
Attribute Reliability in Cognitive Diagnostic Assessment

**Jiawen Zhou
Mark J. Gierl
Ying Cui**

Centre for Research in Applied Measurement and Evaluation
University of Alberta



**Paper Presented at the Paper Session
"Cognitive Diagnosis"**

Annual meeting of annual meeting of the National Council on
Measurement in Education
San Diego, CA

April 14, 2009

Abstract

The attribute hierarchy method is a psychometric procedure for classifying examinees' test item responses into a set of structured attribute patterns associated with different components from a cognitive model of task performance. Results from an AHM analysis yield information on examinees' cognitive strengths and weaknesses. Hence, the AHM can be used for cognitive diagnostic assessment. The purpose of this study is to introduce and evaluate a new concept for assessing attribute reliability using the ratio of true score variance to observed score variance on items that probe specific cognitive attributes. This reliability procedure is evaluated and illustrated using both simulated data and student response data from a sample of algebra items taken from the March 2005 administration of the SAT. The reliability of diagnostic scores and the implications for practice are also discussed.

Acknowledgements

The research reported in this study was conducted, in part, with funds provided to the second author by the College Entrance Examination Board. We would like to thank the College Entrance Examination Board for their support. However, the authors are solely responsible for the ideas, methods, procedures, and interpretations expressed in this study. Our views do not necessarily reflect those of the College Entrance Examination Board.

Cognitive diagnostic assessment (CDA) is a form of testing that employs a cognitive model to, first, develop or identify items that measure specific knowledge and skills and, then, use this model to direct the psychometric analyses of the examinees' item response patterns to promote specific test score inferences. CDAs are designed to measure these specific knowledge structures and processing skills in order to provide examinees with information about their cognitive strengths and weaknesses. A cognitive model in educational measurement refers to a "simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills students at different levels of learning have acquired and to facilitate the explanation and prediction of students' performance" (Leighton & Gierl, 2007, p. 6). Cognitive models are generated by studying the knowledge, processes, and strategies used by examinees as they respond to items. One benefit of developing or identifying items and analyzing data according to a cognitive model stems from the detailed information that can be obtained about the knowledge and skills that examinees' actually use to solve test items. In fact, some cognitive psychologists are now urging educational measurement specialists to develop assessment procedures using cognitive models. Pellegrino, Baxter, and Glaser (1999) claimed that:

...it is the pattern of performance over a set of items or tasks explicitly constructed to discriminate between alternative profiles of knowledge that should be the focus of assessment. The latter can be used to determine the level of a given student's understanding and competence within a subject-matter domain. Such information is interpretative and diagnostic, highly informative, and potentially prescriptive. (p.335)

In short, CDAs have the potential for identifying examinees' problem-solving strengths and weaknesses, particularly when the assessments are created from cognitive models that provide a contemporary representation of the knowledge structures and processing skills that are believed to underlie conceptual understanding in a particular domain. CDA results could also be integrated into the teaching and learning process because this form of assessment supports specific inferences about the examinees' problem-solving skills that could be linked with specific instructional methods designed to improve these cognitive skills.

In an attempt to uncover the diagnostic information that may be embedded in examinees' item response data and to address the challenge posed by Pellegrino et al. (1999), psychometric procedures have been developed to support test score inference based on cognitive models of test performance. These cognitive diagnostic models contain parameters that link item features to the examinees' response patterns so inferences about declarative, procedural, and strategic knowledge can be made. Some early examples include the *rule space model* (Tatsuoka, 1983) and the *linear logistic test model* (Fischer, 1973). More recent examples include the *DINA models* (de la Torre & Douglas, 2004), the *NIDA models* (Junker & Sijtsma, 2001), the *DINO models* (Templin & Henson, 2006), the *Fusion model* (e.g., Roussos, et al., 2007), and the *hierarchical general diagnostic model* (von Davier, 2007).

In 2004, Leighton, Gierl, and Hunka also introduced a procedure for CDA called the *attribute hierarchy method* (AHM). The AHM, a method that evolved from Tatsuoka's rule space model (see Gierl, 2007), is a psychometric procedure for classifying examinees' item responses into a set of structured attribute patterns associated with different components from a cognitive model of task performance. Attributes include different procedures, skills, and/or processes that an examinee must possess to solve an item. These attributes are structured using a hierarchy so the ordering of the cognitive skills is specified. As a result, the attribute hierarchy serves as an explicit construct-centered cognitive model. This model, in turn, provides a framework for designing test items and for linking examinees' test performance to specific inferences about psychological skill acquisition. AHM developments have been documented in the educational and psychological measurement literature, including psychometric advances (e.g., Leighton et al., 2004; Gierl, Leighton, & Hunka, 2007; Gierl, Cui, & Hunka, in press; Cui, Leighton, Gierl, & Hunka, 2006) and practical applications (e.g., Gierl, Wang, & Zhou, 2008; Wang & Gierl, 2007). The AHM has also been used to study differential item functioning (Gierl, Zheng, & Cui, 2008) and to service diagnostic adaptive testing (Gierl & Zhou, 2008).

To-date, however, the AHM has not been applied in an operational diagnostic testing situation because the reliability for attribute-based scoring must be established. Attribute reliability is a fundamental concept in CDA because score reports must provide users with a comprehensive yet

succinct summary of the outcomes from testing, including *score precision*. The authors of the Standards for Educational and Psychological Testing (1999) make this point clear when they state in Standard 5.10:

When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.

Addressing what the test covers, what scores mean, common misinterpretations, and how scores are used is, largely, a descriptive process that is specific to a particular testing situation. Conversely, addressing score precision is a more analytic process that generalizes across testing situations. Hence, the purpose of this paper is to introduce and evaluate an analytic procedure for assessing *attribute reliability*.

Overview of Attribute Hierarchy Method

The AHM (Leighton et al., 2004) is a psychometric method for classifying examinees' item responses into a set of structured attribute patterns associated with different components from a *cognitive model of task performance*. An attribute is a description of the procedural or declarative knowledge needed to perform a task in a specific domain. These attributes form a *hierarchy* that define the psychological ordering among the attributes required to solve a test item. The examinee must possess these attributes to answer items correctly. The attribute hierarchy serves as a cognitive model of task performance which, in educational measurement, refers to a simplified description of human problem solving on standardized tasks at some convenient grain size or level of detail in order to facilitate explanation and prediction of students' performance, including their strengths and weaknesses (Leighton & Gierl, 2007). These models provide an interpretative framework that can guide item development so test performance can be linked to specific inferences about examinees' cognitive skills. An AHM analysis is often conducted as a two-stage process, where the cognitive model is developed first, and then the examinee response data are classified using statistical pattern recognition techniques to produce attribute probability estimates.

