

## **Using the Attribute Hierarchy Method to Identify and Interpret Cognitive Skills that Produce Group Differences**

**Mark J. Gierl, Yinggan Zheng, and Ying Cui**  
*Centre for Research in Applied Measurement and Evaluation,  
University of Alberta*

*The purpose of this study is to describe how the attribute hierarchy method (AHM) can be used to evaluate differential group performance at the cognitive attribute level. The AHM is a psychometric method for classifying examinees' test item responses into a set of attribute-mastery patterns associated with different components in a cognitive model of task performance. Attribute probabilities, computed using a neural network, can be estimated on each attribute for each examinee thereby providing specific information about the examinee's attribute-mastery level. These probabilities can also be compared across groups. We describe a four-step procedure for estimating and interpreting group differences using the AHM. We also provide an example using student response data from a sample of algebra items on the SAT to illustrate our pattern recognition approach for studying group differences.*

Assessment engineering (AE) is emerging as a new research area in educational and psychological measurement (Luecht, 2006a, 2006b). AE is an innovative approach to measurement, where engineering-like principles are used to direct the design and the analysis of assessments as well as the scoring and the reporting of the results. With this approach, an assessment begins with specific, empirically derived cognitive models (e.g., Leighton & Gierl, 2007). Next, assessment task templates are created using established frameworks derived from the cognitive model to produce test items. Finally, psychometric methods are applied to the examinee response data, typically in a confirmatory mode, to produce scores that are interpretable (Luecht, Gierl, Tan, & Huff, 2006).

AE differs from more traditional approaches to test design and analysis in three fundamental ways. First, cognitive models guide task design and item development, rather than content-based test specifications. While the categories in content blueprints can be included in the task templates, the assessment principles used to develop items are based on cognitive principles and, thus, provide more specific information for measuring problem-solving skills. Second, task templates are created to control and manipulate both the content and cognitive attributes of the items. Item writers are required to use the templates during development, thereby producing items that adhere to strict quality controls and that meet high psychometric standards. Third, psychometric models are employed in a confirmatory, versus exploratory, manner to assess the model-data fit relative to the intended underlying structure of the constructs or traits the test is designed to measure. The outcomes from these model-data fit analyses also provide developers with guidelines for

specific modifications to the cognitive models and task templates, as needed, to facilitate the acquisition of data that support the intended assessment inferences.

Leighton, Gierl, and Hunka (2004), introduced the attribute hierarchy method (AHM; see also Gierl, Leighton, & Hunka, 2007). The AHM is a psychometric procedure for classifying examinees' test item responses into a set of structured attribute patterns associated with different components from a cognitive model of task performance. The AHM can be considered an *AE method* because it is guided by a cognitive model; this model, in turn, produces task templates that guide item construction; and finally the examinee response data are analyzed using a confirmatory approach (see, e.g., Gierl & Zhou, in press). Taken together, these three AE features help ensure that test score inferences produced with the AHM are linked directly to the underlying cognitive model that characterizes examinee performance.

Because the AHM uses AE principles in the development of items and in the evaluation of examinee response data, it holds promise in helping researchers and practitioners to better understand the psychology of complex test performance. For instance, the AHM could be used to evaluate differential item functioning (DIF). Many researchers and practitioners now agree that the study of DIF items using the conventional approach—subjecting statistically identified items to the evaluation of content reviewers—leads to inconclusive results about *why* group differences occur (Angoff, 1993; Camilli & Shepard, 1994; Engelhard, Hansche, & Rutledge, 1990; Gierl, 2005; Gierl, Bisanz, Bisanz, & Boughton, 2003; O'Neill & McPeck, 1993; Roussos & Stout, 1996; Stout, 2002; Sudweeks & Tolman, 1993). In fact, the authors of the *Standards for Educational and Psychological Testing* (1999) concluded:

*Although DIF procedures may hold some promise for improving test quality, there has been little progress in identifying the causes or substantive themes that characterize items exhibiting DIF. That is, once items on a test have been statistically identified as functioning differently from one examinee group to another, it has been difficult to specify the reasons for the differential performance or to identify a common deficiency among the identified items. (p. 78)*

Yet, the study of group differences using AE principles with a psychometric method such as the AHM may help resolve this disjunction between the substantive and statistical analyses because each step is unified within a framework that could lead to a better understanding of why group differences occur. Hence, methods like the AHM, which promote a more principled approach to test design and analysis, hold promise for bridging the gap between substantive and statistical DIF outcomes so that group differences can be more easily identified and interpreted. The purpose of this study is to present and illustrate new analytic methods that can be used with the AHM to study the cognitive basis of group differences on tests. We begin with an overview of the AHM, as it applies to AE.

