

Simulation Studies for Evaluating  
the Performance of the Two Classification Methods in the AHM

Ying Cui

Jacqueline P. Leighton

Yinggan Zheng

Centre for Research in Applied Measurement and Evaluation (CRAME)

University of Alberta

### **Abstract**

The attribute hierarchy method (AHM) (Leighton, Gierl, & Hunka, 2004), which is based on the assumption that test items can be described by a set of hierarchically ordered attributes, is designed to estimate examinees' level of competency as well as their profiles for mastering a set of attributes. The AHM, by incorporating the assumption of attribute dependency, brings an important cognitive property into cognitive diagnostic testing. This study used simulated response vectors to evaluate the performance of the two classification methods (method A and B) for the AHM. The performance of the two classification methods was evaluated at the attribute pattern level and the individual attribute level, respectively. Our results suggest that classification method B outperforms method A at both the attribute pattern and the individual attribute level. In addition, the performance of two methods at the individual attribute level is considerably higher than their performance at the attribute pattern level.

## **Simulation Studies for Evaluating the Performance of the Two Classification Methods in the AHM**

By estimating a person's location on an underlying latent continuum, traditional assessments have been effective for selecting students who are most likely to succeed in a particular educational institution or program (Mislevy, 1995). Traditional assessments are typically constructed on logical taxonomies and content specifications but lack explicit cognitive models of the structures and cognitive processes that underlie student performance (Snow & Mandinach, 1991). As a result, test scores from traditional assessments are tied to content areas rather than the examinee's cognitive processes measured by test items.

In addition, test theories used for interpreting scores from traditional assessments are designed to optimize the estimate of a student's single score on an underlying latent scale – the true score scale in classical test theory (CTT) or the latent trait scale in item response theory (IRT). A single aggregate score produced using CTT and IRT provides general information about students' location on a continuum. However, it fails to provide specific information to inform teachers about their students' cognitive strengths and weaknesses which may, in turn, help teachers make instructional decisions intended to help students succeed in educational settings (Nichols, 1994).

Frustrated by the presence of these two limitations with traditional assessment approaches, measurement specialists have become increasingly interested in the development of new diagnostic assessments that are aimed at uncovering the cognitive processes used by students to respond to test items, determining the nature of poor performance, and classifying the poor performance in terms of an accepted typology of

malfunctions (Scriven, 1999). As Nichols (1994) stated:

These new assessments make explicit the test developer's substantive assumptions regarding processes and knowledge structures a performer in a test domain would use, how the processes and knowledge structures develop, and how more competent performers differ from less competent performers. (p. 578)

New diagnostic assessments enable researchers and educators to make inferences about cognitive processes and knowledge that students use when solving test items. A well-designed diagnostic assessment can measure different cognitive processes and knowledge required to solve test items in a domain of interest. Diagnostic assessments can also provide a profile of students' mastery and non-mastery of cognitive skills. The value of diagnostic assessment lies in its ability to reveal each student's specific cognitive strengths and weaknesses and further help design effective interventions for individual students.

Cognitive diagnostic models (CDMs) have been developed to help construct diagnostic assessments and estimate students' attribute mastery patterns associated with different cognitive skills. Leighton, Gierl, and Hunka (2004; see also Gierl, Leighton, & Hunka, 2000) proposed a CDM called the Attribute Hierarchy Method (AHM). The AHM is based on the assumption that test items can be described by a hierarchically-ordered set of attributes. Attributes are defined as basic cognitive processes or skills required to solve test items correctly (Leighton et al., 2004). The attribute hierarchy can be used as a basis for the development of test items, which upon administration to examinees, produces observed response vectors. Examinees' attribute pattern can be estimated by classifying each observed response vector into one of the expected response patterns. Expected response patterns are those response patterns that can be clearly explained by the

presence or absence of the attributes without any errors or “slips”. Expected response patterns and associated attribute patterns can be derived from the hierarchy. Given the importance of correctly identifying examinees’ attribute patterns, the purpose of this study was to use simulation approaches to evaluate the performance of two classification methods (Method A and Method B) for the AHM (see Leighton et al., 2004). To begin, we present a brief overview of the AHM.