Stage 1: Cognitive Model Development

The purpose of the first stage is to generate the expected examinee response patterns for a specific attribute hierarchy. A sample hierarchy is presented in Figure 1. This example is used in Leighton et al. (2004) and it will also be used as one attribute hierarchy in our simulation study. The hierarchy contains two divergent branches, but with a common prerequisite of attribute 1. In the first branch, attribute 2 is prerequisite to attribute 3. In the second branch, attribute 4 is prerequisite to attributes 5 and 6. A formal representation is used where the adjacency, reachability, incidence, reduced incidence, and expected response matrices are specified (cf. Tatsuoka, 1983, 1990, 1991, 1995).

A binary adjacency matrix (A) of order k by k , where k is the number of attributes, specifies the direct relationships among attributes. Then, a reachability matrix (R) of order k by k , where k is the number of attributes, specifies the direct and indirect relationships among attributes. The R matrix is calculated using $R = (A + I)^n$, where n is the integer required for R to reach invariance and can represent the numbers 1 through k , given A , the adjacency matrix, and I , an identity matrix. The incidence matrix (Q) of order k by p where k is the number of attributes and p is the number of potential items, is produced next. The set of potential items is considered a bank or pool of items that probes all combinations of attributes when the attributes are *dependent* and *independent*. The columns of the Q matrix are created by converting the integers ranging from 1 to 2^k-1 to their binary form.¹ This potential item pool is reduced when the attributes are related in a hierarchical structure because the hierarchy imposes *dependencies* among the attributes resulting in a reduced incidence matrix (Q_r). The Q_r matrix is produced by determining which columns of the R matrix are logically included in columns of the Q matrix, using Boolean inclusion. The Q_r matrix represents the cognitive specifications for the test, and it is used as a guide to develop and/or interpret items that measure specific attributes outlined in the hierarchy. The Q_r matrix is of order k by i where k is the number of attributes and i is the reduced number of items resulting from the constraints imposed by the hierarchy. The Q_r matrix for the attribute hierarchy in Figure 1 is given by

¹ The reader who is interested in reviewing the A , R , and Q matrices for the attribute hierarchy in Figure 1 can refer to Leighton et al. (2004), Matrices 1 to 3, respectively.

$$\begin{bmatrix} 11111111111111 \\ 011011011011011 \\ 001001001001001 \\ 00011111111111 \\ 000000111000111 \\ 000000000111111 \end{bmatrix} \cdot$$

Finally, the *expected response matrix* (E) is created, again using Boolean inclusion, where the algorithm compares each row of the expected attribute pattern matrix (which is the transpose of the Q_r matrix) to the columns of the Q_r matrix. The E matrix is of order j by i , where j is the number of examinees and i is the number of items. The rows of the E matrix are those responses that would be *logically* produced by an examinee who possesses the attributes as defined and structured in the attribute hierarchy and presented by the columns of the Q_r matrix. The columns of the E matrix are the items that probe specific attribute combinations. In short, this matrix provides a description of the expected examinee response patterns that would be produced if the cognitive model is true and if examinees respond without slips or errors. The expected response (E) matrix for the attribute hierarchy in Figure 1 is specified as

$$\begin{bmatrix} 10000000000000 \\ 11000000000000 \\ 11100000000000 \\ 10010000000000 \\ 11011000000000 \\ 11111000000000 \\ 10010010000000 \\ 11011011000000 \\ 11111111000000 \\ 10010000100000 \\ 11011000011000 \\ 11111100011100 \\ 10010010010010 \\ 110110110110110 \\ 11111111111111 \end{bmatrix} \cdot$$

Stage #2: Statistical Pattern Recognition

An examinee's observed response pattern is judged relative to the expected response patterns in the E matrix under the assumption that the cognitive model is true and that examinees respond without slips or errors. Hence, the purpose of the statistical pattern recognition stage is to identify the attribute

combinations that the examinee is likely to possess and to estimate the probability that an examinee possesses these specific attribute combinations. These probabilities provide examinees with specific information about their attribute-level mastery as part of the test reporting process.

To estimate the probability that examinees possess specific attributes, given their observed item response pattern, an artificial neural network approach is used (Gierl et al., in press; see also Gierl, Zheng, Cui, 2008, pp. 70-72). A neural network is a type of parallel-processing architecture that transforms a stimulus received by the input unit to a signal for the output unit through a series of mid-level hidden units. The input layer units produces a weighted linear combination of their inputs which are then transformed to non-linear weighted sums that are passed to every hidden layer unit. The hidden layer units, in turn, produce a weighted linear combination of their inputs which are transformed to non-linear weighted sums that are passed to every output layer unit. The network serves as a powerful pattern recognition technique because it can map any relationship between input and output. The input to train the neural network is the expected response vectors produced from the AHM analysis. For each expected response vector there is a specific combination of examinee attributes (i.e., the transpose of the Q_r matrix). The examinee attribute patterns, like the expected response vectors, are meaningful because they are derived from the cognitive model and provide a description of each attribute pattern that should be associated with each expected response pattern. The relationship between the expected response vectors with their associated attribute vectors is established by presenting each pattern to the network repeatedly until it learns the associations. The final result is a set of weight matrices, one for cells in the hidden layer and one for the cells in the output layer, that can be used to transform any response vector to its associate attribute vector.

The functional relationship for mapping the examinees' observed response pattern onto the expected response pattern so their attribute probabilities can be computed, is given as follows. Let

$$F(z) = \frac{1}{1 + e^{-z}},$$

and

$$a_k = \sum_{j=1}^q v_{kj} F\left(-\sum_{i=1}^p w_{ji} x_i\right),$$

then the output for unit k , M_k^* , is given as

$$M_k^* = F(a_k),$$

where q is the total number of hidden units, v_{jk} is the weight of hidden unit j for output unit k , p is the total number of input units, w_{ji} is the weight of input unit i for hidden unit j , and x_i is the input received from input unit i . The output values, scaled from 0 to 1, can be interpreted as probabilities. Using this approach, attribute probabilities can be computed for each observed response pattern thereby providing examinees with specific information about their attribute-level performance (see Gierl, Wang, & Zhou, 2008, pp. 36-41, for an example using SAT Mathematics).