### **Overview of Attribute Hierarchy Method**

The AHM is based on the assumption that test performance depends on a set of hierarchically ordered competencies called *cognitive attributes* (or simply attributes, for short). Attributes are basic cognitive processes or skills required to solve test

items. The examinee must possess these attributes to answer test items correctly. Hence, attributes can be viewed as sources of cognitive complexity in test performance. The AHM was developed to address two specific problems associated with *cognitive model development* and *statistical pattern recognition* (Gierl, 2007).

### *Developing Cognitive Models with the AHM*

To make specific inferences about problem solving, cognitive models are required to operationalize the construct of interest. A cognitive model in educational measurement refers to a simplified description of human problem solving on standardized tasks at some convenient grain size or level of detail in order to facilitate explanation and prediction of examinees' performance, including their strengths and weaknesses (Leighton & Gierl, 2007). These models provide an interpretative framework that can guide item development so that test performance can be linked to specific cognitive inferences about examinees' knowledge, processes, and strategies. These models also provide the means for connecting cognitive principles with measurement practices so that the psychological basis of test performance—including subgroup similarities and differences—can be studied systematically.

A cognitive model of task performance is operationalized using attributes that are specific at a small grain size to highlight the cognitive skills that underlie test performance. With the AHM, a cognitive model also reflects a *hierarchy of cognitive processes* within a domain because the cognitive processes share dependencies and function within a much larger network of inter-related skills. Assessments based on cognitive models should be developed so that test items directly measure specific cognitive processes of increasing complexity in the examinees' understanding of a domain. The items can also be designed with this hierarchical order so that performance is linked directly to information about examinees' cognitive skills relative to the components in the cognitive model in mind. Using this approach, strong inferences about examinees' cognitive skills can be made because the small grain size in these models help the developer measure specific knowledge and skills required to perform competently on testing tasks.

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, such that the  $ij$ th element represents the absence (i.e., 0) or presence (i.e., 1) of a direct connection between attributes  $A_i$  and  $A_j$ . The adjacency matrix is of upper triangular form. Take, for example, the cognitive model in Figure 1 cast into an attribute hierarchy. The  $A$  matrix is given as

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

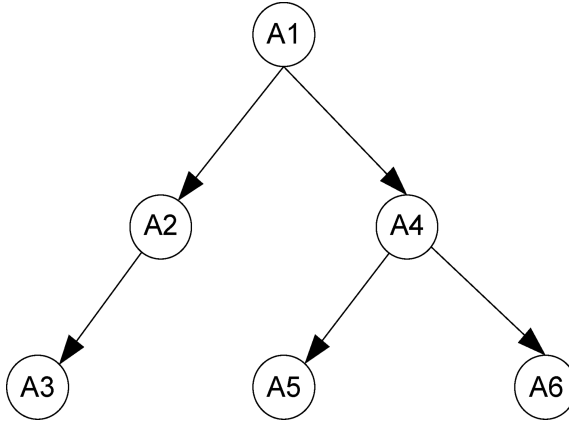


FIGURE 1. A hypothetical hierarchy containing six attributes.

where attribute A1 is a direct prerequisite to attributes A2 and A4. This hierarchical relationship is expressed in the first row of matrix 1 by the positions of a 1 in columns two and four. 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 for the R matrix to reach invariance,  $n = 1, 2, \dots, m$ , and  $I$  is the identity matrix.

Or, stated differently, to obtain the R matrix from the A matrix,  $(A + I)$  is repeatedly powered, using Boolean addition and multiplication rather than using the normal addition and multiplication, until the result is invariant. That is,  $(A + I)^n$  is formed repeatedly with  $n = 1, 2, \dots, m$ , until invariance is obtained. The R matrix for the model in Figure 1 is specified as

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

Next, the potential pool of items is generated. 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. That is, the Q matrix describes the attributes required by the examinee to obtain a correct answer to each item. This matrix can be reduced to form the *reduced Q matrix* ( $Q_r$ ) by imposing the constraints of the attribute hierarchy as defined

in the R matrix. The  $Q_r$  matrix represents the items from the potential pool that fit the constraints defined in the attribute hierarchy. It also represents the attribute combinations that examinees will possess, if the hierarchy is true. 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 for the model in Figure 1 is given as

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}. \quad (3)$$

Given a hierarchy of attributes, the expected response patterns for a group of examinees are generated. The *expected response matrix* (E) is created, again using Boolean inclusion, where the algorithm compares each row of the attribute pattern matrix (which is the transpose of the  $Q_r$  matrix) to the columns of the  $Q_r$  matrix. The expected response matrix, of order  $(j, i)$ , is calculated, where  $j$  is the number of examinees and  $i$  is the reduced number of items resulting from the constraints imposed by the hierarchy. The expected response matrix for the Figure 1 model is

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}. \quad (4)$$

In this example, an additional row is added to the expected response matrix. This row contains only 0s, as it represents the item response pattern for examinees who have not mastered any of the attributes.