### **An Overview of the Attribute Hierarchy Method**

The AHM is a cognitive diagnostic model designed to estimate examinees’ level of competency as well as profiles that reflect their mastery for a set of attributes. The AHM is based on the assumption that test items can be described by a set of hierarchically-ordered attributes. Attributes are defined as basic cognitive processes or skills required to solve test items correctly (Leighton et al., 2004). The first step in using the AHM for cognitive diagnosis is to identify the attributes in the domain or for the task of interest. Correctly identifying the attributes influences the validity of the inferences about examinees made with the AHM. As mentioned by Leighton et al. (2004), methods from cognitive psychology, such as task and protocol analysis, could play an important role in the identification of attributes in a domain. Many studies have been conducted to identify the attributes required for successful performance on test items and tasks. For example, in a language testing study, Buck and Tatsuoka (1998) created the hypothetical attribute set for a 35-item listening comprehension test by using two main sources: an extensive literature review to seek the theoretical and empirical evidence for the attributes that affect performance on listening tests and the results from a series of verbal protocol studies conducted by Buck (1990, 1991, 1994) for examining the second language

listening processes.

In the AHM, attributes are considered to be hierarchically related and therefore can be ordered into a hierarchy based upon their logical and/or psychological properties. As explained by Leighton et al. (2004), the assumption of attribute dependency is consistent with the conclusion that “cognitive skills do not operate in isolation but belong to a network of interrelated competencies (Kuhn, 2001; Vosniadou & Brewer, 1992)” (p. 209). The ordering of the attributes into a hierarchy should be based on “empirical considerations (e.g., a series of well defined, ordered cognitive steps identified via protocol analysis) or theoretical considerations (e.g., a series of developmental sequences suggested by Piaget such as preoperational, concrete operational, and formal operational)” (Leighton et al., 2004, p. 209).

Once the attribute hierarchy is identified, a series of matrices (e.g., the adjacency, reachability, incidence, reduced incidence, and expected response matrices), initially introduced by Tatsuoka (1983, 1995, 1996), can be derived to facilitate the development of test items and the estimation of students’ profiles of their mastery and non-mastery of cognitive skills. A binary adjacency matrix ( $A$ ) of order  $K \times K$  specifies the direct relationship between each pair of attributes, where  $K$  is the number of attributes. To specify the direct and indirect relationship among attributes, a reachability matrix ( $R$ ) of order  $K \times K$  is used. The  $R$  matrix can be obtained using the equation  $R = (A + I)^n$ , where  $I$  is an identity matrix of order  $K \times K$ , and  $n$  is the integer between 1 and  $K$  that leads  $R$  to become invariant. That is, if  $(A + I)$  times itself repeatedly using Boolean algebra until the product become invariant, then the obtained matrix is the  $R$  matrix.

The potential item pool, represented by an incidence matrix ( $Q$ ) of order

$K \times (2^K - 1)$ , contains items that measure all the possible combinations of attributes when the attributes are assumed to be independent of each other. The columns of the  $Q$  matrix are obtained by converting the integers ranging from 1 to  $2^K - 1$  to their binary form. In the  $Q$  matrix, each column represents one item, and the 1s in the column identify which attributes are required for successful performance on this item. When the attributes share dependencies, the size of the potential item pool can be significantly reduced by imposing the constraints of the attribute hierarchy as embodied in the  $R$  matrix. The removal of items that do not match the constraints of the  $R$  matrix ultimately produces a reduced incidence matrix ( $Q_r$ ). The  $Q_r$  matrix can be derived by determining which columns of the  $Q_r$  matrix are logically included in each column of the  $R$  matrix using Boolean addition. By describing the cognitive requirements of the domain of interest with the attribute hierarchy and specifying items needed to measure the domain in the  $Q_r$  matrix, the AHM attempts to make a direct link between student cognition and test design.