Attribute Reliability

Attribute reliability refers to the precision of score decisions, typically on a diagnostic test, about examinees' attribute mastery. One method for estimating the reliability of an attribute is to calculate the ratio of true score variance to observed score variance *on the items that probe each attribute*. With the AHM, an item is designed to measure a combination of attributes as specified in the hierarchy. Consequently, for items that measure more than one attribute, each attribute only contributes to a part of the total item-level variance. Our index, therefore, incorporates the concept of *attribute dependency* into the reliability calculation using the concept of internal consistency. Consider attribute 1 in Figure 1. Attribute 1 is the prerequisite for attributes 2 to 6 because an examinee must possess attribute 1 in order to correctly respond to items measuring any other attribute in the hierarchy. Similarly, if an examinee correctly answers items that directly probe attribute 2, then we can infer that the examinee has also mastered attribute 1 because of their structural relationship in the hierarchy. More generally, we note that attribute 1 is measured directly or indirectly by all test items and, thus, to calculate the reliability of attribute 1, all test items must be included. Now consider attribute 3 in Figure 1. Attribute 3 does not serve as the prerequisite of any other attribute. If an examinee produces a correct answer to items that require attribute 4, for instance, then we could not discern if the examinee had mastered attribute 3

because attributes 3 and 4 are independent. Hence, only items that directly probe attribute 3 can be included in the reliability estimate for attribute 3. Attribute dependency also implies that prerequisite attributes in the initial nodes of the hierarchy, such as attribute 1, are expected to have higher reliability estimates compared to attributes in the final nodes of the hierarchy, such as attributes 3, 5 or 6 in Figure 1 because of the dependencies among the attributes that, in turn, affect the number of items that measure each attribute, either directly or indirectly.

In order to isolate the contribution of each attribute to an examinee's item-level performance, the item score is weighted by the difference of two conditional probabilities. The first probability is associated with attribute mastery (i.e., an examinee who has mastered the attribute can answer the item correctly) and the second probability is associated with attribute non-mastery (i.e., an examinee who has not mastered the attribute can answer the item correctly). The weighted scores for items that measure the attribute are used in the reliability calculation.

Let W_{ik} denote the weight for item i in the calculation of attribute k . A W_{ik} value of 1 indicates that performance on item i is completely determined by attribute k . Hence, the variance of the responses on item i should be used in the calculation of the reliability for attribute k . Conversely, if W_{ik} has a value of 0, indicating that the mastery of attribute k could not increase the probability of solving item i correctly, then item i should not be used to calculate the reliability of attribute k . W_{ik} can be calculated as

$$W_{ik} = p(X_i = 1 | A_k = 1) - p(X_i = 1 | A_k = 0),$$

where $p(X_i = 1 | A_k = 1)$ is the conditional probability that an examinee who has mastered attribute k can answer item i correctly, and $p(X_i = 1 | A_k = 0)$ is the conditional probability that an examinee who has not mastered attribute k can answer item i correctly.

The term $p(X_i = 1 | A_k = 1)$ is calculated as

$$p(X_i = 1 | A_k = 1) = \frac{p(A_k = 1, X_i = 1)}{p(A_k = 1)},$$

where $p(A_k = 1, X_i = 1)$ is the joint probability that an examinee has attribute k and correctly answers item i , and $p(A_k = 1)$ is the marginal probability that an examinee has attribute k . To obtain $p(A_k = 1, X_i = 1)$ and $p(A_k = 1)$, the attribute patterns, the expected response patterns, and the estimate of the population probabilities for each of the expected response patterns must be specified.

The term $p(X_i = 1|A_k = 0)$ should be 0 because examinees are not expected to answer item i correctly since they lack attribute k required by item i . However, in an actual testing situation, it is possible that examinees can still answer the item correctly by guessing or by applying partial knowledge to reach their solution, particularly when the multiple-choice item format is used. Therefore,

$p(X_i = 1|A_k = 0)$ can also be fixed at a specific value (e.g., 0.20) that reflects a “pseudo-guessing” parameter.

Once the W_{ik} s are specified, the weighted scores can be used to calculate attribute reliability by adapting Cronbach coefficient alpha for the AHM procedure. The derived formula is given by

$$\alpha_{AHM_k} = \frac{n_k}{n_k - 1} \left[1 - \frac{\sum_{i \in S_k} W_{ik}^2 \sigma_{X_i}^2}{\sigma^2_{\sum_{i \in S_k} W_{ik} X_i}} \right],$$

where α_{AHM_k} is the reliability for attribute k , n_k is the number of items that are probing attribute k in the Q_r (i.e., the number of elements in S_k), $\sigma_{X_i}^2$ is the variance of the observed scores on item i ,

$\sum_{i \in S_k} W_{ik}^2 \sigma_{X_i}^2$ is the sum of the weighted variance of the observed scores on the items that are measuring

attribute k , and $\sigma^2_{\sum_{i \in S_k} W_{ik} X_i}$ is the variance of the weighted observed total scores, given by

$$\sum_{i \in S_k} W_{ik}^2 \sigma_{X_i}^2$$

The Spearman-Brown formula can also be used to evaluate the effect of changes to test length on the attribute reliability coefficient. The attribute-based Spearman-Brown formula is specified as

$$\alpha_{AHM-SB_k} = \frac{n_k \alpha_{AHM_k}}{1 + (n_k - 1)\alpha_{AHM_k}},$$

where α_{AHM-SB_k} is the Spearman-Brown reliability of attribute k if n_k additional items sets that are parallel to items measuring attribute k are added to the test. This formula can be used to evaluate the effect of adding parallel items to the reduced-incidence matrix on the attribute reliability estimate.

Method and Results

A simulation and real data study were conducted to evaluate and illustrate our attribute reliability indices. Three factors expected to affect attribute reliability were manipulated in the simulation study: cognitive model (linear and divergent), test length (2, 4, and 6 items per attribute), and slip percentage (10, 15, 20, and 25%). The number of attributes for each cognitive model was fixed at six and the sample size for each analysis was based on 5000 simulated examinees. Attribute reliability was then calculated using a random sample of 5000 students for a five-attribute cognitive model in algebra using items from the March 2005 administration of the SAT. The simulation study method and results are presented first, followed by the real data study.