In sum, two AE principles are used in the *Model Development* stage of the analysis. First, the cognitive model of task performance is specified. This model defined the construct measured by the test. The cognitive model is operationalized by describing the attributes and then ordering these attributes hierarchically in the A matrix. Second, the cognitive model guides the creation of the task templates required

for item construction in the form of the  $Q_r$  matrix. This matrix can be interpreted as the *cognitive test specification*, as it contains the attribute-by-item combination for each component of the cognitive model outlined in the A matrix. As a result, the elements in the  $Q_r$  matrix can be used to develop items that measure each specific attribute combination defined in the hierarchy. Then, in the pattern recognition stage, as described in the next section, examinees' observed response patterns can be analyzed according to the cognitive characteristics probed by each item.

*Statistical Pattern Recognition with the AHM*

An examinee's observed response pattern is judged relative to expected response patterns with the AHM under the assumption that the cognitive model is true. Hence, the purpose of the statistical pattern recognition analysis is to estimate the probability that an examinee possess specific attribute combinations based on their response patterns. These probabilities provide examinees with specific information about their attribute-level mastery as part of the test reporting process. To estimate the probability that examinees possess specific attributes, given their observed item response pattern, an artificial neural network approach is used (Gierl, Cui, & Hunka, 2007).

The neural network is considered a confirmatory pattern recognition analysis because the input and output are specified a priori using the cognitive model as operationalized using the attribute hierarchy. The input to train the neural network are the expected response vectors derived from the cognitive model. For each expected response vector, there is also a specific combination of examinee attributes described in the transpose of the  $Q_r$  matrix, which is used as the output to train the neural network. Recall, 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 specified by the hierarchy. The examinee attribute pattern matrix can be obtained by transposing the  $Q_r$  matrix,  $Q_r^T$ , and adding a row of 0s to the  $Q_r^T$  matrix, which is now of the order  $(j, k)$  with  $j$  as the number of examinees and  $k$  is the number of attributes.<sup>1</sup> In other words, the examinee attribute pattern matrix can be interpreted as the attributes (columns) possessed by the examinees (rows). It is given as

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad (5)$$

The attribute pattern matrix contains the possible attribute vectors, where students' mastery or nonmastery of individual attributes is specified, under the assumption that the hierarchy is true. Each element in the attribute vectors specifies the probability that the student has mastered the corresponding attribute, where a higher value indicates that the examinee has a higher probability of possessing the attribute. The attribute pattern matrix, like the expected response matrix, is derived from the cognitive model and, thus, specifies the attribute vectors that should be associated with each expected response pattern, where student responses are clearly explained by the presence or the absence of the attributes without any errors or "slips." As a result, these attribute vectors contain only 1s and 0s, where 1s indicate that the student has mastered the corresponding attributes and 0s suggest that the student has not.

In a real testing situation, however, the observed student response vector might not be completely consistent with the expected response vectors because of slips from 1 to 0 or from 0 to 1, which lead to the uncertainty as embodied by the attribute probability values. To estimate the attribute probabilities associated with each observed response vector, the relationship between the expected response vectors with their associated attribute vectors is established by presenting each expected response vector to the network repeatedly until it learns each association. The final result is a set of weight matrices that can be used to transform an observed response vector to its associated attribute probability vectors. The transformed result can be interpreted as the attribute probability, scaled between 0 and 1, where a higher value indicates that the examinee has a higher probability of possessing a specific attribute (McClelland, 1998).

