Profile Classification for Cognitive Diagnostic Assessment: A Simulation Study

White Paper

Rob Kirkpatrick
Changjiang Wang
Ching-Wei Shin
Yuehmei Chien
Pearson

Joshua Goodman
Pacific Metrics

Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

April, 2013

About Pearson
Everything we do at Pearson grows out of a clear mission: to help people make progress in their lives through personalized and connected learning solutions that are accessible, affordable, and that achieve results, focusing on college-and-career readiness, digital learning, educator effectiveness, and research for innovation and efficacy. Through our experience and expertise, investment in pioneering technologies, and promotion of collaboration throughout the education landscape, we continue to set the standard for leadership in education.For more information about Pearson, visit http://www.pearson.com/.

About Pearson's Research & Innovation Network
Our network **mission** is to spark innovation and create, connect, and communicate research and development that drives more effective learning. Our **vision** is students and educators learning in new ways so they can progress faster in a digital world**.** Pearson's research papers share our experts' perspectives with educators, researchers, policy makers and other stakeholders. Pearson's research publications may be obtained at: http://researchnetwork.pearson.com/.

Profile Classification for Cognitive Diagnostic Assessment: A Simulation Study

Introduction

Cognitive diagnostic assessment (CDA) is a systematic process that seeks to obtain detailed information about the strengths and weaknesses of students' knowledge, skills, and abilities (Rupp, Templin, & Henson, 2010). This information is most often organized in the form of cognitive profiles to assist teachers in designing remedial efforts for individual students. However, with testing time at a premium teachers may want to make use of any test data for additional purposes, such as program evaluation, or tailoring instructional programs for their classrooms as well as individual students. Even in cognitive models with only a small number of cognitive attributes, a common CDA reporting strategy may result in nearly as many unique profiles reported as the teacher has students. Such an outcome may be perceived as impractical by the teacher. As a result, one challenge in using the information produced from CDA is how to reduce the number of reported cognitive profiles to a size that is manageable by the teacher while retaining the maximum usable information. The purpose of this study is to evaluate the efficiency and accuracy of classification methods in reducing the number of cognitive profiles using simulated data. Specifically, this study will focus on two commonly used clustering methods: the hierarchical agglomerative clustering (HAC) and the partitioning around medoids (PAM) methods.

Cognitive Diagnostic Assessment

CDA is a confirmatory framework used to confirm or refute some a priori hypotheses about students' cognition by presenting students with test items measuring specific cognitive attributes (Rupp, et al., 2010). The first step in CDA is the specification of a cognitive model of task performance which encompasses the cognitive attributes as well as the relationships among attributes. This cognitive model is, ideally, identified from the theories in a content domain. When specifying cognitive models, different organization principles may be adopted (Leighton, Gierl, & Hunka, 2004; Wang & Gierl, 2011), which may result in different types of attribute

hierarchies (e.g., linear, convergent, divergent, and unstructured hierarchies). Because these cognitive models are intended to represent the construct of a certain assessment, different specifications of these models will result in different cognitive profiles for students, even though the same test items are administered to students.

After the cognitive model is specified, the attributes measured by each item are specified numerically in a table called the Q-matrix, in which the rows are used to indicate the items and the columns the attributes (Rupp, et al., 2010; Tatsuoka, 1983). The Q-matrix is composed of 1s and 0s, indicating whether or not an attribute is measured by an item. It serves as the link between the cognitive model and students' responses to the test items and allows inferences to be drawn about which attributes have or have not been mastered by the students.

After information about students' cognitive strengths and weaknesses are produced, cognitive profiles are composed for students to assist in remedial efforts. Assuming independence among the cognitive attributes, a cognitive model with k attributes will be able to produce 2^k distinct cognitive profiles. If allowing for hierarchical relationships among attributes (e.g. attribute hierarchies), the number of cognitive profiles may be reduced significantly. However, even when this is true, the number of distinct profiles may still prove overwhelming to classroom teachers, and thus limit a teacher's ability to use the diagnostic information provided for purposes beyond individual student remediation. Thus, methods must be explored to further reduce the number of profiles while at the same time retaining the maximum usable information from CDA.

Method

To evaluate the effectiveness and accuracy of the methods in reducing the number of cognitive profiles, a simulation study was conducted. Student response data were generated under a variety of conditions expected to affect the profile reduction. Two of the factors considered were the structure of the cognitive model and item discrimination power. Additionally, two clustering methods were used in the classification of the cognitive profiles.