Once the  $R$  matrix and the  $Q_r$  matrix are identified, the expected response matrix ( $E$ ) can be created. The  $E$  matrix is composed of the response patterns that can be clearly explained by the presence or absence of the attributes without any errors or “slips” given that the attribute hierarchy is true. The rows of the  $E$  matrix represent the expected examinees who possess cognitive attributes that are consistent with the hierarchy.

*Expected* examinees do not make errors or slips that produce inconsistencies between the observed and expected response patterns.

In real testing situations, it is possible that an examinee, who has not mastered all the attributes required by an item, can still answer the item correctly by guessing or by having partial knowledge. It is also possible that an examinee who has mastered all the

attributes that an item is probing might answer the item incorrectly due to careless mistakes. Therefore, the *observed* examinee response vectors might reflect slips of the form 1 to 0 or 0 to 1. By classifying each observed response vector into one of the expected response patterns, examinees' attribute mastery can be estimated.

Leighton et al. (2004) proposed two methods for the classification of observed response patterns in the AHM. In these two methods, the probability of a correct response to individual items is calculated for each expected response pattern using an IRT model. The three-parameter logistic IRT model is given by:

$$p(x_{ij} = 1 | \theta_i, a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + e^{-1.7a_i(\theta_j - b_j)}},$$

where  $a_j$  is the item discrimination parameter for item  $j$ ,  $b_j$  is the item difficulty parameter for item  $j$ ,  $c_j$  is the pseudo-guessing parameter for item  $j$ , and  $\theta_i$  is the ability parameter for examinee  $i$ . The two-parameter logistic IRT model is a special case of the three-parameter model in which the  $c_j$  parameter is set to 0. The one-parameter model also called Rasch model is another form of the logistic IRT model in which all the items are assumed to have equal discrimination power and no guessing. Item parameters can be estimated based on the expected response patterns using BILOG 3.11 (Mislevy & Bock, 1990).

Once item parameters and the theta value associated with each expected response pattern are estimated, the IRT probability of a correct response for each item can be calculated for each expected response pattern. In Method A, an observed response pattern is compared against each of the expected response patterns to identify the slips from 1 to 0 and from 0 to 1. The likelihood of all slips from 1 to 0 and from 0 to 1 for examinee  $i$

is given by:

$$P_{ijExpected}(\theta_j) = \prod_{k \in S_{i0}} P_{jk}(\theta_j) \prod_{m \in S_{i1}} [1 - P_{jm}(\theta_j)],$$

where  $S_{i0}$  is the subset of items with slips from 0 to 1 for the observed response vector of examinee  $i$ , and  $S_{i1}$  is the subset of items with slips from 1 to 0. The higher the value of  $P_{ijExpected}(\theta_j)$  calculated by comparing the observed response vector to one of the expected response vectors, the more likely the observed response pattern originates from that expected response vector. Therefore, the observed response vector will be classified as originating from expected response vector  $j$  when the maximum value of  $P_{ijExpected}(\theta_j)$  is achieved.

In Method B, the expected response patterns that are logically included in the observed examinee response vector are identified and an examinee is considered to possess all attributes logically included within his/her observed response vector. For those expected response patterns that are not logically included in the observed vector, the likelihood of slips only from 1 to 0 is calculated and compared to a cut-point assigned by researchers. The likelihood of slips from 1 to 0 is given by:

$$P_{ijExpected}(\theta_j) = \prod_{k \in S_{i1}} [1 - P_{jm}(\theta_j)].$$

The observed response vector will be classified as originating from expected response vector  $j$  when the maximum value of  $P_{ijExpected}(\theta_j)$  is achieved.