Simulation Study

A simulation study was conducted to evaluate attribute reliability by manipulating three factors that could directly affect CDA outcomes. The first factor was the type of cognitive model. Two models were evaluated, a linear and a divergent hierarchy. The number of attributes in each model was fixed at six. The first cognitive model is a simple linear hierarchy. This model contains all six attributes aligned in a single branch. This type of model could be used to characterize problem-solving when the knowledge and skills are ordered in a linear manner. The second cognitive model is a more complex divergent hierarchy, as presented in Figure 1. This model contains two independent branches which share a common prerequisite, attribute 1. The first branch includes two additional attributes, 2 and 3, while the second branch includes a self-contained sub-hierarchy with attributes 4 through 6. Two independent branches form the sub-hierarchy: Attributes 4, 5 and attributes 4, 6. This type of model could be used to characterize problem-solving when the knowledge and skills differ as a function of the concepts and content within a domain (see, for example, Gierl, Wang, & Zhou, 2008). Taken together, these two

models represent different types of cognitive structures that could characterize student performance on a diagnostic test.

The second factor was the number of items measuring each attribute at different nodes in the hierarchy. Three different items sets were evaluated—two, four, and six item sets—because length was expected to affect reliability. The two item set yields a diagnostic test with 12 items, as each of the six attributes was measured by two items. Similarly, the four item set produces a diagnostic test with 24 items and the six item set a test with 36 items. Because some CDAs are intended to promote formative assessment outcomes, often, in a classroom environment, testing will likely occur frequently during instruction. Hence, test length may be restricted so it does not detract from instructional time, given that testing will occur more often. To evaluate attribute reliability in this context, we included a short 12-item test (2 item set), a medium length 24-item test (4 item set), and a longer 36-item test (6 item set).

The third factor was the percentage of slips involved in the simulated responses. These slips represent the differences between the expected responses prescribed by the cognitive model in the E matrix and the actual responses produced by examinees. Four slip levels were evaluated to produce different percentages of model-data misfit—10%, 15%, 20%, and 25%, meaning the difference between the expected and actual responses ranged from 10 to 25%. Taken together, 144 conditions were assessed in the simulation study [i.e., (2) cognitive models * (6) attributes * (3) items sets * 4 (slip conditions) = 144).

To generate the observed response data, a two-step process was used. First, the matrices of the AHM for each hierarchy—adjacency, reachability, incidence, reduced incidence, and expected response matrix—were derived. The expected response matrix was used as the basis for generating the simulated examinee response data. For each hierarchy, a sample of 5000 expected item response vectors were generated with the constraint that the ability estimates associated with the expected response patterns be normally distributed. The ability estimates for the expected response patterns were produced using maximum likelihood estimation. Second, slips were added to the expected response data to simulate the observed response patterns. The randomly added slips ranged from 10 to 25%. These slips were based on item probabilities calculated from each expected response pattern using the 2-parameter logistic IRT

model, where the a -parameter was fixed to 1.0. The item parameters used in the simulation study for both linear and divergent models are presented in Table 1. Two types of slips were generated. First, slips were created for the subset of items expected to be answered incorrectly according to the attribute hierarchy (i.e., slips of the form 0 to 1). The percentage of these slips was specified as the item probability. Second, slips were created for the subset of items expected to be answered correctly according to the attribute hierarchy (i.e., slips of the form 1 to 0). The percentage of these slips was specified as one minus the item probability.

As an example, consider the divergent model in Figure 1 where two items are used to measure each attribute. 425 of the 5000 examinees were expected to have attribute pattern [100000] thereby producing the response pattern [110000000000]. The response probabilities for these items were computed using item parameters and the ability level associated with this attribute pattern. If we only consider the probabilities for items 1 and 3 in the expected response pattern, the values are 0.85 and 0.07, respectively. According to their expected response pattern, the 425 examinees are expected to answer the first item correctly. However, the probability of a correct response calculated from the 2-parameter logistic IRT model is 0.85 for item 1, indicating that although students have mastered the attributes required by the item, they still have $1 - 0.85 = 15\%$ chance of producing an incorrect response. Therefore, $425 \times 15\% \approx 64$ response vector associated with attribute pattern [100000] were randomly selected in the simulated data and changed from 1 to 0 for item 1. The 425 examinees are also expected to answer item 3 incorrectly. Using the item parameters and ability level estimate, the probability is 0.07 for item 3. This value suggests that although examinees are not expected to answer the item correctly based on their attribute mastery, they still have 7% chance of producing a correct response. As a result, $425 \times 7\% \approx 30$ slips of the form 0 to 1 were randomly introduced into the simulated data for item 3. Using this data generation procedure, four different slip percentages—10, 15, 20, and 25%—were introduced into the expected response data producing observed responses that differed from the expected responses in the form of either 1 to 0 or 0 to 1.

Simulation Results

The simulation study results are presented in Tables 2 and 3. Each table contains the reliability level for each of the six attributes as a function of the number of items measuring each attribute (2, 4, 6) and the percentage of slips (10%, 15%, 20%, 25%) for the 5000 simulated examinees.

Linear Cognitive Model. Table 2 contains the results for the linear cognitive model. With two items measuring each attribute, reliability was highest for attribute 1 and lowest for attribute 6 in all slip conditions. With 10% error, attribute reliability ranged from 0.84 for attribute 1 to 0.43 for attribute 6. With 15% error, reliability ranged from 0.81 for attribute 1 to 0.28 for attribute 6. With 20% error, reliability ranged from 0.78 for attribute 1 to 0.22 for attribute 6. With 25% error, reliability ranged from 0.74 for attribute 1 to 0.22 for attribute 6.

With four items per each attribute, reliability was highest for attribute 1 and lowest for attribute 6 in all slip conditions. However, test length and model-data fit clearly affected reliability as the outcomes in the four item condition were higher than the two item condition. With 10% error, reliability ranged from 0.92 for attribute 1 to 0.61 for attribute 6. With 15% error, reliability ranged from 0.90 for attribute 1 to 0.46 for attribute 6. With 20% error, reliability ranged from 0.88 for attribute 1 to 0.39 for attribute 6. With 25% error, reliability ranged from 0.86 for attribute 1 to 0.39 for attribute 6.