Four types of model structures, namely linear (Figure 1a), convergent (Figure 1b), divergent (Figure 1c), and unstructured (Figure 1d) structures, were simulated. These four types of structures correspond to the four prototypical attribute hierarchies proposed in Leighton, et al. (2004) and represent decreasing structuredness in the hierarchies. The linear hierarchy is the most structured while the unstructured hierarchy is the least structured. Each of the models contains seven attributes, which is close to the upper bound for the number of cognitive attributes contained in a cognitive model that is practically applicable (Rupp, et al., 2010). Five items were simulated for each cognitive attribute which resulted in a test length of 35 items. This test length was also chosen because attributes can be measured reliably to an acceptable level at this test length (Gierl, Cui, & Zhou, 2009).

The discriminatory power of the items was another factor considered in the simulation. Students who have mastered the attributes required by a highly discriminating item are expected to have a high probability of responding to the item correctly, whereas students who have not are expected to have a low probability (Cui & Leighton, 2009). Two levels of item discrimination, namely high vs. low, will be included in simulation. High item discriminating power is reflected by a relatively large difference between masters and non-masters in terms of their probabilities of producing correct responses to test items, and low item discriminating power is indicated by a relatively small probability difference.

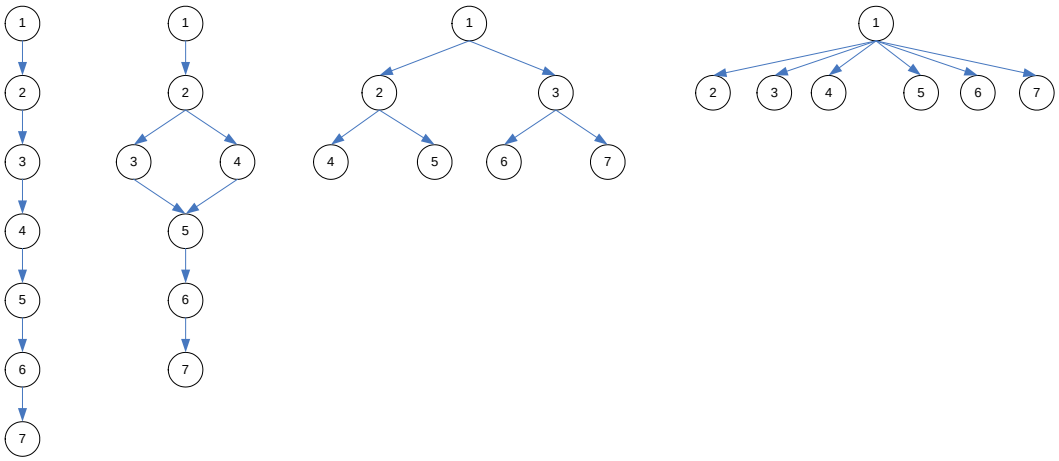

Figure 1. The four cognitive models used for data simulation.

The two methods used for classification were hierarchical agglomerative clustering (HAC) and partitioning around medoids (PAM). These methods represent different statistical principles of classification. The HAC is a bottom-up clustering method. It starts by treating every single attribute profile as a single cluster. Then, in an iterative manner it merges the closest pair of clusters that satisfy certain similarity criteria until a stopping rule is met. In the current study, the Ward's minimum-variance method was used as the similarity measure for clustering. Iterations cease when the desired number of clusters is attained.

Different from the HAC procedure, the PAM procedure aims to find a desired number of elements from the data. These elements, called the medoid of each cluster, represent special features or aspects of the data and are those for which the average dissimilarity to all the elements of the same cluster is minimal. After finding the desired number of medoids, each element of the data set is then assigned to the nearest medoid.

In total, four model structures, two item discriminating levels, and two classification methods were manipulated so as to produce a total of 4 x 2 x 2=16 conditions. A simulation technique was used to create 100 random replications of data representing each condition.

Data generation

The data were generated using the principles of the attribute hierarchy method (AHM, Leighton, et al., 2004), a cognitive diagnostic model designed to accommodate the hierarchical relationships among the cognitive attributes in the construct of an assessment. The four attribute hierarchies in Figure 1 were used as a basis to generate the item response vectors. For each of the four attribute hierarchies, a Q-matrix and the attribute pattern matrices were first generated. An example of a partial Q matrix for the first 7 items derived from the divergent attribute hierarchy (Figure 1c.) is presented below. This Q matrix was repeated 5 times to get the 35-item Q matrix in the simulated test.

