J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 6, Number 6 · February 2008

Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Cognitive Skills in Algebra on the SAT©

Mark J. Gierl, Changjiang Wang, & Jiawen Zhou

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

Volume 6, Number 6

Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Cognitive Skills in Algebra on the SAT©

Mark J. Gierl, Changjiang Wang, & Jiawen Zhou

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Jennifer Higgins
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

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.

J·T·L·A

Abstract:

The purpose of this study is to apply the attribute hierarchy method (AHM) to a sample of SAT algebra items administered in March 2005. The AHM is a psychometric method for classifying examinees' test 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. These attributes form a hierarchy of cognitive skills that represent a cognitive model of task performance. The study was conducted in two steps. In step 1, a cognitive model was developed by having content specialists, first, review the SAT algebra items, identify their salient attributes, and order the item-based attributes into a hierarchy. Then, the cognitive model was validated by having a sample of students think aloud as they solved each item. In step 2, psychometric analyses were conducted on the SAT algebra cognitive model by evaluating the model-data fit between the expected response patterns generated by the cognitive model and the observed response patterns produced from a random sample of 5000 examinees who wrote the items. Attribute probabilities were also computed for this random sample of examinees so diagnostic inferences about their attribute-level performances could be made. We conclude the study by describing key limitations, highlighting challenges inherent to the development and analysis of cognitive diagnostic assessments, and proposing directions for future research.

J·T·L·A

Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Cognitive Skills in Algebra on the SAT©

Mark J. Gierl
Changjiang Wang
Jiawen Zhou
Centre for Research in Applied Measurement and Evaluation, University of Alberta

Introduction

The purpose of this study is to apply the attribute hierarchy method (AHM) (Gierl, Leighton, & Hunka, 2007; Gierl, Cui, & Hunka, 2007; Leighton, Gierl, & Hunka, 2004) to a sample of algebra items from the March 2005 administration of the SAT and to illustrate how the method can promote diagnostic inferences about examinees' cognitive skills. The AHM is a psychometric method for classifying examinees' test 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. The examinee must possess these attributes to answer items correctly. The attributes form a hierarchy of cognitive skills defining the psychological ordering among the attributes required to solve test items. This attribute hierarchy represents a cognitive model of task performance. The hierarchy plays a *foundational role* in the AHM because it directs test development and guides the psychometric analyses so test scores have diagnostic value.

Our paper is divided into four sections. In the first section, we define the phrase *cognitive model* in educational measurement and we explain why these models are essential in the development and analysis of cognitive diagnostic assessments. In the second section we present the AHM. We describe a two-stage approach for diagnostic testing with the AHM where we, first, define the cognitive model of task performance and, second, evaluate the psychometric properties of the model. In the third section, we apply the AHM to a sample of algebra items from the March 2005 administration of the SAT. In the fourth section, we provide a summary, highlight some limitations of the current study, and identify areas where additional research is required.

Section I — Cognitive Models and Educational Measurement

A cognitive diagnostic assessment (CDA) is designed to measure an examinee's knowledge structures and processing skills (i.e., the examinee's cognitive skills). The knowledge structure contains factual and procedural information whereas the processing skills include the transformations and strategies required to manipulate this information (Lohman, 2000). These skills are important to measure on CDAs because they permit us to identify the examinees' cognitive strengths and weaknesses and, thus, make diagnostic inferences about their problem-solving skills. Unfortunately, these types of cognitively-based inferences are difficult, if not impossible, to produce without an explicit interpretative framework because the inferences are at a fine grain size (i.e., specific cognitive skills) rather than a coarse grain size (e.g., a total test score). Cognitive models serve this purpose as they provide the framework necessary to link cognitively-based inferences with specific, fine-grained test score interpretations (Gierl & Leighton, 2007; Leighton & Gierl, 2007a, 2007b). 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, 2007a, p. 6). Cognitive models are generated by studying the knowledge, processes, and strategies used by examinees as they respond to items. Many data sources and analytic procedures can lend themselves to the study of thinking and problem solving, including judgmental and logical analyses, generalizability studies, analyses of group differences, correlational and covariance analyses, and experimental interventions (Messick, 1989). However, verbal report methods are particularly well-suited to the study of *human information processing*. Hence, cognitive models are often created and validated by having examinees think aloud as they solve items to identify the information requirements and processing skills elicited by the tasks (Ericsson & Simon, 1993; Leighton, 2004; Leighton & Gierl, 2007b; Royer, Cisero, & Carlo, 1993; Taylor & Dionne, 2000). The model is then evaluated by comparing its fit to the examinees' observed response data and, sometimes, by comparing model-data fit across competing models. After evaluation and revision, the credibility of the model is established and it may even be generalized to other groups of examinees and to different types of tasks.

A cognitive model is specified at a small grain size because it is designed to magnify and accentuate the specific knowledge structures and processing skills that underlie test performance. With the AHM, the cognitive model also specifies a hierarchy of cognitive processes because attributes are assumed to share dependencies and function within a much larger network of inter-related processes, competencies, and skills (Anderson, 1996; Dawson, 1998; Fodor, 1983; Kuhn, 2001; Mislevy, Steinberg, & Almond, 2003). This key assumption about attribute dependency is significant for test development because the items that measure the attributes must maintain the cognitive structure outlined in the hierarchy and must directly measure specific cognitive processes of increasing complexity. In other words, the items in a cognitive diagnostic assessment must be designed systematically using this hierarchical order, if test performance is to be linked to information about examinees' cognitive skills.

The potential benefits of developing test items and interpreting test scores with reference to a cognitive model are numerous. For example, the development and use of a model provides one approach for identifying and measuring complex cognition so these knowledge structures and processing skills can be connected with test performance and test score interpretations. These outcomes are viable because the model provides a detailed framework for understanding how the examinees' cognitive skills can produce their observed response patterns and subsequent test scores. This type of understanding also permits the developer to provide detailed feedback to the examinees about their problem-solving strengths and weaknesses, given their observed response patterns. Once this model is validated, items that measure specific components of the model can be replicated thereby providing developers with a way of controlling the specific cognitive attributes measured by the test across administrations. But possibly most beneficial is the potential these models hold for linking theories of cognition and learning with instruction. Instructional decisions are based on how students think about and solve problems. Cognitive models provide one method for representing and reporting the examinees' cognitive profile on diverse tasks which could be used to link their weaknesses with instructional methods designed to improve the examinees' skills (National Research Council, 2001; Pellegrino, 2002; Pellegrino, Baxter, Glaser, 1999).

CDAs can also be distinguished from classroom assessments and large-scale tests, both conceptually and empirically. The development of a CDA is guided, at least initially, by educational and psychological studies on reasoning, problem solving, and information processing within a domain so an explicit cognitive model can be identified. This model, in turn, would be evaluated empirically using diverse data sources including examinee response data gleaned from a protocol analysis where the knowledge structures and process skills required by the examinees to perform competently in a specific domain are studied. Classroom assessments, by comparison, are neither developed nor interpreted with the aid of cognitive models.

Instead, classroom assessments tend to focus on content and curriculum-based outcomes so examinees can receive information on their progress within a program of studies on specified learning outcomes in a timely manner. The focus, therefore, is on behavioural outcomes and products (e.g., a total test score) rather than the underlying cognitive skills, their organization, and the processes that lead to different test scores (Leighton & Gierl, 2007a). Similarly, most large-scale tests are not developed from an explicit cognitive model. Instead, these tests are created from specifications or blueprints designed to sample broadly from different content and skill areas. Moreover, the skills measured on a large-scale test represent *cognitive intentions*, as the specifications outline the knowledge and skills the developer expects the examinees to use as they solve items. Hence, these skills serve as hypotheses about how one group (e.g., test developers) believes another group (e.g., students) will think, process information, and solve problems. These hypotheses are rarely, if ever, evaluated empirically. Thus, CDAs can be differentiated from classroom assessments and large-scale tests because they are grounded in a cognitive model that is scrutinized and, eventually, verified through empirical study.

Section II — Incorporating Cognitive Models Into Psychometric Analyses: Overview of Attribute Hierarchy Method

The AHM (Leighton, Gierl, & Hunka, 2004; Gierl, Cui, & Hunka, 2007) is a psychometric method for classifying examinees' test 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 defines the psychological ordering among the attributes required to solve a test item. The attribute hierarchy, therefore, serves as a cognitive model of task performance. These models provide an interpretative framework that can guide item development and psychometric analyses so test performance can be linked to specific cognitive inferences about examinees' knowledge and skills.

Stage 1: Cognitive Model Representation

An AHM analysis proceeds in two-stages. In stage 1, the expected examinee response patterns for a specific hierarchy in the AHM are computed. To specify the relationships among the attributes in the hierarchy using the AHM, the adjacency and reachability matrices are defined. The direct relationship among attributes is specified by a binary *adjacency matrix* (A) of order *(k,k)*, where *k* is the number of attributes. The adjacency matrix is of upper triangular form. The direct and indirect relationships among attributes are specified by the binary *reachability matrix* (R) of order *(k,k)*, where *k* is the number of attributes. To obtain the R matrix from the A matrix, Boolean addition and multiplication operations are performed on the adjacency matrix, meaning $R = (A + I)^n$, where *n* is the integer required to reach invariance, $n = 1, 2, \ldots m$, and *I* is the identity matrix.

The potential pool of items is generated next. This pool is considered to be those items representing all combinations of attributes when the attributes are independent of one other. The size of the potential pool is $2^k - 1$, where *k* is the number of attributes. The attributes in the potential pool of items are described by the *incidence matrix* (Q) of order (*k, p*), where *k* is the number of attributes and *p* is the number of potential items. This matrix can be reduced to form the *reduced incidence matrix* (Q_r) by imposing the constraints of the hierarchy as specified in the R matrix. The Q_r matrix is formed using Boolean inclusion by determining which columns of the R matrix are logically included in each column of the Q matrix. The Q_r matrix is of order *(k, i)* where *k* is the number of attributes and *i* is the reduced number of items resulting from the constraints in the hierarchy. The Q_r matrix also has an important interpretation from a cognitive test design perspective. It represents the cognitive specifications or blueprint for the test because it describes all attribute-by-item combination in the hierarchy. Thus, to operationalize and systematically evaluate each component in the cognitive model, items must be developed to measure each attribute combination in the hierarchy, as outlined in the Q_r matrix.

Given a hierarchy of attributes, the expected response patterns for a group of examinees can then be generated. The *expected response matrix* (E) is created using Boolean inclusion where each row of the attribute pattern matrix, which is the transpose of the Q_r matrix, is compared to the columns of the Q_r matrix. The expected response matrix is of order *(j, i)*, where *j* is the number of examinees and *i* is the reduced number of items resulting from the constraints imposed by the hierarchy. Examples of the adjacency, reachability, reduced incidence, and expected response matrices, as they apply to different types of cognitive models, can be found in Gierl, Leighton, and Hunka (2000, 2007) and Leighton et al. (2004). The adjacency, reduced incidence, and expected response matrices, as they apply to SAT problem solving in algebra, are illustrated later in this study.

Stage #2: Psychometric Analyses of the Cognitive Model

In stage 2, the psychometric analyses of the cognitive model are conducted. The observed response patterns can be classified using item response theory (IRT) based procedures reported in Leighton et al., (2004) or using non-IRT procedures described in Gierl, Cui, & Hunka (2007). In the current study, we use the non-IRT procedures to evaluate model-data fit and to compute attribute probabilities.

Model-Data Fit using the Hierarchy Consistency Index

Response discrepancies can occur when the expected response patterns produced in the E matrix are compared to the observed response patterns for a large sample of examinees. A model-data fit study is conducted to evaluate the consistency between the expected and observed response patterns. The Hierarchy Consistency Index (HCI_j) can be used to evaluate this fit (Cui, Leighton, Gierl, & Hunka, 2006). The HCI_j evaluates the degree to which observed examinee response patterns generated from a large group of examinees is consistent with the expected response patterns generated from the attribute hierarchy. Given K attributes and I items, the element q_{ki} of the Q_r matrix indicates if attribute k is required to solve the i^{th} item. It can be expressed as

$$q_{ki} = \begin{cases} 1 & \text{attribute k required by item i} \\ 0 & \text{otherwise} \end{cases}$$

Attribute mastery occurs when examinees correctly answer the items requiring the attribute. Thus, the HCI for examinee j is specified as

$$HCI_j = 1 - \frac{2 \sum_{i \in S_{correctj}} \sum_{g \in S_i} X_{j_i} (1 - X_{j_g})}{N_{c_j}}$$

where X_{j_i} is examinee j's score (0 or 1) to item i, S includes only those items that have attributes that are logically included in the attributes of item i, and N_{c_j} is the total number of comparisons for correctly answered items by examinee j. If examinee j correctly answers item i, $X_{j_i} = 1$, then the examinee is also expected to answer item g that belongs to S correctly, $X_{j_i} = 1$ $(g \in S_i)$ where S_i is the subset of items that examinee j answered correctly. However, if $X_{j_g} = 0$, then X_{j_i} $(1 - X_{j_g}) = 1$, which is considered a misfit of the response vector j relative to hierarchy. Thus, the numerator contains the number of misfits multiplied by 2. When the examinee's observed response does not match the hierarchy, the numerator is $(2 \times N_{c_j})$ and the HCI_j will have a value of –1. When the examinee's observed response pattern matches the hierarchy, the numerator is 0 and the HCI_j will have a value of 1. Hence, the HCI_j produces an index ranging from a perfect misfit of –1 to a perfect fit of 1. Recently, Cui (2007) demonstrated that HCI_j values above 0.70 indicate good model-data fit.