### Method

For the purpose of this research, the attribute hierarchies illustrated in Figure 1 were separately used as the basis for the simulation. These hierarchies correspond to the

hierarchical structures discussed by Leighton et al. (2004). For each attribute hierarchy, the matrices of the AHM, including the adjacency matrix, the reachability matrix, the  $Q$  matrix, the  $Q_r$  matrix, and the expected response matrix, were derived. A sample of 5000 expected item response vectors was generated based on each of the four expected response matrices with the constraint that the total scores associated with the expected response patterns be normally distributed. Given that each of the four generated samples only consists of expected response patterns which are free from slips, the observed item response vectors were generated by randomly adding slips to each of the expected response patterns. In this study, the percentage of random errors was manipulated to range from 5% to 20% with an increment of 5% of the total number of item responses to examine whether the number of random errors has an impact on the accuracy of two classification methods. In order to generate 5% random errors, for example, 5% of correct responses to each item were randomly selected and changed from 1 to 0. Conversely, 5% of incorrect responses to each item were randomly selected and changed from 0 to 1. Therefore, for each item,  $0.05 \times 5000 = 250$  slips were added on the expected responses to the item. In total, 4 (hierarchical structure) X 4 (percentage of random errors) = 16 conditions were considered in the current study. For each condition, 100 replications were simulated with Mathematica 5.2 (Wolfram Research).

Given that each simulated response vector was produced by creating slips on the expected response vector associated with a known attribute pattern, the performance of two classification methods was evaluated by comparing the estimated attribute pattern to its known true attribute pattern. The performance was evaluated at two different levels: the attribute pattern level and the individual attribute level. At the attribute pattern level,

attribute patterns estimated by two classification methods were separately compared to the true pattern for each response vector. Considering the misclassification of an examinee's attribute pattern could be nearly negligible if only one or two attributes are misclassified, or fairly severe if most of attributes are misclassified, the performance of two methods was also evaluated at the individual attribute level. To accomplish this, the number of attributes that were assigned to the correct value (1 or 0) by each method was counted for each simulated response vector. High overall average classification rates at both attribute pattern and individual attribute level suggest the corresponding classification method works well for the AHM.

### **Results and Discussions**

Table 1 summarizes the classification results of method A and B for the four attribute hierarchies illustrated in Figure 1. The classification rates at both the attribute pattern and the individual attribute level are presented. Due to the limited space, only results for the first hierarchy in Figure 1 will be discussed in detail. Table 2 presents the expected response patterns and their associated attribute patterns derived from hierarchy 1. As shown in the first column in Table 2, the expected response pattern derived from the first attribute hierarchy indicates that 25 items are needed in order to achieve maximum diagnostic information over the attributes. The second column gives the attribute pattern associated with each expected response pattern. The third column provides the frequencies for expected response patterns used to generate the 5000 expected item response vectors.

For the data sets that contain 5% random errors, a total number of  $0.05 \times 5000 \times 25 = 6250$  slips over all the items, including slips from 1 to 0 and from 0 to

1, were created on each simulated data set. The agreement between the true attribute pattern and the estimated attribute pattern occurred 53.68% of the time for method A and 71.21% of the time for method B. At individual attribute level, an average of 6.19 estimated attribute values out of 7 agreed with the true attribute values by Method A and an average of 6.42 by Method B. Therefore, the average classification rates for methods A and B at the individual attribute level were 88.36% and 91.71%, respectively. Thus, method B outperformed method A at both the attribute pattern and the individual attribute level for the data sets that contain 5% random errors.

For each data set generated by randomly adding 10% errors on the expected response patterns derived from the first hierarchy, the total number of slips over all the items was  $0.10 \times 5000 \times 25 = 12500$ , including slips from 1 to 0 and from 0 to 1. The agreement between the true attribute pattern and the estimated attribute pattern occurred 34.94% of the time for method A and 51.12% of the time for method B. The average classification rates for the two methods at the individual attribute level were 81.16% and 84.91%, respectively. Therefore, method B outperformed method A at both the attribute pattern and the individual attribute level for the data sets that contain 10% of random errors.

When 15% errors were added on the expected response patterns of the first attribute hierarchy, a total number of  $0.15 \times 5000 \times 25 = 18750$  slips were generated for each data set. The average classification rates for two classification methods at the attribute pattern level were 25.55% and 37.16%, respectively. At the individual attribute level, the classification rates for method A and method B are 77.29% and 79.32%, respectively.