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (1)$$

An attribute pattern matrix contains all the attribute patterns that are possible based on the attribute hierarchy. After accounting for the interrelationships among the seven attributes in the four attribute hierarchies, a total of 8 to 65 attribute patterns were yielded. By examining the attributes measured by each item in the Q-matrix against the cognitive attributes in each row of the attribute pattern matrix, the expected response matrices were generated. An example of the attribute pattern matrix derived from the divergent attribute hierarchy (Figure 1c.) is presented below. In this matrix, row 3 represents a student who has mastered attributes 1 and 3. When comparing this attribute pattern against the Q matrix (Matrix 1) above for the first 7 items, the expected responses for the first 7 items would be 1010000 because item 1 measures attribute 1 only and item 3 measures attributes 1 and 3 while all the other items measure attributes other than these two attributes. The algorithm used here for generating the attribute pattern matrix and the expected response matrix can be found in Leighton, et. al. (2004).

$$
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 1 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & 1 & 0 & 1 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 1 \\
1 & 1 & 1 & 0 & 0 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 1 & 1 & 0 & 1 & 0 & 0 \\
1 & 1 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 1 & 0 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 1 \\
1 & 1 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 1 & 1 & 0 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 0 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix} \tag{2}
$$

When generating data, item discriminating power was also manipulated. For an item of high discriminating power, the probability of producing a correct response was set at .9 if the student mastered all the attributes required by an item or .1 if the student did not master all the attributes. For the low discriminating power condition, the corresponding probabilities were set at 0.6 and 0.2, respectively. Please see Cui and Leighton (2009) for a rationale of this setting.

Based on the expected response matrices derived from the four attribute hierarchies and the different item discriminating power, eight data matrices were generated corresponding to each of the hierarchy/discrimination combinations. Each data matrix consisted of the item responses of 2000 simulated students to the 35 items. For each attribute pattern in a particular attribute pattern matrix, an approximately equal number of students were generated. For example, in the divergent hierarchy which has 26 distinct attribute patterns, item responses of 2000/26≈77 students were simulated for each pattern.

Analysis

The analysis was conducted in two steps. Step 1 was to obtain the cognitive profile for each student and step 2 was to classify the students into different profile groups and evaluate the results.

Step 1: After the data were generated, the expected response matrices and the student attribute pattern matrices were used to produce students' cognitive profiles, which contain the probabilities of mastery for each attribute. A neural network approach was used to calculate the attribute profiles (Gierl, Cui, & Hunka, 2008).

A neural network can be described as a model which transforms a stimulus received by an input unit to a signal for the output unit through a series of hidden units. The input to train the neural network was the expected response matrices, which contain the expected response patterns, called exemplars in the terminology of the neural network approach. For each expected response pattern there is a corresponding student attribute pattern. The association between the expected response patterns and the student attribute patterns is established by presenting each pattern to the network repeatedly until it learns the association with an acceptable error level. For instance, if 11001001 is an expected response pattern and 1000 is the attribute pattern, then the network is trained to associate the response pattern 11001001 with the attribute pattern 1000. After the training converges at the predetermined error level (in this study, error level was set at 0.001), two matrices of weights are produced. One of the weight matrices associates the input units with the hidden units and the other associates the hidden units with the output units. Using these weight matrices, the attribute probabilities

which compose the cognitive profiles can be produced for the empirically observed response patterns. The R statistical package "neuralnet" was used to conduct this step of analysis.

As dichotomous attribute classification is usually used in CDA (Rupp, et al., 2010), the attribute probabilities in the cognitive profiles produced from the neural network were transformed to either 1 or 0 to indicate mastery or non-mastery of an attribute. A threshold value of 0.8 was used. If the probability of an attribute was equal to or greater than 0.8, the attribute probability was transformed to 1. Otherwise, the attribute probability was transformed to 0. It should be noted that the value of 0.8 was arbitrarily chosen. Other threshold values could also be used. In an operational testing situation, mastery state would likely be determined using some type of attribute-based standard setting procedure.

After the cognitive profiles were transformed to 1s and 0s, the number of profiles in each of the eight attribute hierarchy/discrimination conditions was obtained. In addition, the percentages of the profiles that matched the expected profiles implied by the attribute hierarchies were also calculated. This information helps us understand how well the resultant cognitive profiles are related to the expected cognitive profiles in each condition. It also provides additional information for evaluating the classification results by showing the different degrees of diversity in the cognitive profiles in each condition.

Step 2: Though it is practical and useful to determine the number of profiles by substantive means, for the purposes of this study each simulee was classified into one of five groups based on the attribute probabilities in the cognitive profiles, using both the HAC and the PAM methods. We chose 5 profiles for demonstrative purposes, however, in practice, teachers could select the number of groups they feel comfortable to work with in their classroom.

After classification, the mean probabilities of each attribute were computed across simulees classified to the same cluster to produce a cluster cognitive profile for a particular cluster. These attribute probabilities in the cluster cognitive profiles were transformed to either 1 or 0 in the same manner as described in the previous step. The consistency between the transformed cluster cognitive profiles and the expected cognitive profiles implied by the attribute hierarchy in each condition was examined. Also investigated was the proportion of simulees whose cluster cognitive profiles were consistent with the expected cognitive

profiles. This step was necessary because if, after classification, students' cognitive profiles were no longer consistent with the attribute hierarchy based on which the test was constructed, the validity of the interpretation on these cognitive profiles would be called into question.