With six items per each attribute, reliability was highest for attribute 1 and lowest for attribute 6, as in the previous item-set conditions. Again, the importance of test length and model-data fit was apparent as the overall reliability outcomes in the six item condition were higher than the two or four item conditions. With 10% error, reliability ranged from 0.94 for attribute 1 to 0.70 for attribute 6. With 15% error, reliability ranged from 0.93 for attribute 1 to 0.56 for attribute 6. With 20% error, reliability ranged from 0.92 for attribute 1 to 0.49 for attribute 6. With 25% error, reliability ranged from 0.90 for attribute 1 to 0.49 for attribute 6.

Divergent Cognitive Model. Table 3 contains the results for the divergent cognitive model. Recall, the divergent model is more complex than the linear model because it contains two independent branches which share a common prerequisite, attribute 1. The first branch includes two attributes, 2 and 3, while the second branch forms two sub-hierarchies consisting of attributes 4 and 5 and 4 and 6. With two items measuring each attribute, reliability was highest for attribute 1 and lowest for attributes in the

final nodes of the hierarchy (i.e., attributes 3, 5, and 6) in all slip conditions. With 10% error, reliability for branch one (i.e., attributes 1, 2, and 3) ranged from 0.80 for attribute 1 to 0.68 for attribute 3. For branch two (i.e., attributes 4, 5, and 6), reliability was 0.76 for attribute 4 and 0.68 and 0.69 for attributes 5 and 6, respectively. With 15% error, reliability for branch one ranged from 0.80 for attribute 1 to 0.55 for attribute 3. For branch two, reliability was 0.71 for attribute 4 and 0.54 and 0.55 for attributes 5 and 6, respectively. With 20% error, reliability for branch one ranged from 0.80 for attribute 1 to 0.43 for attribute 3. For branch two, reliability was 0.67 for attribute 4 and 0.43 and 0.47 for attributes 5 and 6, respectively. With 25% error, reliability for branch one ranged from 0.80 for attribute 1 to 0.41 for attribute 3. For branch two, reliability was 0.65 for attribute 4 and 0.41 and 0.42 for attributes 5 and 6, respectively.

With four items measuring each attribute, reliability was highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy across the slip conditions. As with the linear model analysis, test length and model-data fit affected the reliability level as outcomes in the four item condition were higher than the two item condition. With 10% error, reliability for branch one ranged from 0.90 for attribute 1 to 0.81 for attribute 3. For branch two, reliability was 0.87 for attribute 4 and 0.82 for both attributes 5 and 6. With 15% error, reliability for branch one ranged from 0.89 for attribute 1 to 0.70 for attribute 3. For branch two, reliability was 0.84 for attribute 4 and 0.71 for attributes 5 and 6. With 20% error, reliability for branch one ranged from 0.89 for attribute 1 to 0.62 for attribute 3. For branch two, reliability was 0.82 for attribute 4 and 0.62 for attributes 5 and 6. With 25% error, reliability for branch one ranged from 0.89 for attribute 1 to 0.59 for attribute 3. For branch two, reliability was 0.80 for attribute 4 and 0.59 for attributes 5 and 6.

With six items per each attribute, reliability, again, was highest for attribute 1 and lowest for attributes in the final nodes. Overall reliability outcomes were also highest in the six item condition, when compared to the two or four item conditions. With 10% error, reliability ranged from 0.93 for attribute 1 to 0.87 for attribute 3. For branch two, reliability was 0.91 for attribute 4 and 0.87 for both attributes 5 and 6. With 15% error, reliability for branch one ranged from 0.93 for attribute 1 to 0.78 for attribute 3. For branch two, reliability was 0.89 for attribute 4 and 0.79 for attributes 5 and 6. With 20% error,

reliability for branch one ranged from 0.93 for attribute 1 to 0.71 for attribute 3. For branch two, reliability was 0.87 for attribute 4 and 0.71 for attributes 5 and 6. With 25% error, reliability for branch one ranged from 0.92 for attribute 1 to 0.69 for attribute 3. For branch two, reliability was 0.86 for attribute 4 and 0.69 for attributes 5 and 6.

To summarize, test length and slip percentage systematically affected attribute reliability for the linear and divergent models. In both hierarchies, mean reliability increased as the number of items increased across each slip condition. Mean reliability also decreased as the slip percentage increased across each item set condition (see Figures 3a and 3b). In addition, the mean reliability in each simulation condition were comparable for the linear and divergent models. However, the standard deviations were noticeably and consistently larger for the linear model across the simulation study conditions. This outcome can be accounted for by the attribute dependencies inherent in each hierarchy. For instance, in the linear hierarchy with two items per attribute, attribute 1 was directly or indirectly measured by 12 items whereas attribute 6 was directly measured by only two items because of the model structure. For the divergent hierarchy with two items per attribute, attribute 1 was directly or indirectly measured by 12 items whereas attributes 3, 5, and 6 were directly measured by only two items. From this example it becomes clear that the divergent model contains attributes with fewer dependencies and, thus, each attribute is affected by a smaller number of items, either directly or indirectly. As a result, there is less variation among the reliability estimates with the divergent model which yields smaller standard deviations across all study conditions when compared with the linear model.

Real Data Study

To illustrate an application of the AHM within the domain of mathematics, an attribute hierarchy was developed to account for examinee performance in high school algebra. This hierarchy is based on our review of the released items from the March 2005 administration of the SAT. The Mathematics section contains items in the content areas of Number and Operations; Algebra I, II, and Functions; Geometry; and Statistics, Probability, and Data Analysis. But, for our example, only a small subset of items in Algebra I and II were evaluated and used to develop the hierarchy.

Cognitive Model Representation for SAT Algebra. Cognitive models guide diagnostic inferences because they are specified at a small grain size and they magnify the cognitive processes that underlie performance. Ideally, a theory of task performance would direct the development of the cognitive model where it would, first, be created and validated and, then, items would be written to measure each skill in the model. This theoretical position could also be guided by different psychological positions, including the trait, developmental, information-processing, or sociocultural perspectives. But, in the absence of such a theory, retrofitting may be required where a task analysis of the existing items is conducted to generate the initial cognitive model. Because we had no theory of SAT algebra performance, we developed a hypothetical cognitive model that may help explain student performance in algebra. We developed our hypothetical model by solving the SAT algebra items and, in the process, identifying the required mathematical concepts, operations, procedures, and strategies. Then, we categorized these cognitive attributes so they could be ordered in a logical, hierarchical sequence to summarize problem-solving performance. The attribute hierarchy for algebra performance we produced, along with a brief description of the component processes underlying each attribute, is presented in Figure 2. The hierarchy in our example is hypothetical and relatively simple. More complex cognitive models could easily be created from the SAT algebra items by adding attributes and further developing the hierarchical structure (see Leighton et al., 2004, pp. 209-211). Moreover, additional research is needed to refine our hypothetical model. For instance, the model should be validated using examinee performance from think-aloud methods. However, to present a concise hypothetical example using data from an operational test that helps illustrate how attribute reliability can be estimated, our five-attribute algebra hierarchy was used.