For the data sets containing 20% random errors, the average classification rates for classification method A and B at the attribute pattern level were 19.94% and 27.24%, respectively. At the individual attribute level, the classification rates for method A and method B were 73.80% and 74.54%, respectively.

Therefore, for the data sets generated based on the first attribute hierarchy, method B consistently outperformed method A at both attribute pattern level and individual attribute level. When the number of random errors increased, the performance of methods A and B decreased at both the attribute pattern and the individual attribute levels.

### **Conclusions and Future Directions**

This study used simulated response vectors to evaluate the performance of the two classification methods for the AHM. Relatively speaking, classification method B outperformed method A at both the attribute pattern and the individual attribute level. Our results indicate that both methods work poorly at the attribute pattern level. However, the performance of these two classification methods at the individual attribute level is considerably higher than their performance at the attribute pattern level.

In future research, new classification methods for the AHM will be investigated. The classification approach currently being investigated is to use neural networks in the AHM. One benefit of using neural networks is that they do not rely on IRT models and assumptions about the distribution of examinees. Rather, they can be used to estimate the probabilities that examinees have mastered each attribute by minimizing the error associated with the estimation.

## Reference

- Buck, G. (1990). *The testing of second language listening comprehension*. Unpublished doctoral dissertation, University of Lancaster, England.
- Buck, G. (1991). The testing of listening comprehension: an introspective study. *Language Testing, 8* (1), 67-91.
- Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing, 11* (2), 145-170.
- Buck, G., & Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*, 119-157.
- Gierl, M. J., Leighton, J. P., & Hunka, S. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice, 19*, 34-44.
- 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*(3), 205-237.
- Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. F. Chipman, & P. L. Brennan (Eds.), *Cognitively Diagnostic Assessment*, Hillsdale, NJ: Erlbaum.
- Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: Item analysis and test scoring with binary logistic test models [Computer Program]*. Mooreville, IN: Scientific Software.
- Nichols, P. D. (1994). *A Framework for Developing Cognitively Diagnostic Assessment*.

*Review of Educational Research*, 64 (4), 575-603.

Scriven, M. (1999). The Nature of Evaluation Part I: Relation to psychology. *Practical Assessment, Research & Evaluation*, 6 (11).

Snow R. E. & Mandinach, E. B. (1991). *Integrating assessment and instruction: A research and development agenda* (ETS Research Rep. No RR-91-8). Princeton, NJ: Educational Testing Service.

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. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. Nichols, S. F. Chipman, & P. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 327-359), Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1996). Use of generalized person-fit indexes, Zetas for statistical pattern classification. *Applied Measurement in Education*, 9, 65-75.