Attribute Probability

Once we establish that the model fits the data, attribute probabilities for each examinee can be calculated. These probabilities are critical for diagnostic inferences because they provide examinees with specific information about their attribute-level performance. To estimate these probabilities, an artificial neural network is used. The input to train the neural network is the *expected response vectors*. The expected response vector is derived from the attribute hierarchy and serves as the examinees' expected response patterns. These vectors are called exemplars. For each expected response vector there is a specific combination of examinee attributes. The examinee attribute patterns are meaningful because they are derived from the attribute hierarchy. The association between the expected response vectors and the attribute vectors is established by presenting each pattern to the network repeatedly until it learns the association. For instance, if 11001001 is an expected response pattern and 1000 is the attribute pattern, then the network is trained to associate 11001001 with 1000. The final result is a set of weight matrices, one for the 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 associated attribute vector. The transformed result is scaled from 0 to 1, where a higher value indicates that the examinee has a higher probability of possessing a specific attribute.

More specifically, the hidden layer produces a weighted linear combination of their inputs which are then transformed to non-linear weighted sums that are passed to every output unit. The contribution of each input unit i to hidden unit j is determined by connection weight, w_{ji}. The input layer contains the exemplars. The connection weights in the hidden layer transform the input stimuli into a weighted sum defined as

$$S_j = \sum_{i=1}^{p} w_{ji} x_i$$

where S_j is the weighted sum for node j in the hidden layer, w_{ji} is the weight used by node j for input x_i, and x_i is the input from node i of the input layer. For these variables, i ranges from 1 to p for the input node and j ranges from 1 to q for the hidden layer node. The network is designed to learn the value of the weights, w_{ji}, so the exemplars from the input layer are associated (i.e., error is a minimum) with their responses in the output layer. S is then transformed by the logistic function

$$S_j^* = \frac{1}{1 + e^{-S_j}} .$$

Using a similar approach, the hidden layer produces a weighted linear combination of their inputs which are transformed to non-linear weighted sums that are passed to every output layer unit to produce the final results. The effect of each hidden unit j to output unit k is determined by weight, v_{kj}. The output (S_j^*) from every hidden layer unit is passed to every output layer unit where, again, a linearly weighted sum (T_k) is formed using the weights v_{kj}, and the result transformed for output (T_k^*) using a nonlinear function. That is,

$$T_k = \sum_{j=1}^{q} v_{kj} S_j^*$$

where T_k is the weighted sum for each of k output nodes using weights v_{kj}, where j ranges from 1 to q for the hidden layer nodes. T_k is transformed by the logistic function using

$$T_k^* = \frac{1}{1 + e^{-T_k}}$$

resulting in output values that range from 0 to 1.

The quality of the network solution is evaluated by comparing the output targets in the response units (i.e., the examinee attributes) to the pattern associated with each exemplar (i.e., the expected response patterns). The initial solution is likely to be discrepant resulting in a relatively large error. However, the network uses this discrepant result to modify, by iteration, the connection weights leading to a smaller error term. One common learning algorithm used to approximate the weights so the error term is minimized is the generalized delta rule incorporated in a training procedure called *back propagation of error* (Rumelhart, Hinton, & Williams, 1986a, 1986b).

Consider a network with a single node response output that produces classification T^* when, in fact, the correct classification is Y. The squared error for this stimulus input can be given as

$$E = \frac{1}{2} (Y - T^*)^2 .$$

The goal is to adjust the T^* weights, v_{kj}, so the response units in the output layer have a small squared error term. Finding the weights v to minimize the error term E is done by setting the derivative of E with respect to v to 0, and solving for v; that is,

$$\frac{d(E)}{d(v)}$$

is required. Because the v weights in T^* depend on the w weights in S (recall, the neural network transforms any stimulus received by the input unit to a signal for the output unit through a series of mid-level hidden units), the chain rule in calculus must be used to solve

$$\frac{d(E)}{d(v)} = \left(\frac{d(E)}{d(T^*)}\right)\left(\frac{d(T^*)}{d(T)}\right)\left(\frac{d(T)}{d(v)}\right).$$

This equation can be simplified and re-cast into matrix algebraic terms as

$$\frac{d(E)}{d(v)} = -(Y - T^*)\ T^*\ (1 - T^*)\ S_j^*$$

where Y is the correct classification for the stimulus input, T^* holds the responses from the output layer, and S_j^* hold the responses for the hidden layer. The adjustment to weights v for each iteration in the output layer are then specified as

$$v(t+1) = v(t) + C(Y - T^*)\ T^*\ (1 - T^*)\ S_j^*$$

where C is the learning rate parameter. Because C adjusts the learning rate, changes to C affect the accuracy of the solution and can be used to manipulate how the network learns to classify patterns at different rates and with different degrees of accuracy. Typically, C is set to a small number. The learning algorithm, as it is applied to the AHM, is illustrated in the next section.

Section III — Applying AHM to Sample SAT Algebra Items

To illustrate an application of the AHM within the domain of mathematics, we developed a cognitive model to account for examinee performance in algebra. The SAT Mathematics section contains items in the content areas of Number and Operations; Algebra I, II, and Functions; Geometry; and Statistics, Probability, and Data Analysis. For the current study, the hierarchy we developed is based on our review of the released algebra items from the March 2005 administration of the SAT and from a validation study of these items using student verbal report data.

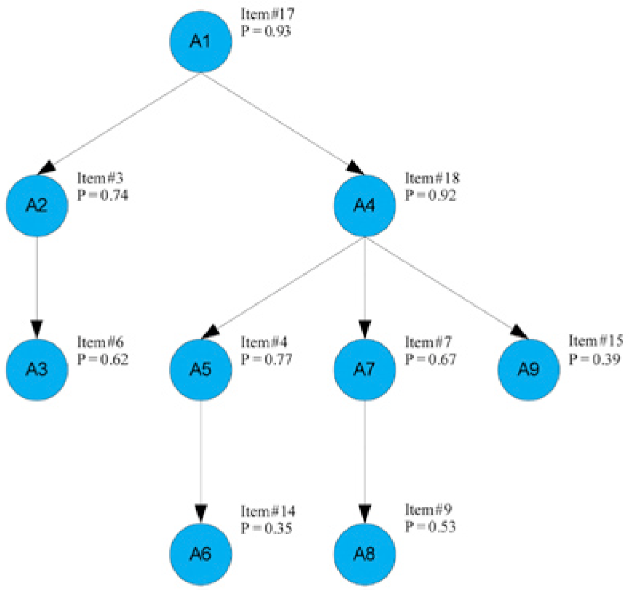
Stage #1: Cognitive Model Representation for SAT Algebra

Developing a Cognitive Model for Algebra Performance

In the previous section we noted that cognitive models guide diagnostic inferences because they are specified at a small grain size and they magnify the cognitive processes that underlie performance. One starting point is to have content specialists conduct a task analysis on a sample of items to specify the cognitive skills assumed to underlie performance, and to order these skills to create a hierarchy. This model can then be validated by presenting examinees in the target population of interest with the tasks and, using verbal report methods, recording the knowledge, processes, and strategies used by these examinees to solve the task.

In conducting the task analysis of the SAT algebra items we (the authors of this paper), first, solved each test item and attempted to identify the mathematical concepts, operations, procedures, and strategies used to solve each item. We then categorized these cognitive attributes so they could be ordered in a logical, hierarchical sequence to summarize problem-solving performance. A cognitive model of algebra performance is presented in Figure 1 (next page). The attributes are labeled A1 to A9. The test items are labeled at the right side of each attribute along with their difficulty level calculated from a random sample of 5000 students who wrote these items on the March 2005 administration. Because the attributes we identified are associated with existing test items, our cognitive model can be described as an item-based hierarchy. This type of hierarchy uses the test item as the unit of analysis. An item-based hierarchy can be compared to an attribute-based hierarchy where the attribute is the unit of analysis. Item-based hierarchies are typically generated when cognitive models are "retrofit" to existing items.

Figure 1: **A Cognitive Hierarchy in Ratios and Algebra**



The hierarchy in our example is relatively simple. More complex cognitive models could easily be created from the SAT algebra items by adding attributes and further developing the hierarchical structure. In addition, only a sample of algebra items from the SAT are used in this example. However, to present a concise example with data from an operational test to illustrate the AHM, our 9-attribute, 9-item hierarchy is presented. (Three other cognitive models for algebra were developed, but are not presented in this paper due to space constraints. The reader is referred to Gierl, Wang, and Zhou, 2007, for details.)

The hierarchy in Figure 1 presents a cognitive model of task performance for the knowledge and skills in the areas of ratio, factoring, function, and substitution (herein called the Ratios and Algebra hierarchy). The hierarchy contains two independent branches which share a common prerequisite – attribute A1. Aside from attribute A1, the first branch includes two additional attributes, A2 and A3, and the second branch includes a self-contained sub-hierarchy which includes attributes A4 through A9. Three independent branches compose the sub-hierarchy: attributes A4, A5, A6; attributes A4, A7, A8; and attributes A4, A9.

As a prerequisite attribute, attribute A1 includes the most *basic arithmetic operation skills*, such as addition, subtraction, multiplication, and division of numbers. For instance, in item 17, examinees are presented with the algebraic expression $4(t + u) + 3 = 19$, and asked to solve for $(t + u)$. For this item, examinees need to subtract 3 from 19 and then divide

16 by 4. Also note that item 17 is only deemed to be a sample item that represents a much broader range of basic prerequisite cognitive skills required by examinees to solve the items in this hierarchy. To evaluate the complete set of prerequisite skills, many more items would be required thereby adding new and more specific attributes along with a new hierarchical structure within the skills associated with attribute A1. In other words, the term "attribute" or "skill" could continually be specified at a smaller grain size and additional attributes or skills could be included between those attributes that are already identified thereby increasing the specificity of the cognitive inferences but also increasing the number of attributes and complexity of the hierarchical structure.

Attributes A2 and A3 both deal with factors. In attribute A2, the examinee needs to have *knowledge about the property of factors*. For example, in item 3, examinees are asked, *If* $(p + 1)(t - 3) = 0$ and p *is positive, what is the value of* t? The examinee must know the property that the value of at least one factor must be zero if the product of multiple factors is zero. Once this property is recognized, the examinee would be able to recognize that because p is positive, $(t - 3)$ must be zero to make the value of the whole expression zero, which would finally yield the value of 3 for t. In attribute A3, the examinee not only requires knowledge of factoring (i.e., attribute A2), but also the *skills of applying the rules of factoring*. Therefore, attribute A3 is considered a more advanced attribute than A2. For example, item 6 states,

$$\text{If } \frac{x+y}{a-b} = \frac{2}{3}\text{, then } \frac{9x+9y}{10a-10b} = ?$$

Only after the examinee factors the second expression into the product of the first expression

$$\left(\frac{x+y}{a-b} = \frac{2}{3}\right) \text{ and } \frac{9}{10}$$

would the calculation of the value of the second expression be apparent.

The self-contained sub-hierarchy contains six attributes. Among these attributes, attribute A4 is the prerequisite for all other attributes in the sub-hierarchy. Attribute A4 has attribute A1 as a prerequisite because A4 not only represents basic skills in arithmetic operations (i.e., attribute A1), but it also involves the *substitution of values into algebraic expressions* which is more abstract and, therefore, more difficult than attribute A1. For instance, in item 18, the examinee needs to substitute the values of variables into an equation (i.e., $w = 4$ and $x = 1$) to compute the value of k. Then, the examinee must substitute the values of k and w into $m = (w - 1)k$ to get the value of m.

The first branch in the sub-hierarchy deals, mainly, with functional graph reading. For attribute A5, the examinee must be able to *map the graph of a familiar function with its corresponding function*. In an item that requires attribute A5 (e.g., item 4), attribute A4 is typically required because the examinee must find random points in the graph and substitute the points into the equation of the function to find a match between the graph and the function. Attribute A6, on the other hand, deals with the *abstract properties of functions*, such as recognizing the graphical representation of the relationship between independent and dependent variables. The graphs for less familiar functions, such as a function of higher-power polynomials, may be involved. Therefore, attribute A6 is considered to be more difficult than attribute A5 and placed below attribute A5 in the sub-hierarchy. Item 14 provides an example: The examinee is required to understand the graph for a higher-power polynomial. The examinee must also recognize the equivalent relationship between $f(x)$ and y, and that the number of times the graph crosses with the line $y = 2$ is the number of values of x that make $f(x) = 2$.