A parallel-processing, multilayer architecture is used in the neural network (Rumelhart, Hinton, & Williams, 1986a, 1986b). This network transforms the stimulus received by the input unit to a signal for the output unit through the hidden units. The contribution of each input unit  $i$  to hidden unit  $j$  is determined by weight,  $w_{ji}$ . Similarly, the contribution of each hidden unit  $j$  to output unit  $k$  is determined by weight,  $v_{kj}$ . Learning is deemed to occur when the output layer, containing the desired response output (i.e., the attribute patterns), is correctly associated with the exemplars, as indicated by the value of the root mean square error. That is, the connection weights in the hidden layer transforms 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 with  $i$  ranging from 1 to  $p$  for the input node and  $j$  ranging from 1 to  $q$  for the hidden layer node.  $S_j$  is then transformed by the logistic function

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

Similarly, the output layer produces a weighted linear combination of their inputs, which are transformed to nonlinear weighted sums that are passed to every output

layer unit to produce the final attribute-level responses. The output,  $S_j^*$ , from every hidden layer unit is passed to every output layer unit where 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. In other words,

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

where  $T_k$ , also called the activation function, is the weighted sum for the  $k$ th output node using weights  $v_{kj}$ , with  $j$  ranging from 1 to  $q$  for the hidden layer nodes.  $T_k$ , like  $S_j$ , is transformed by the logistic function to  $T_k^*$ . Because the activation function is scaled using the logistic transformation, the output values range from 0 to 1.

The final result is a set of weight matrices, one for cells in the hidden layer and one for the cells in the output layer, that can be used to transform any examinee response vector to its associated attribute vector. The functional relationship for mapping the examinees' observed response patterns onto the expected response patterns so that their attribute patterns can be interpreted as probabilities is given as follows. If

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

and

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

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

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

where  $q$  is the total number of hidden units,  $v_{kj}$  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  $i$ . Using this transformation, attribute probabilities can be computed for each observed response pattern thereby providing examinees with specific information about their attribute-mastery level.

From the *Pattern Recognition* stage in the AHM analysis, one AE principle is used: The examinee response data are analyzed using a confirmatory analytic approach. The neural network uses the expected response vectors as input and the examinee attribute patterns as output. Both input and output are derived from the cognitive model specified in the attribute hierarchy. The neural network produces weight matrices to associate the expected response vector with the examinee attribute patterns which, in turn, are used to transform any observed response vector to its associated attribute vector. The transformed result is interpreted as the examinees' attribute probabilities.

### *Using the AHM to Identify and Interpret Differential Performance on Test*

The hierarchy serves as a cognitive model that specifies the attributes examinees use to solve test items. As a result, the attribute hierarchy can guide the study of cognitive factors that produce differential performance by systematically evaluating which attributes elicit group differences. Attribute-level differential functioning, hereafter referred to ADF, can be evaluated on a *studied attribute* by comparing the probabilities that different groups possess this attribute. To ensure that the ability of examinees from the focal and reference groups are comparable before the studied attribute is evaluated, examinees' score are aligned on the *matching attributes*. *ADF occurs when examinees, with the same matching attribute pattern but from different groups, have unequal probabilities of responding to items that measure the studied attribute.*

An ADF analysis has four steps. In step 1, the ADF hypotheses are specified. In step 2, the probability that examinees have mastered the studied attributes in both focal and reference groups is estimated. In step 3, examinees in the reference and the focal group are matched on attributes that are independent of the studied attribute. In step 4, the magnitude and direction of group differences on the studied attributes are estimated using the ADF statistic,  $\beta_{ADF_k}$ , and then tested. To illustrate these four steps, an example using the attribute hierarchy in Figure 1 is used.

*Step 1: Specify the ADF hypotheses.* The attribute hierarchy is used to generate hypotheses about the nature of attribute-related group differences. Typically, the ordering of the attributes provides a logical basis for generating ADF hypotheses because the hierarchy specifies the ordered dependencies among the attributes according to an underlying cognitive model. For example, in Figure 1, attributes A1 and A4 are prerequisites for attribute A5. Suppose that two groups of examinees (e.g., English- versus French-speaking students) perform differently on attribute A5. Because A1 and A4 are prerequisite skills to A5, the source of the difference is unclear. To overcome this interpretative problem, we can evaluate the presence of ADF among the studied attribute as well as the prerequisite attributes systematically. With this approach, each attribute can be evaluated and the source of the group difference isolated.

*Step 2: Estimate the probabilities for the studied attribute.* Once the studied attribute has been specified in the ADF hypothesis, the probability that examinees in the reference and focal group possess this studied attribute must be estimated using the neural network. We call this the *studied* neural network. The input units of the studied neural network are examinees' responses to items that measure the studied attribute as well as all prerequisite attributes because, with the AHM, items that measure the studied attribute also measure the prerequisite attributes due to the dependencies specified in the hierarchy. Note, however, that only items that measure the studied attribute and its prerequisite attributes are included in the input. The output unit is the probability that an examinee possesses the studied attribute. The exemplars used to train the studied neural network are the expected response vectors whereas the target output is the associated attribute patterns. When the error term of the neural network reaches an acceptable level, the hidden and output weights are considered to be estimated adequately. The sum of squared errors between the target output and

the obtained output serves as the loss function to be minimized in calculating the weights. Using this approach, the relationship between the input units and the output unit is established.