Figure 1

Four 7-attribute hierarchies used for simulation

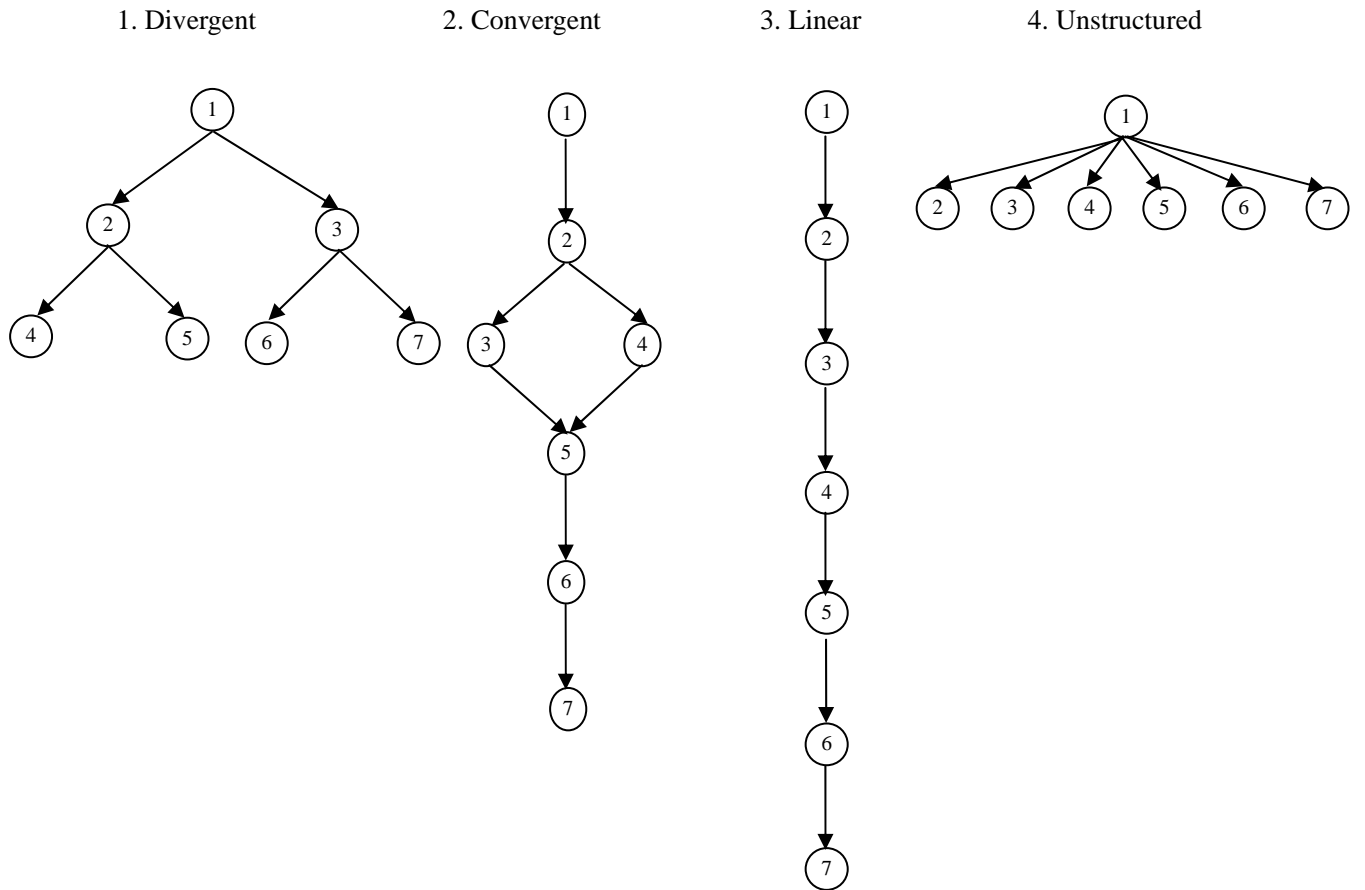


Table 1

*The classification results for methods A and B.*

Hierarchy	Classification Method	Level of analyses	Percentage of random errors			
			5%	10%	15%	20%
H1 (divergent)	A	Attribute Pattern	53.68%	34.94%	25.55%	19.94%
		Individual attribute	88.36%	81.86%	77.29%	73.80%
	B	Attribute Pattern	71.21%	51.12%	37.16%	27.24%
		Individual attribute	91.71%	84.91%	79.32%	74.54%
H2 (convergent)	A	Attribute Pattern	75.99%	58.50%	46.07%	36.89%
		Individual attribute	94.53%	90.24%	86.85%	84.12%
	B	Attribute Pattern	62.40%	52.00%	43.07%	35.39%
		Individual attribute	91.51%	87.08%	82.99%	79.19%
H3 (Linear)	A	Attribute Pattern	65.53%	52.62%	42.71%	35.20%
		Individual attribute	93.15%	89.43%	86.21%	83.42%
	B	Attribute Pattern	80.00%	63.77%	50.92%	40.45%
		Individual attribute	92.88%	86.75%	81.39%	76.75%
H4 (unstructured)	A	Attribute Pattern	32.58%	18.76%	11.82%	7.61%
		Individual attribute	77.99%	71.39%	68.02%	65.31%
	B	Attribute Pattern	67.41%	47.67%	34.87%	26.13%
		Individual attribute	92.61%	86.63%	81.67%	77.36%