Attribute 1 requires the examinee to understand the arithmetic operations implied by $+$, $-$, \times , $/$, $=$, absolute value, square, square root, exponent, $>$, $<$, \leq , \geq , and signed numbers. It also includes the skills required to correctly execute basic computations, such as addition, subtraction, multiplication, and division of whole numbers. The first branch has two attributes, 2 and 3. Attribute 2 includes the skills required to solve linear functions. Attribute 3 includes the skills needed to factor quadratic expressions as well as solve quadratic functions. By the nature of the hierarchical structure, attribute 3 requires the

skills in attributes 1 (basic arithmetic operations and computations) and 2 (solving linear functions). The second branch also has two attributes, 4 and 5. Attribute 4 includes the skills required for simple substitution problems (e.g., substitute the value of one variable for a letter). Attribute 5 includes the skills necessary for complex substitution problems (e.g., substitute numbers and letters into abstract expressions and rules). Again, by the nature of the hierarchical structure, attribute 5 requires the skills in attributes 1 (basic arithmetic operations and computations) and 4 (simple substitution). In our review of the SAT items, only one item was identified for attributes 1, 2, and 4 and two items were identified for attributes 3 and 5. Data from a random sample of 5000 students who wrote these items on the March 2005 administration of the SAT were analyzed.

Reliability Analyses of SAT Algebra Model. Reliability analyses were conducted using the SAT algebra hierarchy in Figure 2. Reliability ranged from 0.56 for attribute 1 to 0.25 for attribute 5, as shown in Table 4 (second column). As in the simulation study, the number of items affects the attribute reliability estimates. Attribute 5 was only measured directly by two items. Hence, the reliability estimate was relatively low at 0.25. Similarly, attribute 3 was measured directly by two items producing a relatively low reliability estimates of 0.41. Attributes 2 and 4 were measured directly by one item and indirectly by two items producing higher reliability estimates of 0.43 and 0.45, respectively. Attribute 1 was measured directly by one item and indirectly by six items resulting in the highest reliability estimate at 0.56.

The effect of test length, which is controlled by the number of parallel items measuring each attribute, can also be evaluated using the attribute-based Spearman-Brown formula.² If we use two parallel items sets (i.e., $7 \times 2 = 14$ items), then reliability increases, as expected (see third column in Table 4). The attribute reliability now ranges from 0.72 for attribute 1 to 0.40 for attribute 5. If we use four parallel

² The attribute-based Spearman-Brown formula can also be applied to the data in our simulation study to estimate the affect of adding parallel items to the reduced-incidence matrix. The formula results are predictive of the simulation study outcomes. For example, when the linear model is used with the four-item set in the 25% error condition, reliability is 0.86, 0.83, 0.77, 0.67, 0.53, and 0.39 for attributes 1 to 6, respectively. When the attribute-based Spearman-Brown formula is used with the linear model and two-item set to predict the outcomes for the four-item set in the 25% error condition, reliability is projected to be 0.85, 0.82, 0.76, 0.65, 0.48, and 0.36 for attributes 1 to 6, respectively.

items sets (i.e., $7 \times 4 = 28$ items), then reliability, again, increases ranging from 0.84 for attribute 1 to 0.57 for attribute 5 (see fourth column in Table 4). If we use six parallel items sets (i.e., $7 \times 6 = 42$ items), then reliability ranges from 0.88 for attribute 1 to 0.67 for attribute 5 (see fifth column in Table 4).

Summary and Discussion

The purpose of this study was to introduce, evaluate, and illustrate two indices for estimating attribute reliability. Attribute reliability refers to the precision of score decisions about examinees' attribute mastery level. Reliability of an attribute can be calculated using the ratio of true score variance to observed score variance on the items that probe specific attributes, meaning that Cronbach's alpha can be recast into the AHM framework to provide information about attribute consistency in the measurement process. It can also be used to evaluate the consequences of increasing test length using the attribute-based Spearman-Brown formula. Our concept of attribute reliability is an example of *construct-centered reliability*, as described by Nichols and Smith (1998), because it incorporates substantive explanations of test score meaning (i.e., those conditions outlined in the cognitive model) to interpret the consistency in the examinees' items response patterns when multiple measurements per attribute are obtained.

The results of our study have at least two practical implications. First, attribute reliability estimates can be used to enhance score reporting through the creation of confidence intervals around attribute-based scores. In the AHM, the statistical pattern recognition stage is used to estimate the probability that an examinee possesses specific attribute combinations. These attribute probabilities serve as scores that provide examinees with specific information about their cognitive strengths and weaknesses and to create score reporting profiles (Gierl et al., in press). The precision of these point-estimate scores can be further enhanced by creating a confidence interval around each probability using the standard error of measurement with the attribute reliability coefficient. The implications on adding parallel items to the reduced-incidence matrix can also be evaluated using the attribute-based Spearman-Brown formula.

Second, attribute reliability outcomes help highlight key measurement issues that must be addressed when comparing CDAs to other forms of assessment. For instance, one promising applications of CDA is

in the area of *formative, classroom-based assessment*. This type of assessment has many unique characteristics. It is implemented periodically during the teaching and learning process. The assessment outcomes are intended to guide teaching and learning hence the content on the test should be closely linked to the curriculum. Testing outcomes support decisions that are direct, specific, and immediate, such as deciding on a student's homework assignment or guiding the teacher's instructional planning. As a result, the assessment should be scored automatically so students and teachers can receive feedback immediately (Gierl & Zhou, 2008).