Next, the consistency between the two clustering methods was examined by calculating the proportions of cognitive profiles that are consistently clustered by both methods. The profiles that matched the mastery state of six or all seven of the seven attributes were considered consistently classified. The results from this analysis would provide information regarding how the profile classification and reduction was affected by different classification methods.

To evaluate the performance of the two different classification methods, two indices, the average silhouette width (ASW) and the Pearson gamma (PG) (Meila, 2007), were used. The ASW index compares the dissimilarity of elements within the same cluster against the dissimilarity between different clusters. Good clustering results would be indicated by low within-cluster dissimilarity and relatively large between-cluster dissimilarity. The ASW index is bounded by -1 and 1, with larger values indicating better classification while smaller values indicating poorer classification. The PG index is the Pearson correlation between the pairwise dissimilarity and a binary vector that is 0 for every pair of observations in the same cluster and 1 for every pair of observation in different clusters. The PG index emphasizes a good approximation of the dissimilarity structure by the clustering in the sense that observations in different clusters should be strongly correlated with large dissimilarity. Both the ASW and the PG indices are usually used to estimate the optimal number of clusters. However, as the number of clusters is fixed at five in the current study, these indices were used to compare the quality of clustering. For both indices, higher values indicate better clustering results. The clustering and validation of clustering results were carried out using the R statistical packages of "cluster" and "fpc".

Results

Descriptive Information of the Cognitive Profiles

The attribute probabilities computed using the neural network approach were transformed to cognitive profiles containing 1s and 0s for the 100 iterations under each of the eight attribute

hierarchy/discrimination conditions. The descriptive information of the number of cognitive profiles under each condition is presented in Table 1. The number of expected cognitive files implied by each type of hierarchy is shown in the second column. The last two columns present the mean and standard deviation of the number of cognitive profiles across the 100 iterations for each simulated condition. In general, less structured hierarchies tend to produce more distinct cognitive profiles. However, there were exceptions. The convergent hierarchy generated fewer cognitive profiles than the more structured linear hierarchy. In addition, the unstructured hierarchy generated fewer cognitive profiles than the more structured divergent hierarchy when the item discrimination is low.

Table 1. Number of Cognitive Profiles Generated in Simulated Conditions

Hierarchy	N of Expected Profiles	Discrimination	N	Mean	STD
Linear	8	**high**	100	36.58	3.42
		Low	100	67.96	3.07
Convergent	9	**high**	100	30.02	2.43
		Low	100	59.03	2.73
Divergent	26	**high**	100	59.16	2.86
		Low	100	82.93	2.35
Unstructured	65	**high**	100	72.46	1.61
		Low	100	72.94	1.96

Table 1 shows that considerably more cognitive profiles were produced than the number of expected profiles implied by each hierarchy. It would be desirable to know how many of those additional cognitive profiles were produced for each condition. Table 2 shows the percentage of cognitive profiles that were consistent with the expected profiles implied by each hierarchy under different item discrimination conditions. Not unexpectedly, the high discrimination condition typically produced greater percentages of cognitive profiles that matched the expected profiles. The unstructured hierarchy was again the exception, where both discrimination conditions produced almost the same proportion of cognitive profiles that are consistent with the

expected profiles. When combined with the information presented in Table 1, we see that even though numerous cognitive profiles were not consistent with the expected profiles, the majority of simulees were classified to clusters which had cognitive profiles that were consistent with the expected cognitive profiles. This included 75% to 99% of simulees for the low discrimination conditions and 93% to 99% of simulees for the high discrimination conditions. The information in Tables 1 and 2 helps us understand how well the resultant cognitive profiles are related to the expected cognitive profiles in each condition. It also provides additional information for evaluating the classification results by showing the different degrees of diversity in the cognitive profiles in each condition.

Table 2 Percentages of Cognitive Profiles Consistent with the Expected Profiles

Hierarchy	Number of Expected Profiles	Discrimination	N	Mean % of Consistent Profiles	STD
Linear	8	**high**	100	95.74	0.40
		low	100	74.81	0.91
Convergent	9	**high**	100	94.37	0.51
		low	100	76.98	0.89
Divergent	26	**high**	100	92.64	0.54
		low	100	75.95	0.96
Unstructured	65	**high**	100	99.15	0.22
		low	100	99.23	0.21

Overall Classification Results

A global analysis was done to investigate whether the cluster cognitive profiles were consistent with the expected cognitive profiles implied by the attribute hierarchies. Also investigated was the proportion of simulees whose cluster cognitive profiles were consistent with the expected cognitive profiles. More simulees classified to cognitive profiles that are consistent with the expected profiles would indicate that it is feasible to do the profile classification and reduction. Figure 2 graphically presents such results and it indicates that when the discrimination is high,

both methods performed well: 100% or close to 100% of the simulees were classified to cognitive profiles that were consistent with the expected profiles. However, when the item discrimination was low, fewer simulees were classified to cognitive profiles that were consistent with the expected profiles for the linear and convergent hierarchy, although the consistency rate remained at over 85%.