The second branch in the sub-hierarchy considers the skills associated with advanced substitution. Attribute A7 requires the examinee to *substitute numbers into algebraic expressions*. The complexity of attribute A7 relative to attribute A4 lies in the concurrent management of multiple pairs of numbers and multiple equations. For example, in item 7, examinees are asked to identify which equation matches the pairs of x and y values. To solve this item, the examinee needs to substitute three pairs of x and y values into the five equations provided to find the correct pair. Attribute A8 also represents the *skills of advanced substitution*. However, what makes attribute A8 more difficult than attribute A7 is that algebraic expressions, rather than numbers, need to be substituted into another algebraic expression. For instance, in item 9, the examinee is given $x = 3v$, $v = 4t$, $x = pt$, and then asked to find the value of p. Examinees need to substitute x and v into the equation, set up an equation as $x = 3v = 12t = pt$, and then substituting a numeric value for t (such as 1) and for v (such as 4) which leads to the result that $p = 12$.

The last branch in the sub-hierarchy contains only one additional attribute, A9, related to *skills associated with rule understanding and application*. It is the rule, rather than the numeric value or the algebraic expression, that needs to be substituted in the item to reach a solution. In item 15, for example, examinees are presented with $x\Delta y = x^2 + xy + y^2$, and then asked to find the value of $(3\Delta 1)\Delta 1$. To solve this item, the examinee must first understand what the rule Δ represents, and then substitute the rule into the expression, $(3\Delta 1)\Delta 1$, twice, to produce the solution.

Protocol Analysis

Methods, Sample, and Coding Procedures

To validate the cognitive model in the Ratios and Algebra hierarchy, a protocol analysis was conducted (see Gierl, Leighton, Wang, Zhou, Gokiert, & Tan, 2007). The sample algebra items were administered in November 2005 to 21 high school students from New York City. Each volunteer was individually assessed in an empty conference room at the College Board main office. Students were asked to think aloud as they solved the items. After students reported an answer for the algebra item, they were then asked, "How did you figure out the answer to the problem?" unless the student volunteered the information. Each session was audiotaped and lasted, on average, 45 minutes.

The sample was drawn from all potential New York City students who took the PSAT as 10th graders, with the following six constraints: (1) the assessment was administered without special testing accommodations; (2) students live and attend school in New York City; (3) students scored between 55–65 on Math; (4) students scored between 60–80 on Critical Reading; (5) students opted-in to the Student Search Service; and (6) students had only taken the PSAT once. We intentionally selected students with above average PSAT Critical Reading scores, as these students were expected to have stronger verbal skills and, thus, be more proficient at verbalizing their thinking processes. At the same time, we attempted to select students with average to above average math skills so a range of mathematical proficiencies would be included in the think aloud sample. These selection decisions may limit the generalizability of our results, but it did help ensure that the verbal reports were clearly articulated and, thus, easier to code. A statistical analyst at the College Board sampled from this population producing a list of 75 male and 75 female students who were eligible to participate. All 150 students were contacted by mail. Of the 150 students contacted, 26 agreed to participate (17.3% of total sample); of the 26 who agreed, 21 (12 males; 9 females) students attended their scheduled testing session at the College Board main office. Sixteen of the 21 students were White; one student was an Asian/Pacific Islander; one student was Black/African American; and one student was Hispanic. Two students did not respond to the ethnicity self-report item. Each student who participated received $50 and a public transportation voucher for travel to and from the College Board.

Next, flow charts were created to represent students' cognitive processes as reported in the think aloud protocols. These flowcharts were used to evaluate both the item attributes and their hierarchical ordering. The cognitive flow charts were created and coded in three steps. In the first step, two graduate assistants on this project (Wang and Zhou) listened to each audiotape and created a flow chart for each student protocol. The elementary cognitive processes reported by students for each item were graphed using both the students' verbal responses and their written responses. Flow charts were used because they provided a systematic method for representing problem solving where both the components (i.e., elementary cognitive processes) and the overall structure of the components could be graphically presented. Flow charts also highlighted individual differences where the elementary steps and solution strategies for each student could be compared and contrasted. Standard flow chart symbols, as found in cognitive and computer science, were followed. The flow charts contained four different symbols:

1. *Start/Stop Box* – this is a parabola that starts and stops the flow chart. In this study students began by reading the questions out loud. Therefore the start box represents this point in the problem-solving sequence. The protocol was complete when students reported their answer. Thus the stop box contained the students' final answer. Only the solution path used to reach the final answer was graphed and presented in this study.

2. *Process Box* – this is a rectangle with one flowline leading into it and one leading out of it. Each process box contained an elementary cognitive process reported by the students as they solved the items.

3. *Connector* – this is a circle connecting two flowlines in a diagram. In most cases, connectors represented junctions or links in the flow chart where students differed from one another.

4. *Flowline* – this is a line with a one-way arrow used to connect process boxes with one another or process boxes with start/stop boxes. Flowlines indicated the direction of information processing as students worked toward their solutions. Information was assumed to flow as a sequential rather than parallel process therefore only one elementary event is processed at a time and only one arrow per box is presented.

In the second step, the elementary cognitive processes in the flow charts were coded into more general categories associated with specific problem-solving strategies. For example, in a problem such as

$$4(x - 1) - 3x = 12, \text{ then } x = ?$$

students often used as many as five different elementary processes. However, these processes were indicative of a more general problem-solving strategy – namely, solve *x* by isolating the variable on one side of the equation. Both the elementary cognitive processes and the problem-solving strategies used by students to solve each of the 21 SAT algebra items were documented. Although both correct and incorrect responses were coded, only the correct responses are presented in this report. The decision to focus on correct responses stems from the nature of our psychometric procedure, the AHM, which is used to model correct response patterns. While the incorrect responses can be a valuable sources of diagnostic information (cf. Luecht, 2006), these data cannot be modeled, currently, with the AHM.

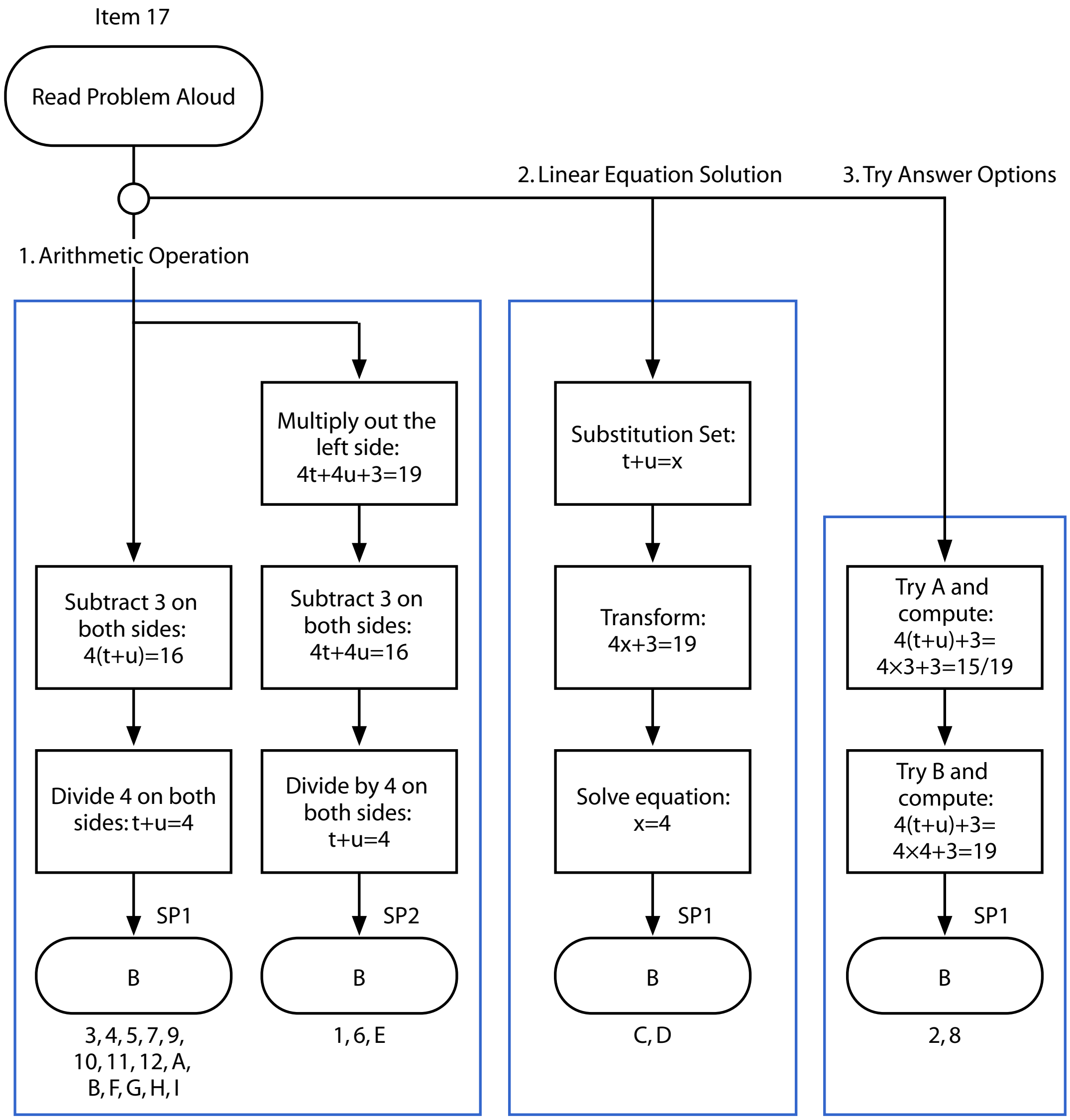
In the third step, to evaluate how well the attributes and the hierarchy specified in the Figure 1 (page 15) cognitive model matched the cognitive processes reported by students, the attribute descriptions were compared to the cognitive flow charts for each item. Two reviewers (Wang and Zhou) were asked to independently compare the student think aloud flow chart data to the cognitive models. Once the comparison was complete, the reviewers met to discuss their results with one another and with the first author of this study. All disagreements were discussed, debated, and resolved. Resolution of our disagreements required an iterative process where our attribute descriptions were continually refined and reworded, in light of our discussions and interpretations of the verbal report data.

Results

The results from the protocol analysis for Ratios and Algebra hierarchy are presented in Figures 2 to 10. Each solution path is labeled in the stop box at the bottom of the flow chart. Solution path one, for example, is labeled SP1. Males are assigned numbers from 1 through 12 whereas females are assigned letters from A through I.

As a prerequisite attribute, A1 represents *basic arithmetic skills* with operations (e.g., addition, subtraction, multiplication, and division of numbers). Item 17 serves as an example to illustrate this attribute. All 21 students correctly answered the item. Three strategies were involved: arithmetic operation, linear equation solution, and try answer options (Figure 2, next page). Seventeen of the 21 students adopted the first strategy, two adopted the second strategy, and the remaining two adopted the third strategy. Of the 17 students who used the first strategy, 14 students used SP1 and the remaining three used SP2 depending on the order of arithmetic operations.

Figure 2: The Problem-solving Steps Used by Examinees to Solve Item 17


Item 17
Read Problem Aloud
1. Arithmetic Operation
2. Linear Equation Solution
3. Try Answer Options
Multiply out the left side: 4t+4u+3=19
Subtract 3 on both sides: 4(t+u)=16
Subtract 3 on both sides: 4t+4u=16
Divide 4 on both sides: t+u=4
Divide by 4 on both sides: t+u=4
SP1
SP2
B
B
3, 4, 5, 7, 9, 10, 11, 12, A, B, F, G, H, I
1, 6, E
Substitution Set: t+u=x
Transform: 4x+3=19
Solve equation: x=4
SP1
B
C, D
Try A and compute: 4(t+u)+3= 4×3+3=15/19
Try B and compute: 4(t+u)+3= 4×4+3=19
SP1
B
2, 8

Item 17

If $4(t + u) + 3 = 19$, then $t + u =$

(A) 3

(B) 4

(C) 5

(D) 6

(E) 7

Attributes A2 and A3 both deal with factors. Attribute A2 includes *knowledge about the properties of factors*. For example, item 3 measures attribute A2. Sixteen of the 21 students correctly answered this item. Two strategies were adopted: applying knowledge of factors and plugging in numbers (Figure 3). Thirteen of the sixteen students used the first strategy while the remaining three students used the second strategy. However, the second strategy – plugging in numbers – does not reflect the skills associated with the attribute of knowledge about the properties of factors.

Figure 3: The Problem-solving Steps Used by Examinees to Solve Item 3

Item 3

If $(p + 1)(t - 3) = 0$ and p is positive, what is the value of t?