Yet, one significant challenge in formative, classroom-based assessment stems from establishing the reliability of diagnostic scores because of the granularity of the cognitive model, on the one hand, and the frequency of testing, on the other hand. The granularity of cognitive diagnostic models is typically fine grained because many attributes are required to characterize complex problem solving. To measure these attributes reliably, two or more items per attribute are needed, depending on the standard of score precision. At the same time, CDAs applied to formative classroom assessment are intended to align learning, instruction, and assessment. This alignment comes about, in part, through the frequent use of testing so assessment results can provide the timely feedback necessary to guide teaching and learning. The results from our study suggest that it may be difficult, or even impossible, to measure attributes reliably with small numbers of items. In our simulation study, test length had a predictable affect on attribute reliability, as longer tests produced higher reliabilities. For example, if we use a reliability estimate of 0.70 as our minimum standard of score precision, the 24- and 36-item linear model produced acceptable results for all attributes in the 10% slip condition, with one exception (attribute 6 in the 24-item condition). The divergence model was more robust as the 24- and 36-item conditions produced acceptable results for all attributes in the 10% and 15% slip conditions. The reliability results were even acceptable in most of the 36-item condition with 20% and 25% error. These findings indicate that researchers and practitioners must carefully consider their diagnostic model structure as well as its fit to the underlying data and test length when estimating attribute reliability. But our simulation and real data study results also reveal that short diagnostic tests (e.g., 12 item or less) will likely not yield adequate score precision to guide decisions about learning and instruction. This conclusion leads to a critical

question: If developers are willing to create the detailed probes necessary for evaluating specific cognitive skills, are users willing to allot more testing time to acquire this detailed information? The relatively large number of items required to measure fine-grained attributes reliably combined with the need for more frequent testing highlights one practical trade-off that both developers and users must recognize when applying CDAs in an area such as formative, classroom-based assessment.

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Cui, Y., Leighton, J. P., Gierl, M. J., & Hunka, S. (2006, April). *A person-fit statistic for the attribute hierarchy method: The hierarchy consistency index*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis, *Psychometrika*, *69*, 333-353.
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374.
- Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule space model and attribute hierarchy method. *Journal of Educational Measurement*, *44*, 325-340.
- Gierl, M. J., Leighton, J. P., & Hunka, S. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. (pp. 242-274). Cambridge, UK: Cambridge University Press.
- Gierl, M. J., Cui, Y., & Hunka, S. (in press). Using connectionist models to evaluate examinees' response patterns on tests. *Journal of Modern Applied Statistical Methods*.
- Gierl, M. J., Zheng, Y., & Cui, Y. (2008). Using the attribute hierarchy method to identify and interpret the cognitive skills that produce group differences. *Journal of Educational Measurement*, *45*, 65-89.
- Gierl, M. J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT®. *Journal of Technology, Learning, and Assessment*, *6* (6). Retrieved [date] from <http://www.jtla.org>.
- Gierl, M. J., & Zhou, J. (2008). Computer adaptive-attribute testing: A new approach to cognitive diagnostic assessment. *Zeitschrift für Psychologie—Journal of Psychology*, *216*, 29-39.

- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272.
- Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*, 3-16.
- Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement, 41*, 205-236.
- Nichols, P. D., & Smith, P. L. (1998). Contextualizing the interpretation of reliability data. *Educational Measurement: Issues and Practice, 17*, 24-36.
- Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practices. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* (pp. 307-353). Washington, DC: American Educational Research Association.
- Roussos, L., Dibello, L., Stout, W., Hartz, S., Henson, R., & Templin, J. (2007). The fusion model skills diagnostic system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment in education: Theory and applications* (pp. 275-318). Cambridge, UK: Cambridge University Press.
- 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. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Fredrickson, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.
- Tatsuoka, K. K. (1991). *Boolean algebra applied to the determination of the universal set of knowledge states*. (Research Report 91-2-ONR). Princeton, NJ: Educational Testing Service.
- Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.

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

von Davier, M. (2007). *Hierarchical general diagnostic models* (Research Report No. RR-07-19). Princeton, NJ: Educational Testing Service.

Wang, C., & Gierl, M. J. (2007, April). *Investigating the cognitive processes underlying student performance on the SAT Critical Reading subtest*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Table 1

Item Parameters Estimated from the Expected Response Matrix for the Linear and Divergent Models

Attribute	<i>Model</i>			
	<i>Linear</i>		<i>Divergent</i>	
	a	b	a	b
1	1.00	-2.50	1.00	-2.50
2	1.00	-1.50	1.00	0.00
3	1.00	-0.50	1.00	2.50
4	1.00	0.50	1.00	0.00
5	1.00	1.50	1.00	2.50
6	1.00	2.50	1.00	2.50

Table 2

Attribute Reliability Using Different Percentages of Slips in the Observed Response Patterns for a Linear Cognitive Model as a Function of Test Length

Item Set	Attribute	Slip Percentages			
		10%	15%	20%	25%
2	1	0.84	0.81	0.78	0.74
	2	0.82	0.78	0.74	0.69
	3	0.78	0.73	0.67	0.61
	4	0.71	0.64	0.56	0.48
	5	0.60	0.50	0.41	0.32
	6	0.43	0.28	0.22	0.22
	Mean	0.69	0.62	0.57	0.51
	SD	0.16	0.20	0.22	0.21
4	1	0.92	0.90	0.88	0.86
	2	0.91	0.88	0.86	0.83
	3	0.88	0.85	0.81	0.77
	4	0.84	0.80	0.74	0.67
	5	0.77	0.69	0.61	0.53
	6	0.61	0.46	0.39	0.39
	Mean	0.82	0.76	0.72	0.67
	SD	0.12	0.17	0.19	0.18
6	1	0.94	0.93	0.92	0.90
	2	0.94	0.92	0.90	0.88
	3	0.92	0.90	0.87	0.83
	4	0.89	0.86	0.81	0.76
	5	0.84	0.77	0.70	0.63
	6	0.70	0.56	0.49	0.49
	Mean	0.87	0.82	0.78	0.75
	SD	0.09	0.14	0.16	0.16

Table 3

Attribute Reliability Using Different Percentages of Slips in the Observed Response Patterns for a Divergent Cognitive Model as a Function of Test Length