Take, for instance, attribute A2 as the studied attribute in Figure 1. According to the first and the second columns of the  $Q_r$  matrix presented in matrix 3, item 1 is solely measuring attribute A1 whereas item 2 is measuring both attributes A1 and A2. These two items are used in estimating the probability that examinees possess attribute A2. Item 3, on the other hand, as shown in the third column of the  $Q_r$  matrix, is measuring attribute A3 in addition to attributes A1 and A2. Therefore, examinees' responses to item 3 would not be used as an input in the studied neural network. Using the same logic, items 5, 6, 8, 9, 11, 12, 14, and 15 are not used to estimate the probability of examinees' mastery of attribute A2 because these items all measure at least one attribute in addition to attributes A1 and A2. This step is critical because it provides a way of isolating the studied attribute A2 relative to the matching attributes and, thus, leads to a statistical outcome in step 4 that is easily interpreted. The output unit of the studied neural network is the probability that an examinee possesses attribute A2. Each exemplar used to train the neural network should contain the responses to items 1 and 2 as the input of the neural network, and the corresponding probability for the mastery of attribute A2 as the output. The first two columns of the expected response matrix presented in matrix 4 contain the expected responses to items 1 and 2. The second column of the attribute pattern matrix presented in matrix 5 holds the mastery level of attribute A2 associated with each expected response pattern. These vectors are used to specify the exemplars (Table 1), which, in turn, are presented to the neural network to estimate the hidden and output weights. For each

TABLE 1  
*Exemplars for the Studied Neural Network in Figure 1*

Exemplars	Input Units		Output Unit
	Item 1	Item 2	A2
1	0	0	0
2	1	0	0
3	1	1	1
4	1	1	1
5	1	0	0
6	1	1	1
7	1	1	1
8	1	0	0
9	1	1	1
10	1	1	1
11	1	0	0
12	1	1	1
13	1	1	1
14	1	0	0
15	1	1	1
16	1	1	1

examinee, the probability of mastering attribute A2 can be calculated by placing his or her responses to items 1 and 2 into the studied neural network.

*Step 3: Defining the matching attributes.* In step 3, examinees in the reference and focal group are matched on all attributes independent of the studied attribute (i.e., on all attributes, other than those that require A2). This is a critical decision, as we noted in step 2, because it ensures that group differences on the matching attributes are controlled before the two groups are compared on the studied attribute. To control group difference on the matching attributes, each examinee is classified into one of the *matching attribute patterns*, thereby specifying the examinee's mastery level on the matching attributes using the neural network. We call this the *matching* neural network. The input units of the matching neural network are the examinees' responses to items that measure at least one of the matching attributes, but not the studied attribute. The output units are the probabilities that an examinee possesses each of the matching attributes. The exemplars used to train the matching neural network are obtained from the expected response vectors and the associated attribute patterns. By presenting each exemplar to the network repeatedly until the specified error criterion is reached, the hidden and output weights for the matching neural network can be estimated.

For the attribute hierarchy shown in Figure 1, if attribute A2 is the studied attribute, then the matching attributes include A4, A5, and A6. According to the  $Q_r$  matrix, items 4, 7, 10, and 13 all measure at least one of the matching attributes, but not A2. These items are used as the input units in the matching neural network to estimate examinees' probability vector for the matching attributes. The output units of the matching neural network are the probabilities that an examinee possesses attributes A4, A5, and A6, respectively. Each exemplar used to train the neural network should contain the responses for items 4, 7, 10, and 13, as well as the probabilities for the mastery of attribute A4, A5, and A6. Therefore, columns 4, 7, 10, and 13 of the expected response matrix and columns 4, 5, and 6 of the attribute pattern matrix are used to specify the exemplars used to estimate the hidden and output weights of the matching neural network. These exemplars are shown in Table 2. Once the hidden and output weights are estimated, each examinee's probability vector on the matching attributes can be calculated by placing his or her responses for items 4, 7, 10, and 13 into the matching neural network.