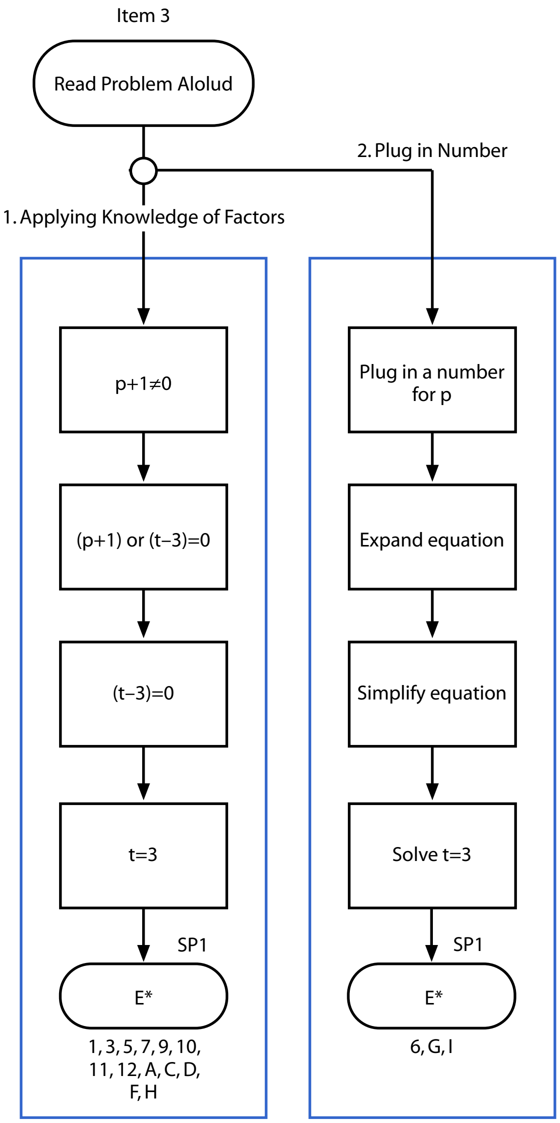
(A) –3

(B) –1

(C) 0

(D) 1

(E) 3



In addition to the knowledge about properties of factors in attribute A2, A3 involves the *skills of applying the rules of factoring*. Item 6 measures attribute A3. Nineteen students correctly answered item 6. Three strategies were used: applying the rules of factoring, plugging in random numbers, and solving equation. Of the 19 students who correctly answered this item, 14 adopted the first strategy, four adopted the second strategy, and the remaining one adopted the third strategy (Figure 4). However, the second strategy, plugging in random numbers, again, does not reflect the skills measured by this attribute.

Figure 4: The Problem-solving Steps Used by Examinees to Solve Item 6

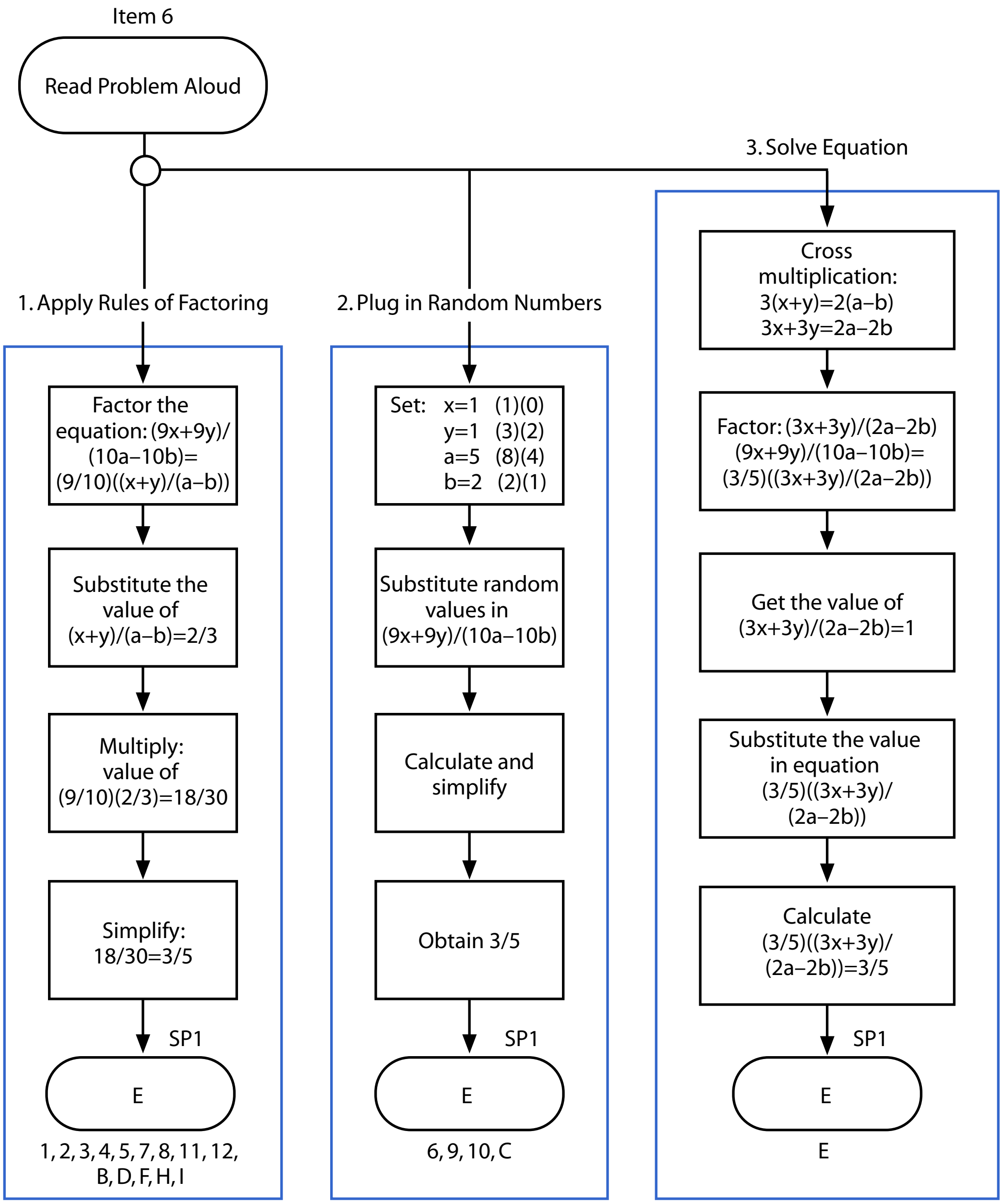

Item 6
Read Problem Aloud
1. Apply Rules of Factoring
2. Plug in Random Numbers
3. Solve Equation
Factor the equation: (9x+9y)/(10a–10b)=(9/10)((x+y)/(a–b))
Substitute the value of (x+y)/(a–b)=2/3
Multiply: value of (9/10)(2/3)=18/30
Simplify: 18/30=3/5
SP1
E
1, 2, 3, 4, 5, 7, 8, 11, 12, B, D, F, H, I
Set: x=1 (1)(0) y=1 (3)(2) a=5 (8)(4) b=2 (2)(1)
Substitute random values in (9x+9y)/(10a–10b)
Calculate and simplify
Obtain 3/5
SP1
E
6, 9, 10, C
Cross multiplication: 3(x+y)=2(a–b) 3x+3y=2a–2b
Factor: (3x+3y)/(2a–2b) (9x+9y)/(10a–10b)=(3/5)((3x+3y)/(2a–2b))
Get the value of (3x+3y)/(2a–2b)=1
Substitute the value in equation (3/5)((3x+3y)/(2a–2b))
Calculate (3/5)((3x+3y)/(2a–2b))=3/5
SP1
E
E

Figure 4: The Problem-solving Steps Used by Examinees to Solve Item 6 (continued)

Item 6

If $\frac{x+y}{a-b} = \frac{2}{3}$, then $\frac{9x+9y}{10a-10b} =$

(A) $\frac{9}{10}$

(B) $\frac{20}{23}$

(C) $\frac{20}{27}$

(D) $\frac{2}{3}$

(E) $\frac{3}{5}$

The self-contained sub-hierarchy contains six attributes. Among these attributes, A4 is the prerequisite for all other attributes in the sub-hierarchy. Attribute A4 has A1 as a prerequisite because A4 not only represents basic skills in arithmetic operations (i.e., A1), but it also involves the substitution of values into algebraic expressions which is more abstract and, therefore, more difficult than attribute A1. Item 18 measures attribute A4. All 21 students correctly answered the item. One dominant strategy, substitution, was adopted by all students (Figure 5). Depending on the order of substitution, two solution paths were identified for the dominant strategy of substitution. Twenty students substituted the values of variables consecutively into the algebraic expressions to obtain the final answer (SP1). The remaining student substituted an algebraic expression into another algebraic expression first and then substituted the values of variables to obtain the final answer (SP2).

Figure 5: The Problem-solving Steps Used by Examinees to Solve Item 18

Item 18

$k = 3wx$
$m = (w - 1)k$

If k and m are defined by the equation above, what is the value of m when $w = 4$ and $x = 1$?