This global analysis also produced information regarding the performance of the HAC and the PAM method. For example, the PAM method slightly outperformed the HAC method in that it classified more simulees into clusters which had cognitive profiles consistent with the expected cognitive profiles when the item discrimination is high. However, when the discrimination was low, the HAC method outperformed the PAM method for the more structured linear and convergent hierarchies, while the PAM method outperformed the HAC method for the less structured divergent hierarchy.

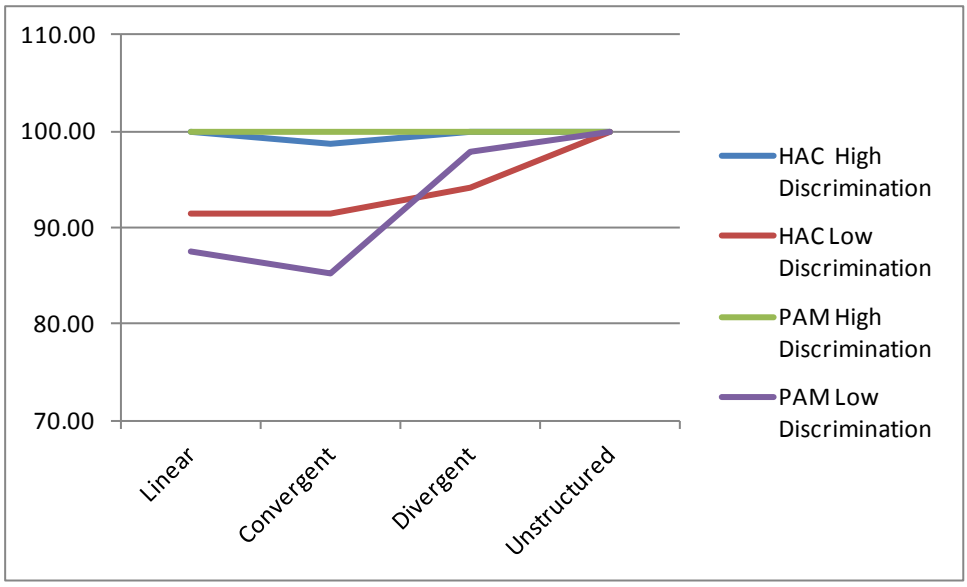

Figure 2. Percentages of cognitive profiles consistent with expected profiles using the HAC and PAM methods

Consistency between the Two Clustering Methods

After the simulees were classified into 5 different clusters, the consistency between the results from the two methods was examined. The degree of consistency between the results from both

methods would indicate how robust the profile classification and reduction was in terms of the classification methods used. Table 3 presents the percentages of cognitive profiles that matched on 6 out of the 7 attributes using the two clustering methods as well as those that matched on all 7 attributes. As the hierarchy became more unstructured from linear hierarchy to unstructured hierarchy, the classification consistency between the results from the two methods decreased in terms of the percentage of cognitive profiles that match on all 7 attributes. In addition, the classification on hierarchies with higher discriminating items tend to show better consistency than the classification on hierarchies with lower discriminating items, with the exception of the unstructured hierarchy, where better consistency was achieved when the hierarchy contained lower discriminating items. If cases of exact match and those which had 6 out of 7 attributes matched using the two methods are both regarded as consistent classification, more than 80% of the cognitive profiles were classified consistently for the linear and convergent hierarchies with higher discriminating items. Lower percentages of consistency were produced for other hierarchy/discrimination conditions. The lowest consistency rate occurred for the condition of unstructured hierarchy with high discriminating items. These results were graphically presented in Figure 3.