Item Set	Attribute	<i>Slip Percentages</i>			
		10%	15%	20%	25%
2	1	0.80	0.80	0.80	0.80
	2	0.70	0.64	0.60	0.59
	3	0.68	0.55	0.43	0.41
	4	0.76	0.71	0.67	0.65
	5	0.68	0.54	0.43	0.41
	6	0.69	0.55	0.47	0.42
	Mean	0.72	0.63	0.57	0.55
SD	0.05	0.11	0.15	0.16	
4	1	0.90	0.89	0.89	0.89
	2	0.84	0.79	0.76	0.77
	3	0.81	0.70	0.62	0.59
	4	0.87	0.84	0.82	0.80
	5	0.82	0.71	0.62	0.59
	6	0.82	0.71	0.62	0.59
	Mean	0.84	0.78	0.72	0.70
SD	0.03	0.08	0.12	0.13	
6	1	0.93	0.93	0.93	0.92
	2	0.89	0.86	0.84	0.84
	3	0.87	0.78	0.71	0.69
	4	0.91	0.89	0.87	0.86
	5	0.87	0.79	0.71	0.69
	6	0.87	0.79	0.71	0.69
	Mean	0.89	0.84	0.79	0.78
SD	0.03	0.06	0.10	0.11	

Table 4

Attribute Reliability for SAT Example Across Four Different Test Lengths

Attribute	Test Length			
	7 Items	14 Items	28 Items	42 Items
1	0.56	0.72	0.84	0.88
2	0.43	0.60	0.75	0.82
3	0.41	0.58	0.74	0.81
4	0.45	0.62	0.77	0.83
5	0.25	0.40	0.57	0.67

Figure 1. The six-attribute divergent hierarchy.

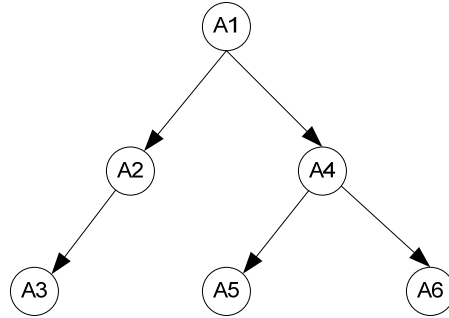
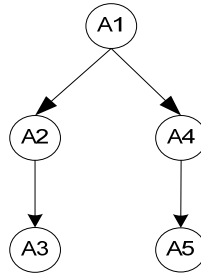
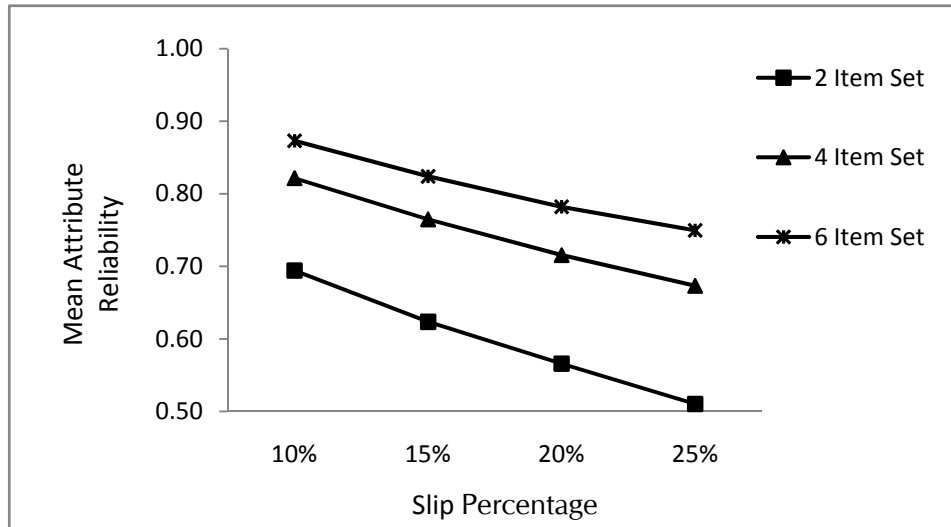


Figure 2. A five-attribute algebra hierarchy used to evaluate the cognitive skills required to solve seven items from the March 2005 administration of the SAT.

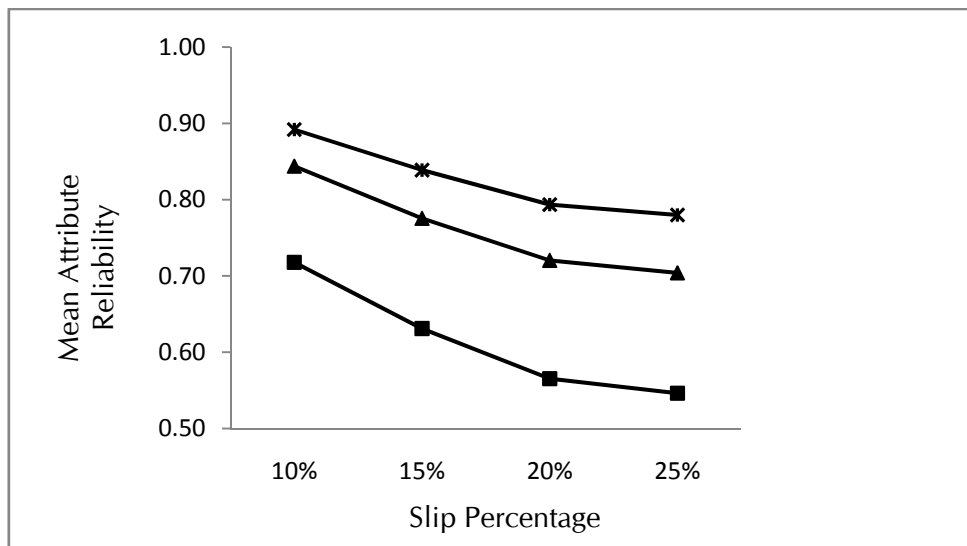


- A1: The understanding of the arithmetic operations implied by $+$, $-$, \times , $/$, $=$, absolute value, square, square root, exponent, $>$, $<$, \leq , \geq , and signed numbers. Also, the skills required to correctly execute basic computations, such as addition, subtraction, multiplication and division of whole numbers (pre-requisite attribute).
- A2: The skills required to solve linear functions.
- A3: The skills needed to factor quadratic expressions as well as solve quadratic functions.
- A4: The skills required for simple substitution problems (e.g., substitute the value of one variable for a letter).
- A5: The skills necessary for complex substitution problems (e.g., substitute numbers and letters into abstract expressions and rules).

Figure 3. Mean attribute reliability values two cognitive models as a function of test length and slip percentage.



a. Linear Hierarchy



b. Divergent Hierarchy