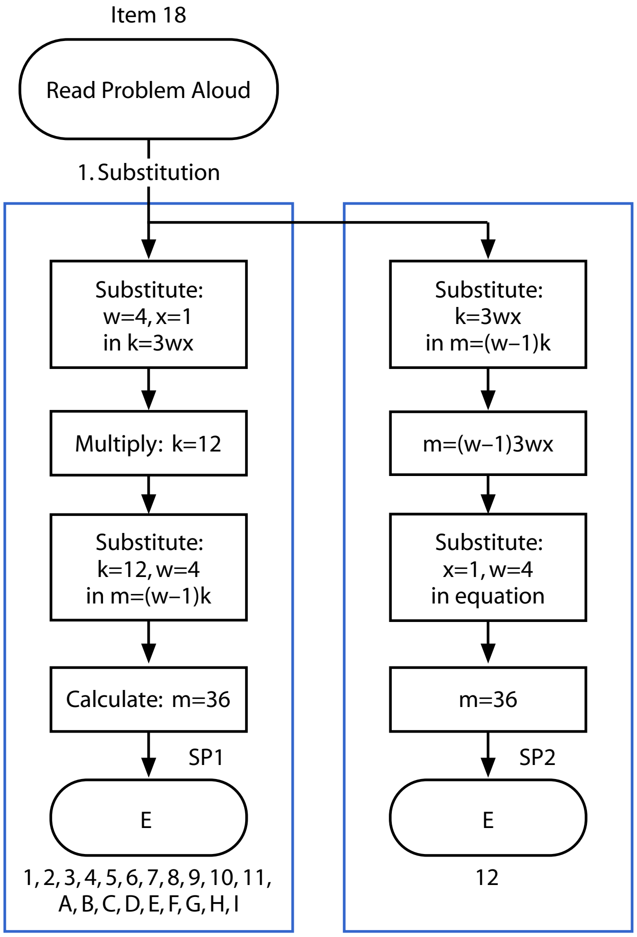
(A) 0

(B) 3

(C) 12

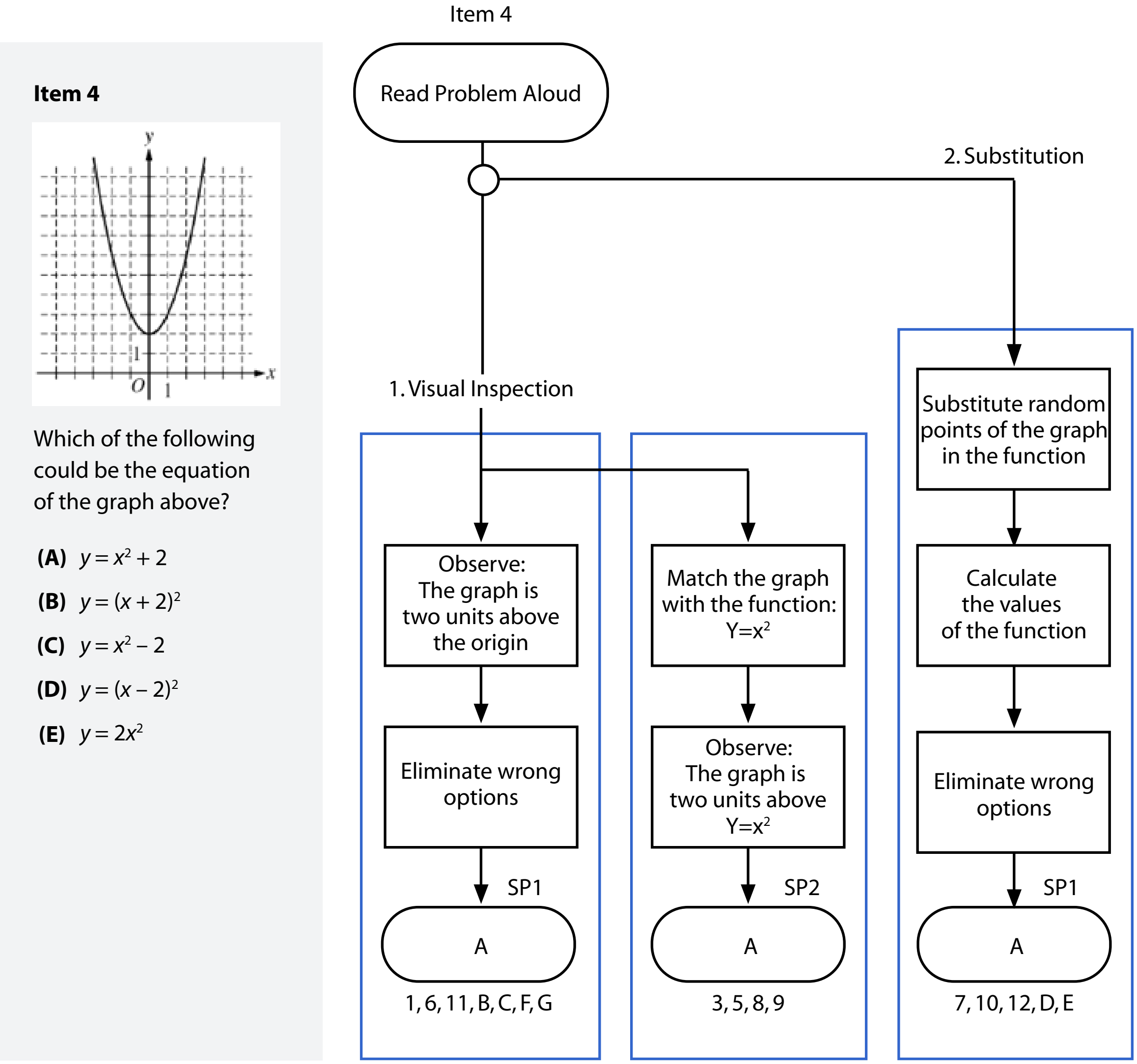
(D) 24

(E) 36


Item 18
Read Problem Aloud
1. Substitution
Substitute: w=4, x=1 in k=3wx
Multiply: k=12
Substitute: k=12, w=4 in m=(w–1)k
Calculate: m=36
SP1
E
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, A, B, C, D, E, F, G, H, I
Substitute: k=3wx in m=(w–1)k
m=(w–1)3wx
Substitute: x=1, w=4 in equation
m=36
SP2
E
12

The first branch, which contains A5 and A6, in the sub-hierarchy deals, mainly, with functional graph reading. Attribute A5 represents the *skills of mapping a graph of a familiar function with its corresponding function*. This attribute involves the knowledge about the graph of a familiar function and/or substituting points in the graph. Item 4 is an instance of attribute A5. Sixteen students correctly answered this item. In solving the item, two strategies were used: visual inspection and substitution of random points (Figure 6). Of the 16 students, 11 used the first strategy while the remaining five students used the second strategy. Of the 11 students who used the first strategy, two solution paths were generated. Seven students produced the answer by observing the graph and eliminating the wrong options and solving an equation (SP1) and four students produced the answer by finding the relationship between the graph and the graph of $y = x^2$ (SP2).

Figure 6: The Problem-solving Steps Used by Examinees to Solve Item 4



Item 4

Which of the following could be the equation of the graph above?

(A) $y = x^2 + 2$

(B) $y = (x + 2)^2$

(C) $y = x^2 - 2$

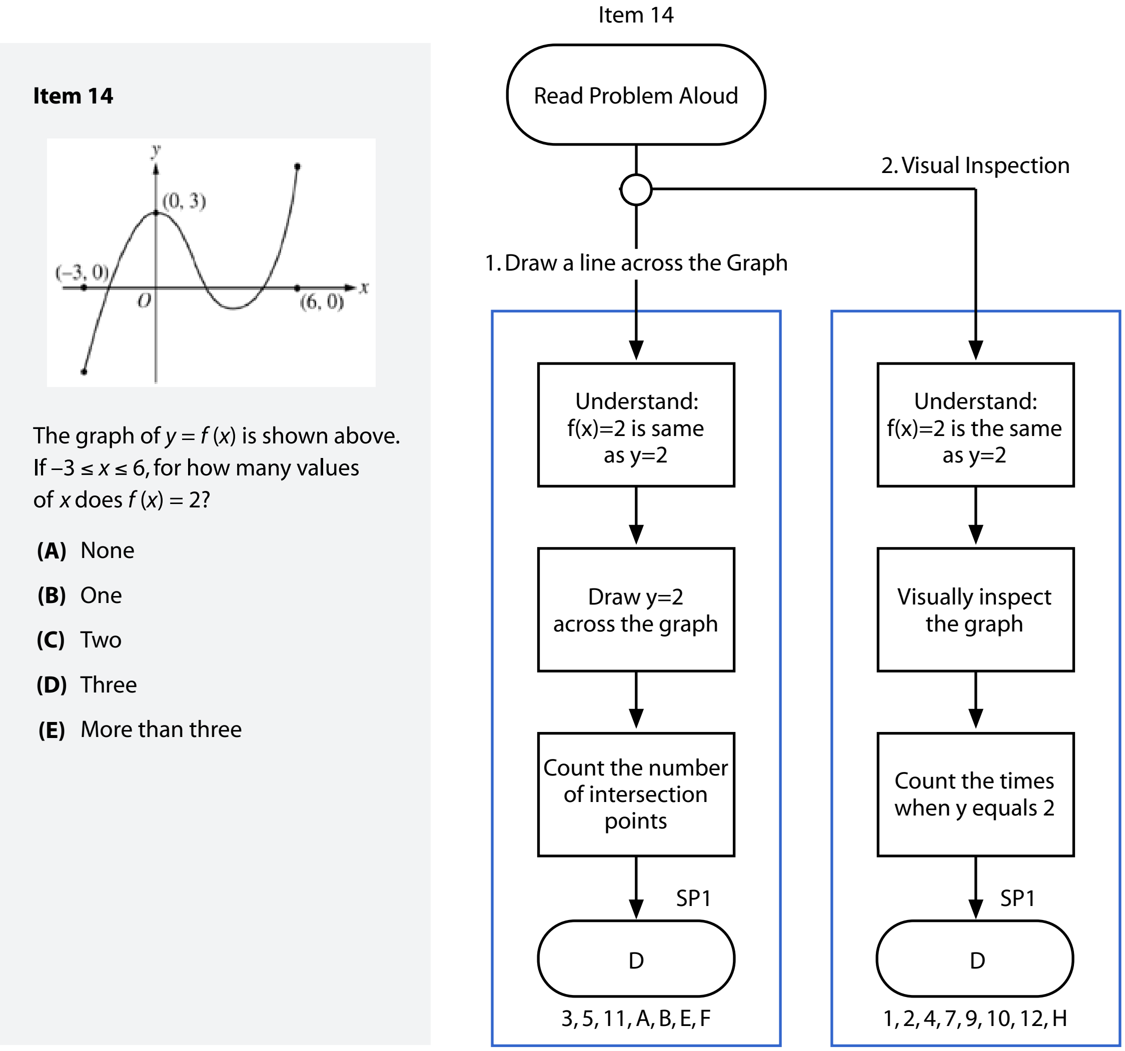
(D) $y = (x - 2)^2$

(E) $y = 2x^2$

y
x
O
1
1
Item 4
Read Problem Aloud
2. Substitution
1. Visual Inspection
Observe: The graph is two units above the origin
Eliminate wrong options
SP1
A
1, 6, 11, B, C, F, G
Match the graph with the function: Y=x²
Observe: The graph is two units above Y=x²
SP2
A
3, 5, 8, 9
Substitute random points of the graph in the function
Calculate the values of the function
Eliminate wrong options
SP1
A
7, 10, 12, D, E

Attribute A6, on the other hand, deals with the *abstract properties of functions, such as recognizing the graphical representation of the relationship between independent and dependent variables*. The graphs of less familiar functions, such as a periodic function or function of higher-power polynomials, may be involved. Therefore, A6 is considered more difficult than A5. Item 14 measures attribute A6. Fifteen students correctly answered this item. In solving the item, two strategies were used: drawing lines across the graph and visual inspection (Figure 7). Of the 15 students, seven used the first strategy while the remaining eight students used the second strategy.

Figure 7: The Problem-solving Steps Used by Examinees to Solve Item 14



Item 14

The graph of $y = f(x)$ is shown above. If $-3 \leq x \leq 6$, for how many values of x does $f(x) = 2$?

(A) None

(B) One

(C) Two

(D) Three

(E) More than three

y
(0, 3)
(−3, 0)
O
(6, 0)
x
Item 14
Read Problem Aloud
2. Visual Inspection
1. Draw a line across the Graph
Understand: f(x)=2 is same as y=2
Draw y=2 across the graph
Count the number of intersection points
SP1
D
3, 5, 11, A, B, E, F
Understand: f(x)=2 is the same as y=2
Visually inspect the graph
Count the times when y equals 2
SP1
D
1, 2, 4, 7, 9, 10, 12, H

.

The second branch in the sub-hierarchy considers the skills associated with advanced substitution. Attribute A7 requires the *skills to substitute numbers into algebraic expressions*. The complexity of A7 relative to A4 lies in the concurrent management of multiple pairs of numbers and multiple equations. Item 7 is an example of attribute A7. Twenty out of 21 students correctly answered the item. Two strategies were adopted: multiple substitution and pattern recognition (Figure 8). Nineteen out of the 20 students adopted the first strategy and obtained the correct answer by substituting the number pairs in the functions provided in the answer options. The remaining student obtained the correct answer by recognizing the pattern implied by the number pairs and then matching the pattern with the functions provided in the answer options.

Figure 8: The Problem-solving Steps Used by Examinees to Solve Item 7

Item 7

(x, y)
(0, 100)
(1, 99)
(2, 96)

Which of the following equations describes y in terms of x for all ordered pairs in the table above?