Table 3. Percentages of Cognitive Profiles Consistently Classified by the Two Methods

		% of Match on 6 Attributes		% of Match on 7 Attributes	
Hierarchy	**Discrimination**	**Mean**	**STD**	**Mean**	**STD**
Linear	**High**	36.84	17.76	47.20	16.73
	Low	36.53	18.53	32.80	16.68
Convergent	**High**	40.22	16.23	43.54	16.94
	Low	38.44	14.33	33.33	14.25
Divergent	**High**	39.22	16.03	29.07	13.89
	Low	35.57	13.84	27.20	12.98
Unstructured	**High**	37.07	13.04	14.64	8.36
	Low	39.75	11.33	23.49	11.98

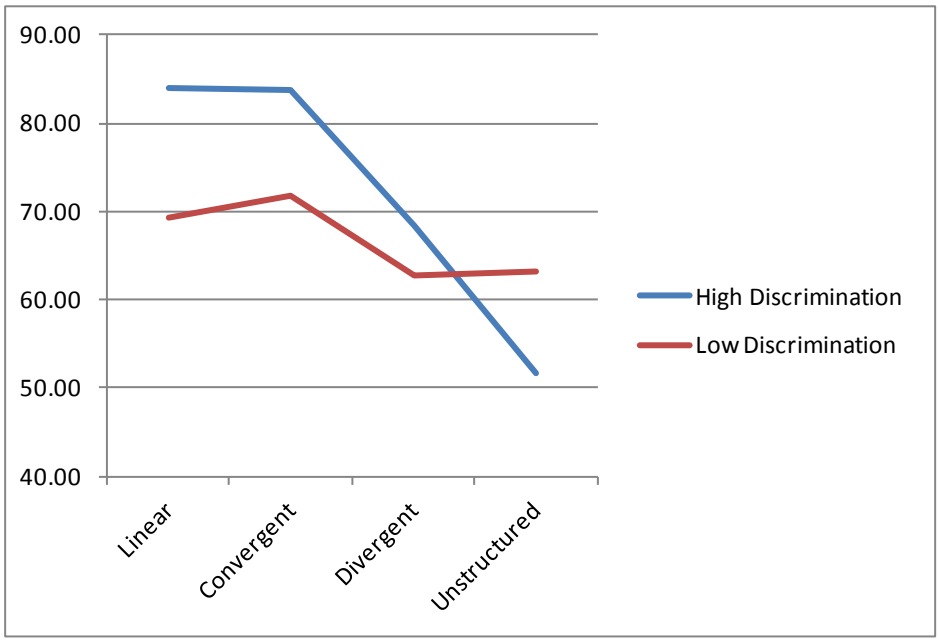

Figure 3. Consistency Rate for the Eight Hierarchy/Discrimination Conditions

Evaluation of the Two Classification Methods

In a previous section, it was indicated that the two classification methods performed differently under different simulated conditions in terms of the percentage of cognitive profiles that were consistent with expected profiles. In this section, the classification results and the performance of the two methods were examined using the ASW index and the PG index. The values of the indices for each condition are presented in Table 4 as well as graphically in Figures 4 and 5.

Table 4. The ASW and PG Indices for the HAC and PAM Methods

Method	Discrimination	Hierarchy	ASW		PG	
			Mean	STD	Mean	STD
HAC	High	Linear	0.42	0.03	0.69	0.01
		Convergent	0.40	0.03	0.68	0.02
		Divergent	0.22	0.03	0.52	0.03
		Unstructured	0.13	0.01	0.39	0.02

	Low	**Linear**	0.29	0.03	0.62	0.03
		Convergent	0.26	0.03	0.57	0.03
		Divergent	0.15	0.02	0.44	0.03
		Unstructured	0.11	0.01	0.37	0.02
PAM	**High**	**Linear**	0.52	0.03	0.67	0.01
		Convergent	0.48	0.02	0.67	0.01
		Divergent	0.24	0.02	0.54	0.02
		Unstructured	0.15	0.01	0.41	0.01
	Low	**Linear**	0.31	0.02	0.59	0.02
		Convergent	0.32	0.01	0.58	0.01
		Divergent	0.19	0.01	0.48	0.02
		Unstructured	0.15	0.01	0.43	0.01