The output of the matching neural network, at this stage in the analysis, is a vector of probabilities indicating whether the examinees possess the matching attributes, ranging from 0 to 1. In order to classify each examinee into one of the expected matching attribute patterns of 0s and 1s, a least square residual method is used. With this method, examinees' probability vectors on the matching attributes are compared against each of the expected matching attribute patterns using

$$Res_{jm}^2 = \sum_{k=1}^{K'} (p_{jk} - q_{mk})^2,$$

where  $Res_{jm}^2$  is the sum of squared residual of comparing examinee  $j$ 's probability vector against matching attribute pattern  $m$ ,  $K'$  is the number of matching attributes,

TABLE 2  
*Exemplars for the Matching Neural Network in Figure 1*

Exemplars	Input Units				Output Units		
	Item 4	Item 7	Item 10	Item 13	A4	A5	A6
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	1	0	0	0	1	0	0
6	1	0	0	0	1	0	0
7	1	0	0	0	1	0	0
8	1	1	0	0	1	1	0
9	1	1	0	0	1	1	0
10	1	1	0	0	1	1	0
11	1	0	1	0	1	0	1
12	1	0	1	0	1	0	1
13	1	0	1	0	1	0	1
14	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1
16	1	1	1	1	1	1	1

$p_{jk}$  is examinee  $j$ 's probability of mastering attribute  $k$  estimated using the neural network, and  $q_{mk}$  is the mastery level of attribute  $k$  (1 or 0) in expected matching attribute pattern  $m$  derived from the attribute hierarchy. If expected matching attribute pattern  $m$  is the examinee's true pattern, then we expect that the difference between  $p_{jk}$  and  $q_{mk}$  should be close to 0 and, consequently,  $Res_{jm}^2$  should be very small. When a least square residual is obtained, the examinees are classified into one of the corresponding expected matching attribute patterns.

*Step 4: Testing for ADF.* To identify the magnitude and direction of the group difference on the studied attribute, the Simultaneous Item Bias (SIB) test statistic (Shealy & Stout, 1993) was adapted for the ADF framework to statistically evaluate the nature of the attribute-based group differences. The SIB test statistic is used to evaluate group differences on the studied item by aligning examinees at each score level on the matching subtest and then weighting these differences by the proportions of examinees in each group at each matching subtest level. The ADF test statistic is used in a similar manner: It evaluates group differences on the studied attribute probabilities by aligning examinees with the same matching attribute pattern and then weighting these differences by the proportion of examinees with the same matching attribute pattern. The derived formula for calculating the group difference on attribute  $k$  can be calculated by

$$\hat{\beta}_{ADF_k} = \sum_{m=1}^M \hat{f}_m (\bar{p}_{Rm} - \bar{p}_{Fm}),$$

where  $M$  is the total number of matching attribute patterns,  $\hat{f}_m$  is the proportion of examinees (including reference and focal group examinees) that are classified into matching attribute pattern  $m$ ,  $\bar{p}_{Rm}$  is the mean of the studied attribute probabilities for reference group examinees who are classified into matching attribute pattern  $m$ , and  $\bar{p}_{Fm}$  is the respective value for focal group examinees who are classified into matching attribute pattern  $m$ . The standard error of  $\hat{\beta}_{ADF_k}$  is given as

$$\hat{\sigma}(\hat{\beta}_{ADF_k}) = \left( \sum_{m=1}^M \hat{f}_m^2 \left( \frac{\hat{\sigma}_{Rm}^2}{n_{Rm}} - \frac{\hat{\sigma}_{Fm}^2}{n_{Fm}} \right) \right)^{1/2},$$

where  $\hat{\sigma}_{Rm}^2$  is the estimated variance of the studied attribute probabilities for reference group examinees who are classified as having matching attribute pattern  $m$ ,  $\hat{\sigma}_{Fm}^2$  is the respective value for focal group examinees who are classified as having matching attribute pattern  $m$ , and  $n_{Rm}$  and  $n_{Fm}$  are the number of examinees who are classified as having matching attribute pattern  $m$  in the reference and focal group, respectively.

Once the group difference on attribute  $k$  and the standard error are calculated, the significance of the  $\hat{\beta}_{ADF_k}$  can be evaluated using the test statistic,  $\frac{\hat{\beta}_{ADF_k}}{\hat{\sigma}(\hat{\beta}_{ADF_k})}$ . Shealy and Stout (1993) demonstrated that SIB has an approximate normal distribution with mean 0 and variance 1 under the null hypothesis. The null hypothesis is rejected if SIB exceeds the 100  $(1 - \alpha/2)$  percentile point from the normal distribution using a nondirectional hypothesis test. The  $\hat{\beta}_{ADF_k}$  index can be interpreted in a similar way. A value greater than 1.96 suggests that the studied attribute favors the reference group while a value smaller than  $-1.96$  indicates the studied attribute favors the focal group.