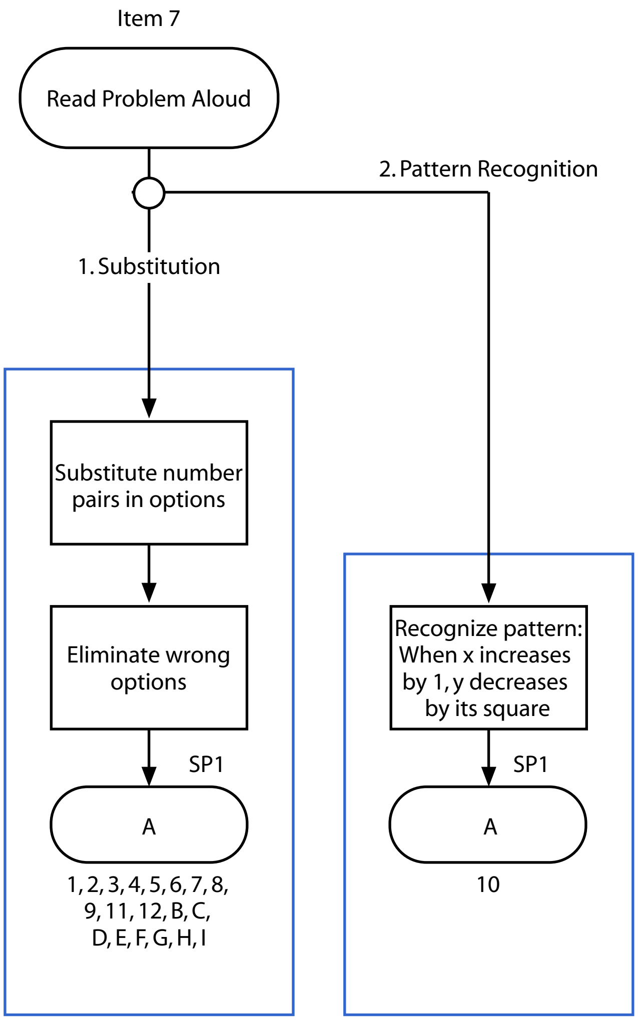
(A) $y = 100 - x^2$

(B) $y = 100 - x$

(C) $y = 100 - 2x$

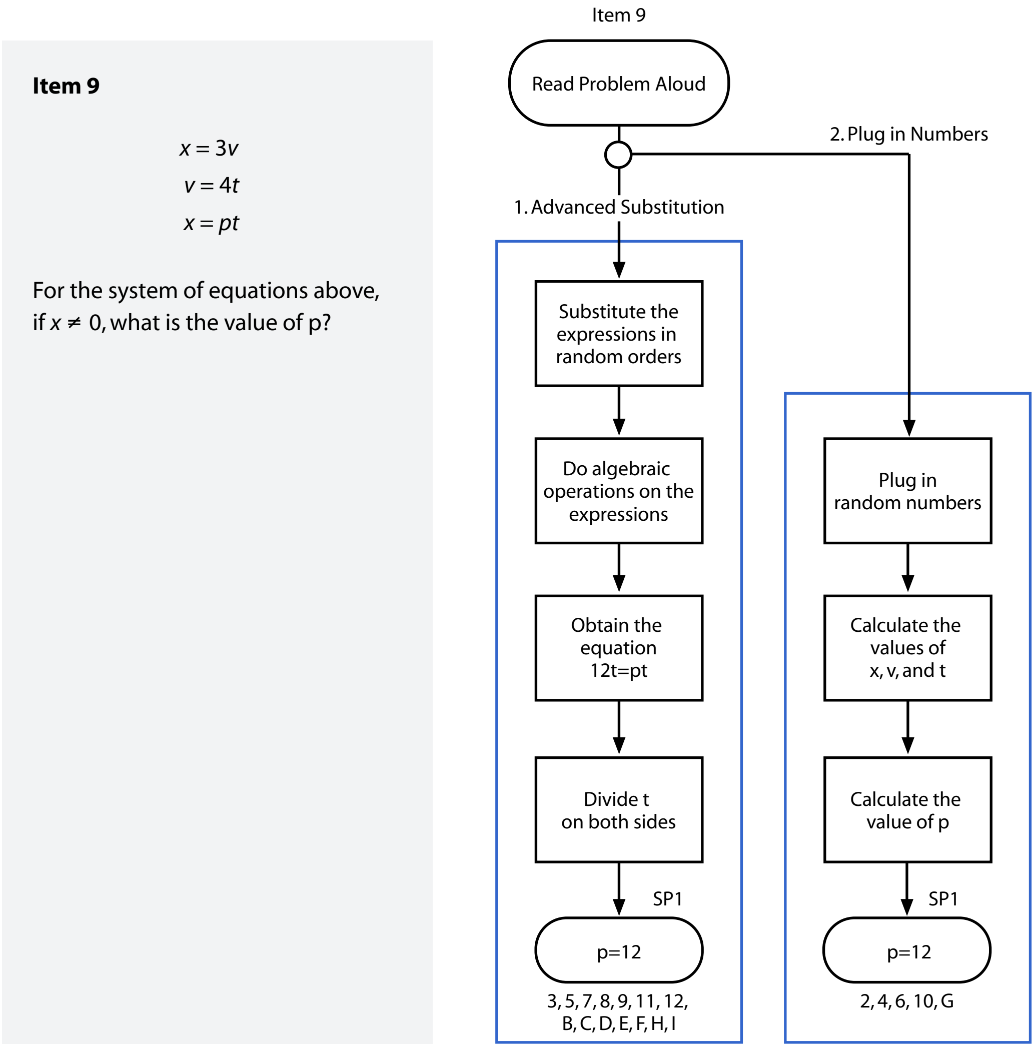
(D) $y = 100 - 4x$

(E) $y = 100 - 100x$


Item 7
Read Problem Aloud
2. Pattern Recognition
1. Substitution
Substitute number pairs in options
Eliminate wrong options
SP1
A
1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, B, C, D, E, F, G, H, I
Recognize pattern: When x increases by 1, y decreases by its square
SP1
A
10

Attribute A8 also represents the *skills of advanced substitution*. However, what makes A8 more difficult than A7 is that algebraic expressions, rather than numbers, need to be substituted into another algebraic expression. Item 9 measures attribute A7. Nineteen out of 21 students correctly answered the item. Two strategies were adopted: substitution and plugging in numbers (Figure 9). Fourteen students adopted the first strategy and five students adopted the second strategy. The five students who produced the correct answer by plugging in random numbers used a strategy unrelated to our inferences about their mastery of attribute A8, skills of substitution. Hence, this strategy, which leads to the correct answer, is inconsistent with our attribute description.

Figure 9: The Problem-solving Steps Used by Examinees to Solve Item 9



Item 9

$x = 3v$

$v = 4t$

$x = pt$

For the system of equations above, if $x \neq 0$, what is the value of p?

Item 9
Read Problem Aloud
2. Plug in Numbers
1. Advanced Substitution
Substitute the expressions in random orders
Do algebraic operations on the expressions
Obtain the equation 12t=pt
Divide t on both sides
SP1
p=12
3, 5, 7, 8, 9, 11, 12, B, C, D, E, F, H, I
Plug in random numbers
Calculate the values of x, v, and t
Calculate the value of p
SP1
p=12
2, 4, 6, 10, G

The last branch in the sub-hierarchy contains only one additional attribute, A9, related to *skills associated with rule understanding and application*. Item 15 measures attribute A9. Eighteen out of 21 students correctly answered the item and they adopted one dominant strategy: understanding and application of the rule (Figure 10).

Figure 10: The Problem-solving Steps Used by Examinees to Solve Item 15

Item 15

For all numbers x and y, let $x \Delta y$ be defined as $x \Delta y = x^2 + xy + y^2$. What is the value of $(3 \Delta 1) \Delta 1$?

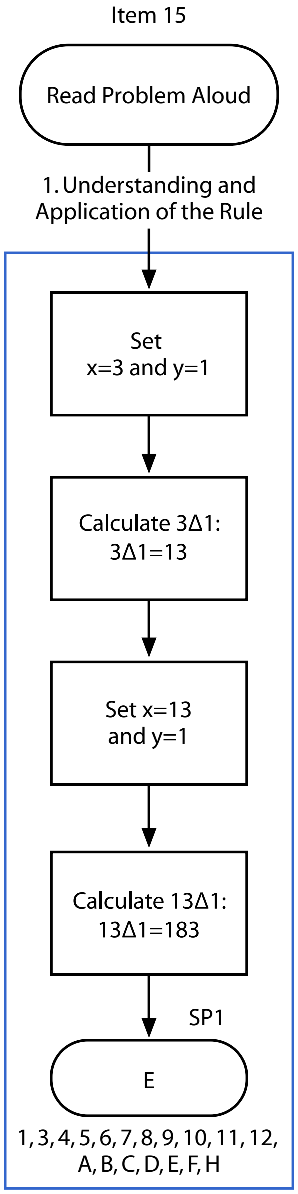
(A) 5

(B) 13

(C) 27

(D) 170

(E) 183



Results from the verbal reports reveal that the content-based cognitive model developed initially provided an excellent approximation to the actual student results. The verbal reports did not result in any structural changes to the Ratios and Algebra hierarchy. The reports did, however, allow us to develop a more concise description of each attribute as well as identify examples of how different strategies and solution paths within a single attribute can yield the same solution. Attribute grain size is a constant concern when developing a diagnostic test because the attribute must characterize the knowledge and skills used by all examinees as they solve items. If the attribute grain size is too fine, then strategy diversity produces multiple attributes that are plausible. If the attribute grain size is too coarse, then the diagnostic inferences are broad and, potentially, uninformative about the examinees' cognitive skills. Because of these concerns, the attribute definitions must be closely monitored, systematically developed, and carefully described. A summary of the attributes is presented in Table 1.

Table 1: Summary of the Attributes Required to Solve the Items in the Ratios and Algebra Hierarchy

Attribute A1	Represents the most basic arithmetic operation skills
Attribute A2	Includes knowledge about the properties of factors
Attribute A3	Involves the skills of applying the rules of factoring
Attribute A4	Includes the skills required for substituting values into algebraic expressions
Attribute A5	Represents the skills of mapping a graph of a familiar function with its corresponding function
Attribute A6	Deals with the abstract properties of functions, such as recognizing the graphical representation of the relationship between independent and dependent variables
Attribute A7	Requires the skills to substitute numbers into algebraic expressions
Attribute A8	Represents the skills of advanced substitution – algebraic expressions, rather than numbers, need to be substituted into another algebraic expression
Attribute A9	Relates to skills associated with rule understanding and application

After the attributes required to solve the test items are identified and the cognitive model of task performance is operationalized as a structured hierarchy, the analytic procedures in stage 1 can be conducted. That is, the adjacency, reachability, incidence, reduced-incidence, and expected response matrices are generated. The adjacency matrix for Ratios and Algebra hierarchy in Figure 1 (page 14) is shown in Matrix 1.

Matrix 1: Adjacency Matrix for the Ratios and Algebra Hierarchy

$$\begin{vmatrix} 010100000 \\ 001000000 \\ 000000000 \\ 000010101 \\ 000001000 \\ 000000000 \\ 000000010 \\ 000000000 \\ 000000000 \end{vmatrix}$$

In the adjacency matrix, a 1 in the position (k, k) indicates that attribute is directly connected in the form of a prerequisite to attribute k. The first row indicates, for example, that attribute A1 is a prerequisite to attributes A2 and A4.

In the reduced-incidence matrix[1] (Matrix 2), of order (k, i), for the Ratios and Algebra hierarchy, we identified 9 attributes (k rows) that were structured using the hierarchy. These attributes, in turn, were used to code 9 items (i columns). Item 17 (column 1), for instance, measured attribute A1 whereas item 6 (column 3) measured attributes A1, A2, and A3.

Matrix 2: Reduced-Incidence Matrix for the Ratios and Algebra Hierarchy

$$\begin{vmatrix} 111111111 \\ 011000000 \\ 001000000 \\ 000111111 \\ 000011000 \\ 000001000 \\ 000000110 \\ 000000010 \\ 000000001 \end{vmatrix}$$

The expected response matrix for the attribute hierarchy is specified in Matrix 3 (next page).

Matrix 3: Expected Response Matrix for the Ratios and Algebra Hierarchy

$$\begin{vmatrix}
000000000 \\
100000000 \\
110000000 \\
111000000 \\
100100000 \\
110100000 \\
111100000 \\
100110000 \\
110110000 \\
111110000 \\
100111000 \\
110111000 \\
111111000 \\
100100100 \\
110100100 \\
111100100 \\
100110100 \\
110110100 \\
111110100 \\
100111100 \\
110111100 \\
111111100 \\
100100110 \\
110100110 \\
111100110 \\
100110110 \\
110110110 \\
111110110 \\
100111110 \\
110111110 \\
111111110 \\
100100001 \\
110100001 \\
111100001 \\
100110001 \\
110110001 \\
111110001 \\
100111001 \\
110111001 \\
111111001 \\
100100101 \\
110100101 \\
111100101 \\
100110101 \\
110110101 \\
111110101 \\
100111101 \\
110111101 \\
111111101 \\
100100111 \\
110100111 \\
111100111 \\
100110111 \\
110110111 \\
111110111 \\
100111111 \\
110111111 \\
111111111
\end{vmatrix}$$

This matrix, of order (*j*, *i*), indicates that 58 different responses are expected by examinees 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 expected response matrix are the items that probe specific attribute combinations. When an examinee's attributes match those attributes measured by an item, a correct answer is expected.

Stage #2: Psychometric Analyses of SAT Algebra Model

In stage 2, psychometric analyses are conducted on the cognitive model. Data from a random sample of 5000 students who wrote these items on the March 2005 administration of the SAT were analyzed. First, the fit of the hierarchy in Figure 1 (page 14) was evaluated relative to the actual student response data from the random sample using the Hierarchy Consistency Index (HCI_j). The HCI_j assesses the degree to which an observed examinee response pattern is consistent with the attribute hierarchy. An HCI_j greater than 0.70 indicates good model-data fit. Second, attribute probabilities were computed. Attribute probabilities provide examinees with specific information about their attribute-level performance.

Applying the algebra hierarchy to the 5000 student sample, the mean HCI_j was high at 0.80. Because the index ranges from a maximum misfit of –1 to a maximum fit of 1, the value of 0.80 indicates strong model-data fit.

To compute the attribute probabilities, a neural network was used to evaluate the sample data. The input to train the neural network is the *expected response vectors* produced from the AHM analysis presented in matrix 3 (previous page). The expected response vector is derived from the algebra hierarchy which serves as our cognitive model of task performance. For each expected response vector there is a specific combination of examinee attributes (i.e., the transpose of the reduced-incidence matrix in matrix 2, page 31). 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 each association. Using nine hidden units, the network converged, meaning that an acceptable error level was achieved using a network defined with 9 input, 9 hidden, and 9 output units. The value for the root mean square was 0.0004 after 500 iterations.

Once the network has converged and the weights w_{ji} and v_{kj} established, the functional relationship between any examinee response vectors and their associated attributes can be defined by the following expressions

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

and

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

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

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

where q is the total number of hidden units, v_{kj} 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 elements of the final output vector (one element per attribute), M^*, are interpreted as probability estimates for the attribute (McClelland, 1998).

Six examples are presented in Table 2 (next page). The first three examples illustrate the attribute probabilities for observed response patterns that are consistent with the attribute hierarchy. Take, for instance, an examinee who possesses the first three attributes, A1 to A3 thereby producing the response pattern 111000000 (i.e., example 1). This observed response pattern is consistent with one of the 58 expected response patterns in matrix 3 (see row 4 of Matrix 3). The attribute probability levels for this response pattern are 0.96, 1.00, 0.99, 0.04, 0.00, 0.00, 0.00, 0.00, and 0.00 for attributes A1 to A9, respectively. Examples 2 and 3 illustrate the attribute probabilities associated with observed response patterns that are also consistent with the hierarchy.