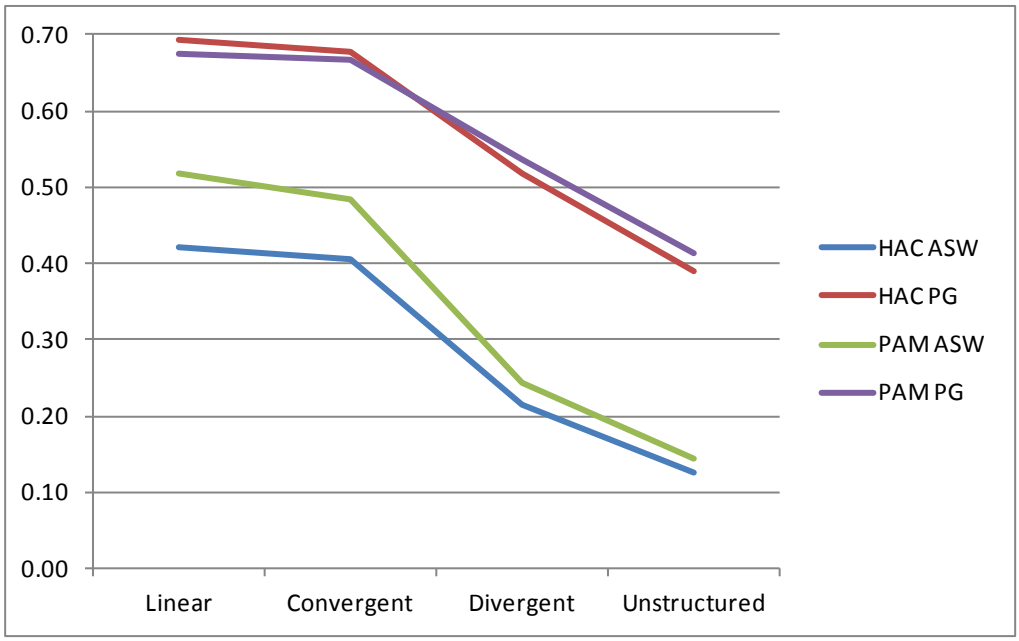

Figure 4. The ASW and PG indices for high discriminating conditions.

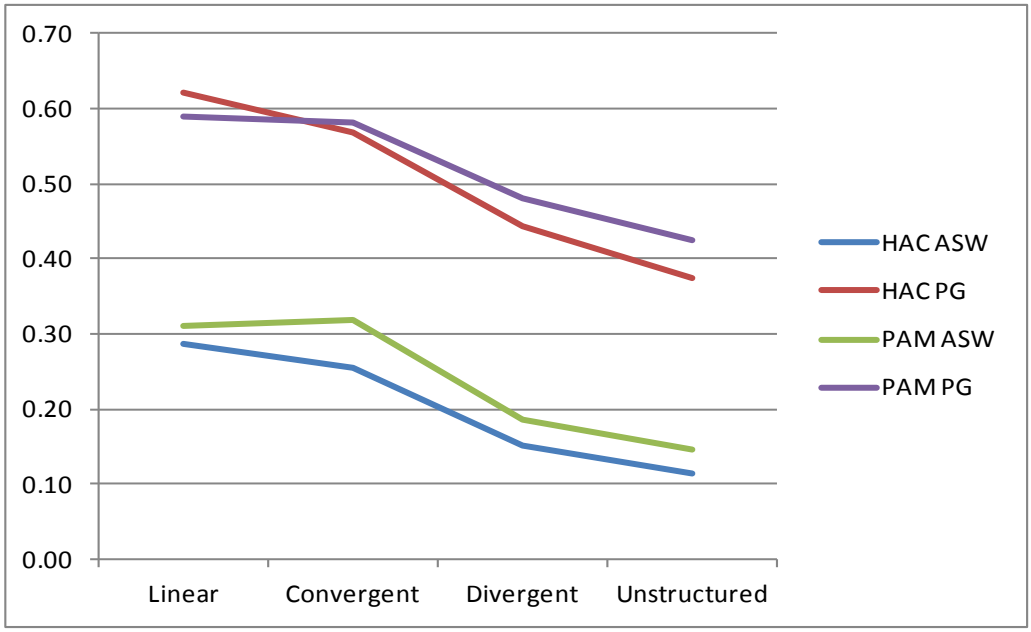

Figure 5. The ASW and PG indices for low discriminating conditions.

From Table 4 and Figures 3 and 4, it can be seen that as the hierarchies become less structured, the clustering results became poorer, regardless the item discrimination conditions or the cluster validation index being used. However, for more structured hierarchies, the high discrimination conditions resulted in better classification results than the low discrimination conditions.

If the ASW index is used as the sole evaluation criterion, the PAM method yielded better clustering results for all the hierarchy/ discrimination conditions. However, if both indices are used for the evaluation, then the HAC method yielded better clustering results for the linear hierarchy under both discrimination conditions and the convergent hierarchy under the high discrimination condition, while the PAM method yielded better clustering results for the remaining conditions. Hence, there was no clear cut indication of which method performed better than the other method.

Discussion

The current study aimed to investigate the efficiency and accuracy of classification methods in reducing the number of cognitive profiles using simulated data. Two classification methods, the HAC and the PAM methods, were used to classify the cognitive profiles generated by four types of attribute hierarchies containing items with two different levels of discrimination. The four