### Using the AHM to Study Differential Gender Performance on the SAT

Two examples are provided to illustrate how the AHM can be used to study differential gender performance using student response data from a large operational testing program. The examples are based on the observed response data from a random sample of 3016 students (1508 females and 1508 males) who wrote the algebra items on the March 2005 administration of the SAT. The Mathematics section contains items in the content areas of Number and Operations; Algebra I, II, and Functions; Geometry; and Statistics, Probability, and Data Analysis. For our analysis, only a subset of items in Algebra I and II were evaluated.

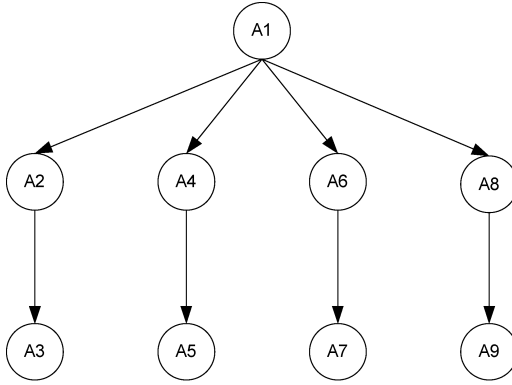
#### *Preliminary Considerations—Specifying and Validating the Cognitive Model of Algebra Performance*

Cognitive models guide test score inferences from the AHM because they are specified at a small grain size and they magnify the cognitive processes that underlie performance. Ideally, a theory of task performance would direct the development of the cognitive model. But, in the absence of theory, a cognitive model must still be specified to create the attribute hierarchy. Another starting point is to develop a cognitive model from a task analysis of the items in the domain when a theory or

model is unavailable and then validate the model using a sample of examinees from the intended target population. Recently, Gierl, Wang, and Zhou (2007) conducted a task analysis of the SAT Algebra I and II items by solving each test item and attempted to identify the mathematical concepts, operations, procedures, and strategies used to solve each problem. They produced a hierarchy, shown in Figure 2a, which serves as a cognitive model for skills in algebra, including areas such as exponents, geometric series, equation solution, and function graph reading. In this hierarchy, attribute A1 is the prerequisite attribute, which includes the basic language skills enabling examinees to understand the test item and basic mathematical knowledge and skills enabling examinees to use operations such as addition, subtraction, multiplication, and division. This hierarchy has four branches: attributes A1, A2, and A3; attributes A1, A4, and A5; attributes A1, A6, and A7; and attributes A1, A8, and A9. These four branches are independent from one another except that they all require the prerequisite attribute A1.

The first branch deals with basic exponential operations. Attribute A2 includes the basic knowledge of exponential and power addition operations. For example, examinees must apply the rule “For any real numbers  $p$  and  $q$ ,  $x^p \cdot x^q = x^{p+q}$ .” In addition to the skills in A2, attribute A3 requires the mastery of a more difficult exponential rule, “For any real numbers  $p$  and  $q$ ,  $(x^p)^q = x^{p \cdot q}$ ,” which requires knowledge of power multiplication and flexible application of multiple rules in an exponential operation. The second branch deals with geometric series. Attribute A4 requires knowledge about geometric series (e.g., the nature of the between-term ratio) and/or the consecutive numerical computation (e.g., multiplication and division). In addition to understanding a geometric series conceptually (i.e., attribute A4), attribute A5 requires a more in-depth understanding of the property of geometric series where the examinee must apply abstract operations to tasks involving geometric series. The third branch deals with equation solutions. Attribute A6 requires the basic mathematical skills in solving for a linear equation (e.g., subtraction or division on both sides of the equation). In addition to the knowledge and skills required in A1, this attribute requires the management of the basic mathematical skills in A1 on both sides of a linear equation. Attribute A7 requires the solution for a quadratic equation, which typically includes the skills required for solving linear equations. Moreover, in solving quadratic equations, additional skills, such as factoring, are needed. The fourth branch deals with functional graph reading. Attribute A8 requires the examinee to map a graph of a familiar function (e.g., a parabola) with its corresponding function. Attribute A9 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 also be involved. Therefore, attribute A8 is considered as a prerequisite for attribute A9.