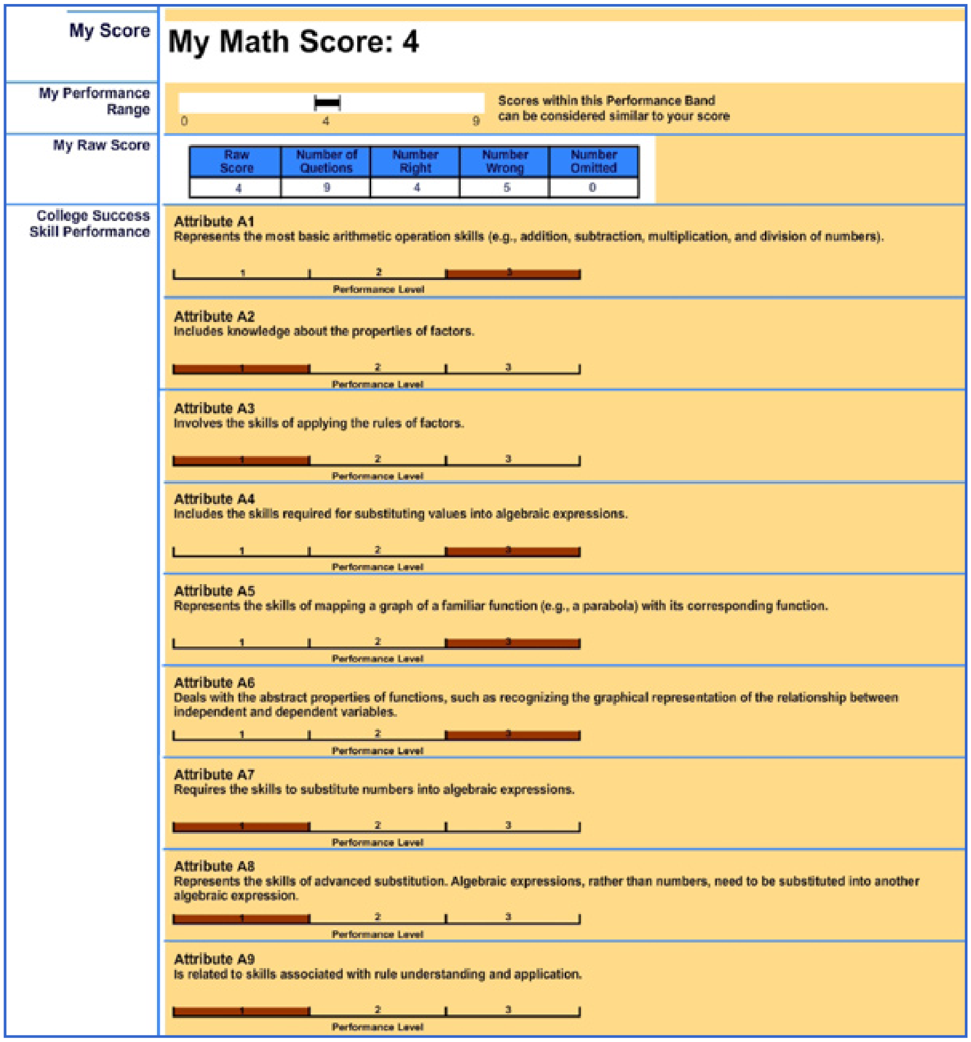
Table 2: Attribute Probabilities for Six Different Observed Examinee Response Patterns using the Ratios and Algebra Hierarchy

Pattern	Attribute Probability								
	1	2	3	4	5	6	7	8	9
Consistent									
1. A1 to A3	0.96	1.00	0.99	0.04	0.00	0.00	0.00	0.00	0.00
2. A1, A4 to A6	0.99	0.01	0.02	0.98	1.00	0.98	0.02	0.00	0.02
3. A1, A4 to A8	0.97	0.02	0.00	0.99	1.00	0.99	1.00	0.99	0.02
Inconsistent									
4. A1, A5 (missing A4)	0.98	0.01	0.00	0.66	0.82	0.04	0.00	0.00	0.00
5. A1, A7, A8 (missing A4)	0.92	0.03	0.00	0.77	0.00	0.00	1.00	1.00	0.00
6. A1, A5 to A8 (missing A4)	0.85	0.04	0.00	0.94	1.00	0.99	1.00	0.99	0.00

Examples 4 to 6 illustrate attribute probabilities for observed response patterns that are inconsistent with the attribute hierarchy. In other words, these response patterns are not one of the 58 expected response patterns in the expected response matrix. This inconsistency is overcome using the network because its purpose is to define the functional relationship for mapping the examinees' observed response pattern onto the expected response pattern using $M_k^* = F(a_k)$. For these three examples, attribute A4, which is the prerequisite attribute for these examples, is missing. In example 4, the examinee correctly solves the items measuring A1 and A5, but incorrectly solves the item measuring A4. The attribute probabilities for this observed response pattern are 0.98, 0.01, 0.00, 0.66, 0.82, 0.04, 0.00, 0.00, and 0.00 for attributes A1 to A9, respectively, indicating that the examinee possesses A1 and A5, and may possess A4. In this case, the evaluation of A4 is difficult because the examinee only solved *one item correctly that required A4*. In example 5, the examinee correctly solves the items measuring A1, A7, and A8, but incorrectly solves the item measuring A4. The attribute probabilities for this observed response pattern are 0.92, 0.03, 0.00, 0.77, 0.00, 0.00, 1.00, 1.00, and 0.00 for attributes A1 to A9, respectively, indicating that the examinee possesses A1, A7, and A8. The examinee may also possess A4 because it is unlikely that an examinee could solve two items measuring attributes A7 and A8 without possessing the pre-requisite A4. The same inference was based on only one item in example 4 (i.e., probability of A4 in example 4 was 0.66 whereas the probability of A4 in example 5 is 0.77). If all four items requiring the pre-requisite attribute are correctly solved, as in example 6 (i.e, A5 to A8), but the prerequisite is incorrectly solved (i.e., A4), then the probability is high that the examinee, in fact, possesses this prerequisite. Or, stated differently, it is *unlikely* that the examinee could solve the items associated with A5 to A8 without possessing A4, if the hierarchy is true. The attribute probabilities for this observed response pattern are 0.85, 0.04, 0.00, 0.94, 1.00, 0.99, 1.00, 0.99, and 0.00 for attributes A1 to A9, respectively, indicating that the examinee possesses A1, A5, A6, A7, A8, and, likely, A4.

A key advantage of the AHM is that it supports individualized diagnostic score reporting using the attribute probability results. The score reports produced by the AHM have not only a total score but also detailed information about what cognitive attributes were measured by the test and the degree to which the examinees have mastered these cognitive attributes. This diagnostic information is directly linked to the attribute descriptions, individualized for each student, and easily presented. Hence, these reports provide specific diagnostic feedback which may direct instructional decisions. To demonstrate how the AHM can be used to report test scores and provide diagnostic feedback, a sample report is presented in Figure 11 (next page).

Figure 11: A Sample Diagnostic Score Report for an Examinee Who Mastered Attributes A1, A4, A5, and A6



My Score

My Math Score: 4

My Performance Range

Scores within this Performance Band can be considered similar to your score

My Raw Score

Raw Score	Number of Quetions	Number Right	Number Wrong	Number Omitted
4	9	4	5	0

College Success Skill Performance

Attribute A1
Represents the most basic arithmetic operation skills (e.g., addition, subtraction, multiplication, and division of numbers).

Attribute A2
Includes knowledge about the properties of factors.

Attribute A3
Involves the skills of applying the rules of factors.

Attribute A4
Includes the skills required for substituting values into algebraic expressions.

Attribute A5
Represents the skills of mapping a graph of a familiar function (e.g., a parabola) with its corresponding function.

Attribute A6
Deals with the abstract properties of functions, such as recognizing the graphical representation of the relationship between independent and dependent variables.

Attribute A7
Requires the skills to substitute numbers into algebraic expressions.

Attribute A8
Represents the skills of advanced substitution. Algebraic expressions, rather than numbers, need to be substituted into another algebraic expression.

Attribute A9
Is related to skills associated with rule understanding and application.

In this example, the examinee mastered attributes A1 and A4 to A6 (Table 2, Consistent Pattern #2, page 35). Three performance levels were selected for reporting attribute mastery: non-mastery (attribute probability value between 0.00 and 0.35), partial mastery (attribute probability value between 0.36 and 0.70), and mastery (attribute probability value between 0.71 and 1.00). Other intervals for the attribute probability values could also be selected to determine mastery level. In an operational testing situation, the performance level for each mastery state would likely be determined using standard setting procedures. The results in the score report reveal that the examinee has clearly mastered four attributes, A1 (basic arithmetic operations), A4 (skills required for substituting values into algebraic expressions), A5 (the skills of mapping a graph of a familiar function with its corresponding function), and A6 (abstract properties of functions). The examinee has not mastered the skills associated with the remaining five attributes.

Section IV — Discussion

Summary of Study

The purpose of the study was to apply the attribute hierarchy method to a sample of algebra items from the March 2005 administration of the SAT to illustrate how the AHM could promote diagnostic inferences using data from an operational testing program. To begin, we defined the phrase *cognitive model* in educational measurement and explained why these models are important in the development and analysis of diagnostic assessments. Then, we presented the AHM. We described a two-stage approach for diagnostic testing where we defined the cognitive model of task performance and we evaluated the psychometric properties of the model. Finally, we applied the AHM to a sample of algebra items from the March 2005 administration of the SAT to demonstrate how the method could be used with actual student response data.

The Evolution of Cognitive Models

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, 2007a, p. 6). These models provide an interpretative framework to guide test development and psychometric analyses so test performance can be linked to specific cognitive inferences about the examinees. Recently, Mislevy (2006) described six aspects or steps in model-based reasoning in science. These six steps, presented in Table 3 (next page), provide an summary for considering our progress in developing cognitive models in algebra on the SAT.

Table 3: Six Steps in Model-based Reasoning from Mislevy (2006)

Model Formation	Establishing a correspondence between some real-world phenomenon and a model, or abstracted structure, in terms of entities, relationships, processes, etc. Includes scope and grain size to model and determining which aspects of the phenomenon to include and exclude.
Model Elaboration	Combining, extending, adding detail to a model, establishing correspondences across overlapping models. Often done by assembling smaller models into larger assemblages or fleshing out more general models with details.
Model Use	Reasoning through the structure of a model to make explanations, predictions, conjectures, etc.
Model Evaluation	Assessing the correspondence between the model components and their real-world counterparts, with emphasis on anomalies and important features not accounted for by the model.
Model Revisions	Modifying and elaborating a model for a phenomenon in order to establish a better correspondence. Often initiated by model evaluation procedures.
Model-based Inquiry	Working interactively between phenomena and models, using all of the previous steps. Emphasis on monitoring and taking actions with regard to model-based inferences vis-à-vis real-world feedback.

The first step is *model formation*. The researcher must establish a correspondence between some real-world phenomenon and a model. The empirical considerations for modeling cognitive skills using the AHM with hierarchical structures are described in Leighton et al. (2004) and Gierl, Wang, and Zhou (2007). The psychological considerations for modeling cognitive skills using psychometric methods and linking these skills to diagnostic inferences are outlined in Leighton and Gierl (2007a). The second aspect is *model elaboration*. In this step, models are developed and detailed. Over the course of two studies – Gierl, Wang, and Zhou, 2007 and Gierl, Leighton, Wang, Zhou, Gokiert, and Tan, 2007 – we have developed four cognitive models of algebra performance that describe different aspects of problem solving using sample items from Algebra I and II. One of the four models was illustrated in the current study. These models were elaborated using results from task analyses conducted by content specialists and from verbal think aloud protocols by SAT examinees using Algebra I and II items. Although the models have similarities (i.e., some models share attributes and items) and differences, they provide a concise yet detailed description of the *types of skills* that could be evaluated in algebra on the SAT. The third aspect, *model use*, provides structure to the model so that explanations and predictions can be made. By ordering the algebra attributes within a hierarchy of cognitive skills, our model specifies how the attributes are structured *internally* by SAT examinees when they solve test items. *Model evaluation* is the fourth step. Here, the correspondence between the model components and their real-world counterparts is assessed. The purpose of the current study was to evaluate the plausibility of a cognitive model in ratios and algebra by comparing representations of content specialists and SAT examinees to establish the correspondence between the model and examinees' problem-solving procedures. The protocol results indicate that the attribute descriptions capture the knowledge structures and processing skills examinees use to solve the ratio and algebra problems (although the attribute descriptions were refined in light of the verbal response data – see *model revision* step). The *HCI* results also demonstrate that there is a strong concordance between the expected response patterns produced by the attribute hierarchy and the observed response patterns generated by a random sample of 5000 students. In step five, *model revisions* can occur. Our evaluation of the cognitive model in Figure 1 (page 14) using student response data from verbal reports produced refinements in our attribute descriptions so the attributes characterized the knowledge structures and processing skills outlined in the examinee flow charts. Finally, in step 6, *model-based inquiry* can begin. In this step, the model is applied to student response data, where outcomes and actions are guided by model-based inferences. In other words, when steps 1 through 5 have been satisfied, the model can be used in step 6.