types of hierarchies imply different number of expected cognitive profiles (Table 1). When slips and guessing were introduced to the data files by manipulating the item discrimination, considerably more distinct cognitive profiles were produced than the number of expected cognitive profiles (Table 1). However, a majority of the simulees were classified to clusters which had cognitive profiles that were consistent with the expected cognitive profiles (Table 2). The cognitive profiles in each condition were classified into five groups using each of the two classification methods. In CDA, the cognitive model, in this case the attribute hierarchy, plays the fundamental role because it guides test construction and implies different cognitive profiles based on which inferences can be made about student's cognitive strengths and weaknesses. Thus the feasibility of profile classification would be questionable if the cognitive profiles after classification are not consistent with the expected cognitive profiles implied by the attribute hierarchies. In the current study, results indicated that when the items in the simulated tests had high discrimination, 100% or close to 100% of the classified cognitive profiles were consistent with the expected cognitive profiles. When the items had low discriminations, a minimum of 85% of the classified cognitive profiles were consistent with the expected cognitive profiles, indicating that it is feasible to further classify the numerous cognitive profiles implied by the attribute hierarchies. Further, when item discrimination was low, the more structured linear and convergent hierarchies suffered more in terms of the percentage of cognitive profiles that are consistent with the expected profiles. This might be due to the fact that the more structured hierarchies tend to imply a smaller set of expected cognitive profiles. When the item discrimination was low and deviation from the expected profile occurred, the resultant cognitive profiles tend to fall out of the small set of expected cognitive profiles.

The two classification methods, the HAC and the PAM methods, were used to classify the cognitive profiles of the students. Results show that the two methods could consistently classify larger percentages of cognitive profiles for more structured hierarchies than for less structured hierarchies. In addition, classification on cognitive profiles inferred from tests with high-discriminating items tends to be more consistent from the two methods. If both the cases of exact match and the cases where six out of the seven attributes matched on the classified cognitive profiles are regarded as classification consistency, 70% to over 80% of the cognitive profiles from more structured hierarchies were consistently classified. Such results indicate that, if further

classification and reduction of cognitive files are needed, more structured attribute hierarchies would be preferred.

The performance of two classification methods was also compared. Results indicated that the two classification methods performed differently under different simulated conditions in terms of the percentage of classified cognitive profiles that were consistent with expected profiles. The ASW and the PG indices were also used to quantify the classification results and compare the two methods. If the ASW index was used alone, then the PAM method outperformed the HAC method in all simulated conditions. When both indices were used, although PAM method outperformed the HAC method in most conditions, for the more structured hierarchies, the HAC method had advantages. Hence, there was no clear cut indication of which method performed better than the other method although the PAM method performed better in more of the simulated situations. Such results may be related to the nature of the two classification methods, the two validation indices, as well as the interrelationships among the expected cognitive profiles implied by the attribute hierarchies. Further studies could be conducted to investigate the interplay among the different factors.

Conclusion

The study explored the effectiveness and accuracy of two classification methods in reducing the number of cognitive profiles under eight different conditions. Results from the study indicated that it is feasible to further reduce the numerous cognitive profiles into fewer profiles. However, no clear cut indication of which method performed consistently better than the other method. Therefore, future research can build on the results and further evaluate the effects of more parameters and conditions so that the more effective method could be selected when profile classification is needed. It was also found that less structured hierarchies did not perform as well as the more structured hierarchies. However, it was not clear whether it was the structure of the hierarchies or simply the number of expected cognitive profiles (keep in mind that the more structured linear and convergent hierarchies imply far fewer expected cognitive profiles than the other two less structured divergent and unstructured hierarchies) that was the key factor causing such results. Thus, the reasons for the differential profile classification results should be further explored.

Another notable point is, when generating data, approximately equal number of students were generated for each attribute pattern in a particular attribute pattern matrix. This method of data generation does not necessarily produce simulee samples with normally distributed student abilities, which is usually assumed in the setting of educational assessment. Thus, further studies can be conducted to investigate profile classification results with this different student ability distribution.

In practice, the study serves as one building block to bridge the gap between CDA research and educational practice, namely, how to apply the cognitive information produced from CDA to the teaching and learning process and practice (Wang & Gierl, 2011). With the results from the study, practitioners will have some referential resources when attempting to apply CDA results into educational practice. However, in the current study, details were not given about the features of cognitive profiles which tend to cluster together during profile classification. This information is important because it may shed light on the interrelationships among the cognitive attributes, which may eventually lead to more accurate specification of cognitive models in CDA. Thus, in future studies, these features can be further examined.

Reference

Cui, Y., & Leighton, J. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(4), 429-449.

Gierl, M. J., Cui, Y., & Hunka, S. (2008). Using connectionist models to evaluate examinees' response patterns on tests. *Journal of Modern Applied Statistical Methods, 7*, 234-245.

Gierl, M. J., Cui, Y., & Zhou, J. (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(3), 293-313.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205-236.

Meila, M. (2007). Comparing clusterings: An information based distance. *Journal of Multivariate Analysis, 98*, 873-895.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and practice*. New York: Guilford.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptioins based on item response theory. *Journal of Educational Measurement, 20*(4), 345-354.

Wang, C., & Gierl, M. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement, 48*(2), 165-187.