structure within the same country, there seemed to be both disparities and similarities in the clusters. We again emphasize at this point that the appearance of overlapping clusters in low-performing regions does not necessarily mean that the cluster structure is less distinct, but rather that the structure is less hierarchical. We consider that more study directed at examining the relationship between overlapping clusters and student mastery of attributes is needed.

Our study presented another advantage associated with examining clusters of attributes. This was a reduction in the computational burden arising out of traditional techniques of cluster analysis. Unlike other models of cognitive diagnosis that use restricted latent classes for classification and therefore require heavy computational power, all analyses conducted in this study can be run using most statistical software packages under their usual running speeds (i.e., fewer than three seconds for all convergence when using standard computing memory and processors). Once again, we need to note that the models we used included cluster attributes, not students. Therefore, researchers interested in investigating or exploring the cluster structures of attributes to determine preliminary results should find the method presented in this paper an efficient way to examine how students tend to draw on the multiple attributes assumed relevant for solving mathematics problems. This type of exploratory work can furthermore aid the creation and validation of the Q-matrix, especially given that this method does, at times, involve a cumbersome process. The clustering information provided by an analysis such as this can furthermore be used as diagnostic feedback with respect to construction of that analysis. It can also be used to help identify different cognitive diagnosis models useful for describing the structure of the attributes.

In short, the main finding of this study indicates that clustering methods, especially those using EXTREE, can be useful for detecting the overlapping clusters that other clustering methods do not show; perhaps they may indicate areas that low-performing students can focus on in order to improve their achievement. This is another area that requires further research. Although the results that emerged from our *K*-means and HACA analyses were similar, the clusters formed by EXTREE provided a greater depth of information. This outcome may indicate that students from higher-performing regions have a greater applied understanding of the attributes needed to solve a particular item-based problem because they perceive and use these attributes as distinct and independent from others. The tendency toward clustering of both similar and unrelated attributes evident in the lower-performing countries likely indicates student uncertainty and a lack of mastery of the required attributes. Utilization of the fine-grained attributes specified in this analysis can therefore not only help improve student performance, but also serve as a reliable method for sorting and providing achievement-based information of a kind that educational researchers and policymakers will find most useful. For these individuals, opportunity to examine and tease out the attributes within overlapping clusters should help them focus more effectively on the specific areas of learning identified as necessary to improve students' mastery and understanding of when to apply a particular attribute.

References

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Newbury Park, CA: Sage.

Beller, M. (1990). Tree versus geometric representation of tests and items. *Applied Psychological Measurement*, *14*(1), 13–28.

Birenbaum, M., Tatsuoka, C., & Yamada, Y. (2004, May). *Diagnostic assessment in TIMMS-R: Comparison of eighth graders' mathematics knowledge states in the United States, Japan, and Israel*. Paper presented at the First IEA International Research Conference Lefkosia, Cyprus. Retrieved from http://www.iea.nl/fileadmin/user_upload/IRC2004/Birenbaum_ Tatsuoka_Yamada.pdf

Borg, I., & Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications* (2nd ed.). New York, NY: Springer-Verlag.

Cai, J. (2007). What is effective mathematics teaching? A study of teachers from Australia, Mainland China, Hong Kong, and the United States. *Mathematics Education*, *39*, 265–270.

Cai, J., & Silver, E. A. (1995). Solution processes and interpretations of solutions in solving a division-with-remainder story problem: Do Chinese and U.S. students have similar difficulties? *Journal for Research in Mathematics Education*, *26*, 491–497.

Chen, Y.-H., Gorin, J. S., Thompson, M. S., & Tatsuoka, K. K. (2008). Cross-cultural validity of the TIMSS-1999 mathematics test: Verification of a cognitive model. *International Journal of Testing*, *8*, 251–271.

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

Chiu, C., & Seo, M. (2009). Cluster analysis for cognitive diagnosis: An application to the 2001 PIRLS reading assessment. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 137–159.

Corter, J. E. (1995). Using clustering methods to explore the structure of diagnostic tests. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 305–326). Hillsdale, NJ: Lawrence Erlbaum Associates.