Diagnostic Inferential Errors

The attribute hierarchy serves as a representation of the underlying cognitive model of task performance. These models provide the means for connecting cognitive principles with measurement practices, in the spirit prescribed by Pellegrino, Baxter, and Glaser (1999):

> …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)

To develop these models, we must also assume that student performance is goal-directed, purposeful, and principled based on the instructional events that precede testing. Students are not expected to guess, plug in numbers from the multiple-choice alternatives to incomplete equations and expressions, or randomly apply option alternatives to information in the multiple-choice stem. We must make these assumptions because random performance is impossible to predict and, therefore, model. Moreover, random performances, even when they do lead to the correct answer, cannot inform instruction.

Unfortunately, our assumption about purposeful student performance is not always accurate, as the results of our study make clear. Students are motivated to produce the right answer even by the wrong means and the multiple-choice item format permits guessing. Two strategies unaccounted for with our cognitive models were used by students to correctly solve algebra items: plug in numbers and try answer options. A summary of the prevalence of these strategies is presented in Table 4.

Table 4: Summary of the Strategies Used to Correctly Solve Items But Excluded from the Cognitive Model

Attribute (Item)	Strategy	Number of Students
A1 (17)	Try answer options	2 (2 Males)
A2 (3)	Plug in numbers	3 (1 Male; 2 Females)
A3 (6)	Plug in numbers	4 (3 Males; 1 Female)
A8 (9)	Plug in numbers	5 (4 Males; 1 Female)

Although the number of strategies excluded from our cognitive models is not large and their use is not frequent, these problem-solving approaches will produce errors in our diagnostic inferences because we must assume that students possess the attributes outlined in the cognitive model if they produce a correct response. That is, we assume the correspondence between the cognitive model and the response outcome is perfect. One purpose of the current study was to evaluate this assumption using SAT items and examinees. Our results revealed that the algebra models provides an acceptable approximation to the cognitive skills initially identified by content specialists and used by students to solve the 21 algebra items. But, we also acknowledge that the correspondence between the cognitive model and the response outcome is not perfect.

Limitations of the Current Study

The primary limitations of the current study stems from the use of a *post-hoc or retrofitting approach* when identifying and applying the cognitive model of task performance to algebra items on the SAT. A post-hoc approach is limited because the attributes must be associated with existing test items (as no new items are developed when data are retrofit to a cognitive model) producing an *item-based* hierarchy rather than an *attribute-based* hierarchy. While item-based hierarchies are convenient because test items and examinee response data are available, they are also very limited because the cognitive model must be generated post-hoc and only existing items can be used to operationalize the attributes. Moreover, a post-hoc approach does not guarantee that either an appropriate cognitive model can be identified or an adequate number of items can be located on the test to measure the attributes in the cognitive model. In the current study, the HCI_j index suggested that we had adequate model-data fit for the Ratios and Algebra hierarchy presented in Figure 1 (page 14). However, only one item was associated with each attribute because these items provided the best representation of the attributes. If other SAT algebra items were included, the total number of indicators per attribute would increase, but at the expense of the cohesion in our attribute descriptions and in our HCI_j result.

But even when multiple items per attribute are identified, item-based hierarchies are inherently restricted because the distribution of items is uneven across attributes given that the items were never developed from a cognitive model. This uneven distribution of items detracts from the usefulness of an AHM analysis because some attributes will rarely, if ever, be observed resulting in precarious cognitive inferences for some problem-solving skills. And yet this limitation should be expected whenever item development proceeds without an explicit cognitive model of test performance because a large-scale test like the SAT was neither intended nor developed to evaluate hypotheses about the specific cognitive bases of group performance. As a result, the cognitive analysis of any existing test using retrofitting procedures will invariably produce a tenuous fit between the model (assuming that the model can be identified, initially) and the test data because the tests were not designed from an explicit cognitive framework.

These concerns raise an important question: Can an existing test be retrofit so it will yield cognitive diagnostic inferences? Two answers are offered. *Yes: an existing test can be retrofit so it will provide diagnostic information about the examinees*. However, the item-based hierarchy must be maintained, and the developer must increase the number of items measuring each attribute in the cognitive model. If the number of items can be increased, then this approach should yield less inferential error because a larger sample of examinee behaviour would be available for each attribute (i.e., four items per attribute provide a broader sample of the examinees' cognitive skills than one item per attribute). The intention in sampling the same cognitive skills on multiple test items is that the anomalous strategies we encountered – plug in numbers and try answer options – would not consistently lead to the correct solution. As a result, the statistical pattern recognition approach we used to produce the attribute probabilities would yield a lower value for examinees who use these anomalous strategies. However, it is worth repeating that when using a retrofitting approach, new items should be developed to measure each attribute after the initial cognitive model is identified and validated. This approach is also based on the assumption that an existing test actually contains a cognitive model that can be identified, validated, and used to produce diagnostic information.

Our second answer is more pessimistic: *No, an existing test cannot be retrofit without serious compromises and limitations* (as described in the previous paragraph). Instead, we advocate that the proper design be used to produce a cognitive diagnostic assessment where an attribute-based hierarchy is created by, initially, defining the cognitive model of task performance and then generate items systematically using the reduced incidence matrix from the AHM analysis to measure each attribute in the hierarchy. In other words, we use principled test design procedures (e.g., Luecht, 2006; Messick, 1984, 1989; Mislevy, Steinberg, & Almond, 2003) to specify the attribute-based cognitive model, and then create multiple, replicable test items to systematically measure each attribute in the model. Once the items are developed and the cognitive models are validated, confirmatory psychometric procedures can be used to compute the attribute probabilities for each examinee.

Directions for Future Research

Two directions for future research are proposed. First, we must increase our understanding of how to specify an appropriate grain size or level of analysis with a cognitive diagnostic assessment. Unfortunately, the factors required to identify the "appropriate" grain size are poorly defined. We noted, for example, that prerequisite skills could be broken into much more specific attributes. But we also claimed that more items would be required to measure these skills thereby adding new and more specific attributes along with new hierarchical structures. In other words, attributes and hierarchies can continually be specified at a smaller grain size thereby increasing the specificity of the cognitive inferences but also increasing the number of items required on the test to tap these attributes. Grain size should be closely linked to the specificity of the cognitive inference desired and to the reporting methods used.

Grain size also requires an important but seldom recognized tradeoff. One important benefit of a cognitive diagnostic assessment is that it yields specific inferences about examinees' cognitive skills. To measure these skills, items must be developed to probe each attribute systematically. However, a cognitive analysis at a fine grain size will, by necessity, limit construct representation and content coverage when a finite number of items is administered. Hence, the trade-off that must be struck stems from a shift in the breadth of construct and content coverage typical of classroom assessments and large-scale tests to depth of cognitive coverage typical of a cognitive diagnostic assessment.

Second, we need more concrete examples of how to implement cognitive diagnostic assessments in operational testing situations (cf. Mislevy, 2006). To overcome the problems associated with a cognitive retrofitting approach, we advocated for a more principled approach to test design and analysis where the cognitive model of task performance is identified and evaluated, then test items are developed to measure the attributes in the model, and, finally, model-based statistics are used to analyze the data and generate the test scores. This order of events – where the cognitive model is first identified and then the test items are developed – is needed because the hierarchical organization of attributes should guide the development of test items and, subsequently, the interpretation of test performance when using the AHM. In other words, by using the attribute hierarchy to develop test items, the developer achieves control over the specific attributes measured by each item which, in turn, leads to more specific inferences about the examinees' cognitive skills. Moreover, when a cognitive model is developed before the test items, the reduced incidence matrix can guide test development and test score interpretation because this matrix represents the *cognitive blueprint* for the exam. Hence, our goal is to design and analyze a test using a more principled approach so all of the benefits associated with the AHM can be realized and demonstrated in a practical testing context.

References

Anderson, J.R. (1996). ACT: A simple theory of complex cognition. *American Psychologist, 51*, 355–365.

Cui, Y. (2007). *The hierarchy consistency index: Development and analysis*. Unpublished Doctoral Dissertation. University of Alberta: Edmonton, Alberta, Canada.

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.

Dawson, M.R.W. (1998). *Understanding cognitive science*. Malden, MA: Blackwell.

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.

Fodor, J.A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Gierl, M.J., Cui, Y., & Hunka, S. (2007, April). *Using connectionist models to evaluate examinees' response patterns on tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Gierl, M.J., & Leighton, J.P. (2007). Linking cognitively-based models and psychometric methods. In C. R. Rao & S. Sinharay (Eds.) *Handbook of statistics: Psychometrics, Volume 26* (pp. 1103–1106). North Holland, UK: Elsevier.

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.

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., Leighton, J.P., Wang, C., Zhou, J., Gokiert, R., & Tan, A. (2007). *Developing and validating cognitive models of algebra performance on the SAT*© (Research Report). New York: The College Board.

Gierl, M.J., Wang, C., & Zhou, J. (2007). *Using the attribute hierarchy method to develop cognitive models and evaluate problem-solving skills in algebra on the SAT*© (Research Report). New York: The College Board.

Kuhn, D. (2001). Why development does (and does not occur) occur: Evidence from the domain of inductive reasoning. In J. L. McClelland & R. Siegler (Eds.), *Mechanisms of cognitive development: Behavioral and neural perspectives* (pp. 221–249). Hillsdale, NJ: Erlbaum.

Leighton, J.P. (2004). Avoiding misconceptions, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*, 6–15.

Leighton, J.P., & Gierl, M.J. (2007a). 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. (2007b). Verbal reports as data for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. (pp. 146–172). Cambridge, UK: Cambridge University Press.

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.

Lohman, D.F. (2000). Complex information processing and intelligence. In R.J. Sternberg (Ed.), *Handbook of intelligence* (pp. 285–340). NY: Cambridge University Press.

Luecht, R.M. (2006). *Engineering the test: From principled item design to automated test assembly*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Luecht, R.M. (2007). Using information from multiple-choice distractors to enhance cognitive-diagnostic score reporting. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and application* (pp. 319–340). Cambridge, UK: Cambridge University Press.

McClelland, J.L. (1998). Connectionist models and Bayesian inference. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition*. Oxford: Oxford University Press. 21–53.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*, 215–237.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education, Macmillian.

Mislevy, R.J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–306). Washington, DC: American Council on Education.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–62.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, and R. Glaser (Eds.). Board on Testing and Assessment, Center for Education. Washington, DC: National Academy Press.

Pellegrino, J.W. (2002). Understanding how students learn and inferring what they know: Implications for the design of curriculum, instruction, and assessment. In M.J. Smith (Ed.), *NSF K–12 Mathematics and Science Curriculum and Implementation Centers Conference Proceedings* (pp. 76–92). Washington, DC: National Science Foundation and American Geological Institutue.

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.

Royer, J.M., Cisero, C.A., & Carlo, M.S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research, 63*, 201-243.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986a). Learning representations by back-propagating errors. *Nature, 323*, 533–536.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986b). *Parallel distributed processing* (Vol. 1). Cambridge, MA: MIT Press.

Taylor, K.L., & Dionne, J.P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92*, 413–425.

Endnote

[1] Only a sample of the attribute-by-item patterns is used in our item-based hierarchy. Hence, this matrix is called a sample *Q_r matrix*. It does not measure all combination of the cognitive components in the Figure 1 cognitive model. For example, we do not have one item that measures attributes A1, A2, A3, and A4. A sample Q_r matrix is employed because we are retrofitting a cognitive model to nine existing items – no new items were developed for our study. The complete Q_r matrix for the cognitive model in Figure 1 (page 15) is of order (9, 58).

Acknowledgements

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

We thank Bihua Xiang, Andrew Turnbull, Aimee Levy, and Mary Roberts for their help preparing this manuscript.

Author Biographies

Mark J. Gierl is Professor of Educational Psychology and Canada Research Chair in Educational Measurement, Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5; mark.gierl@ualberta.ca. He earned his Ph.D. in quantitative methods from the University of Illinois, Urbana-Champaign, in 1996. His primary research interests are cognitive diagnostic assessment; assessment engineering, including construct mapping, automated item generation, and automated test assembly; differential item and bundle functioning; psychometric methods for evaluating test translation and adaptation.

Changjiang Wang is a psychometrician in the Psychometric & Research Services at Harcourt Assessment in San Antonio, Texas. He earned his Ph.D. in educational psychology at the University of Alberta, Canada in 2007. His primary research interests include application of classical test theory and item response theory to practical testing problems such as item/test bias, use of cognitive psychology for understanding test performance, and language testing.

Jiawen Zhou is a Ph.D. student in the Centre for Research in Applied Measurement and Evaluation at the University of Alberta in Edmonton, Alberta, Canada. Her research interests include cognitive diagnostic assessment, automated item generation, and computer adaptive testing.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College