Corter, J. E., & Tversky, A. (1986). Extended similarity trees. *Psychometrika*, *51*, 429–451.

Dogan, E., & Tatsuoka, K. K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R. *Educational Studies in Mathematics*, *68*(3), 263–272.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika*, *49*, 175–186.

Foy, P., & Olson, J. F. (2009). *TIMSS 2007 user guide for the international database*. Chestnut Hill, MA: Boston College.

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement*, *44*(4), 325–340.

Givvin, K. B., Hiebert, J., Jacobs, J. K., Hollingsworth, H., & Gallimore, R. (2005). Are there national patterns of teaching? Evidence from the TIMSS 1999 video study. *Comparative Education Review*, *49*, 311–343.

Haberman, S. J., & von Davier, M. (2006). A note on models for cognitive diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1031–1038). Amsterdam, the Netherlands: Elsevier.

Hair, J. F., Black, B., Babin, B., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed). Upper Saddle River, NJ: Prentice Hall.

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

Hiebert, J., Stigler, J. W., Jacobs, J. K., Givvin, K. B., Garnier, H., Smith, M., ... Gallimore, R. (2005). Mathematics teaching in the United States today (and tomorrow): Results from the TIMSS 1999 video study. *Educational Evaluation and Policy Analysis*, *27*, 111–132.

Hook, W., Bishop, W., & Hook, J. (2007). A quality math curriculum in support of effective teaching for elementary schools. *Educational Studies in Mathematics*, *65*, 125–148.

House, D. J. (2005). Motivational qualities of instructional strategies and computer use for mathematics teaching in Japan and the United States: Results from the TIMSS 1999 assessment. *International Journal of Instructional Media*, *32*, 89–105.

IES National Center for Education Statistics. (n. d.). *Trends in International Mathematics and Science Study (TIMSS)*. Washington, DC: Author. Retrieved from http://nces.ed.gov/timss/index.asp

Kuhs, T. M., & Ball, D. L. (1986). *Approaches to teaching mathematics: Mapping the domains of knowledge, skills, and dispositions* [Research memo]. East Lansing, MI: Michigan State University, Center on Teacher Education.

Lubienski, S. T. (2000). Problem solving as a means toward mathematics for all: An exploratory look through a class lens. *Journal for Research in Mathematics Education*, *31*, 454–482.

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.

McNeely, M. (Ed.). (1997). *Guidebook to examine school curricula*. Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education.

Menon, P. (2000). Should the United States emulate Singapore's education system to achieve Singapore's success in the TIMSS? *Mathematics Teaching in the Middle School*, *5*, 345–348.

Mullis, I. V.S., Martin M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: Boston College.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (2009). *TIMSS 2007 technical report*. Chestnut Hill, MA: Boston College.

Papanastasiou, E. (2002). Factors that differentiate mathematics students in Cyprus, Hong Kong, and the USA. *Educational Research and Evaluation*, *8*, 129–146.

Rautalin, M., & Alasuutari, P. (2007). The curse of success: The impact of OECD's Programme for International Student Assessment on the discourses of the teaching profession in Finland. *European Educational Research Journal*, *6*(4), 348–363.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 219–262.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting and clustering. *Science*, *210*, 390–398.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*, 87–123.

Simola, H. (2005). The Finnish miracle of PISA: Historical and sociological remarks on teaching and teacher education. *Comparative Education*, *41*(4), 455–470.

Sireci, S. G., & Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. A*pplied Psychological Measurement*, *16*(1), 17–31.

StataCorp. (2007). Stata statistical software: Release 10 [computer program]. College Station, TX: StataCorp LP.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*, *1*2, 55–73.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, *41*(4), 901–926.

Um, E., Dogan, E., Im, S., Tatsuoka, K., & Corter, J. E. (2003, April). *Comparing Eighth Grade diagnostic test results for Korean, Czech, and American student*s. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

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

Wishart, D. (2005). Number of clusters. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1442–1446). Chichester, UK: Wiley.