Once the cognitive model was specified, Gierl, Wang, & Zhou (2007) identified items on the SAT that could measure the attributes. Each attribute was associated with, at least, one test item, except attribute A1. Attribute A1 is considered a basic prerequisite skill that is required to solve all SAT algebra items. Hence, this attribute represents the culmination of many basic mathematical skills and operations that examinees must possess to solve problems in algebra



a. Attribute Hierarchy

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

b. Adjacency Matrix

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

c.  $Q_r$  Matrix

FIGURE 2. The attribute hierarchy for a cognitive model of task performance in algebra on the SAT along with the adjacency and  $Q_r$  matrices.

at the high school level. To ensure that the items measured the appropriate attributes and that the hierarchical ordering among the attributes approximated the cognitive model of algebra performance used by SAT examinees, a sample of 21 Grade 11 Juniors were asked to think aloud as they solved each item on the test. The protocol analyses revealed that the cognitive model created by the task analysis provided excellent fit to the cognitive model produced from the student reports (see Gierl, Leighton, Wang, Zhou, Gokiert, & Tan, 2007 for details). The cognitive model also provided strong fit to the examinees' observed response data, as measured by the hierarchy consistency index ( $HCI_j$ ; Cui, Leighton, Gierl, & Hunka, 2006). The  $HCI_j$  evaluates the degree to which an observed examinee response pattern is consistent with the attribute hierarchy. The index ranges from  $-1$  to  $+1$ . A value of  $-1$  indicates a complete misfit where the examinee correctly answers one item but fails to answer any item that requires the subset of attributes measured by the correctly answered item, whereas a value of  $+1$  indicates the examinee's observed response pattern matches the hierarchy (see Cui et al., 2006 for a complete description of this person-fit index). The  $HCI_j$  for the algebra model in Figure 2a was high for our student sample, at .80, indicating strong model-data fit.

At this point, we must acknowledge that the AE approach used in this study—where items are fit to the cognitive model—is not ideal because the items were drawn from an existing test. The use of existing items with new models may yield inadequate results because the items were not initially created from an explicit cognitive model and, thus, may poorly represent items that would have been generated from the hierarchy. In the current study, we, in effect, “retrofit” the items to the model because test development had already occurred and we only had access to existing items. Fortunately, retrofitting procedures can yield satisfactory results when additional validation studies are undertaken—for instance, the use of verbal reports (Gierl, Leighton, et al; 2007)—to establish a link between the model and the items (see also Haynie, Haertel, Lash, Quellmalz, & DeBarger, 2006). Nevertheless, a more efficient approach is to specify the model first, and then create the items based on the cognitive structure using, for example, the item-by-attribute combinations outlined in the  $Q_r$  matrix. Although this important shortcoming exists in our illustrative example, the cognitive model identified and validated in this example allows us to present the four key steps associated with studying ADF using the AHM.

*Step 1: Specify the ADF hypotheses.* Two ADF hypotheses are tested in the current example to evaluate the examinees' understanding of equation solutions (i.e., the third branch of the hierarchy). Hypothesis 1 focuses on attribute A6. This attribute was selected because it serves as an important prerequisite for attribute A7. Attribute A6 includes the skills required to solve a linear equation. Hypotheses 2, in turn, probes attribute A7, which measures the skills required to solve a quadratic equation. Attributes A6 and A7 produce an important dependency with one another because the skills required to solve a quadratic equation typically include the skills required for solving a linear equation. Also, to solve quadratic equations, additional skills, such as factoring, are needed.

*Step 2: Estimate the probabilities for the studied attribute.* Once the studied attribute has been specified in each ADF hypothesis, the probability that examinees

in the focal and reference group possess this studied attribute must be estimated using the *studied* neural network. Recall, the input unit of the studied neural network is examinees' responses to items that measure the studied attribute (including all prerequisite attributes) whereas the output unit is the probability that an examinee possesses the studied attribute. The exemplars used to train the studied neural network are the expected response vectors. The target output is the associated attribute patterns.

To investigate gender difference on attribute A6, examinees' probabilities of possessing A6 are estimated using the studied neural network. According to the  $Q_r$  matrix presented in Figure 2c, item 5 is measuring attribute A6 and its prerequisite attribute, A1 (see column 5). Therefore, the input unit of the studied neural network is examinees' responses to item 5. The output unit of the studied neural network is the probability that an examinee possessed attribute A6. The exemplar used to train the studied neural network is specified using the column vector associated with item 5 in the expected response pattern matrix, while the output target is the corresponding vector for attribute A6 in the attribute pattern matrix. Exemplars are then presented to the neural network to estimate the hidden and output weights. Once these weights are estimated, each examinee's probability of mastering attribute A6 is calculated by placing his or her response to item 5 into the studied neural network.

Similarly, for attribute A7, the  $Q_r$  matrix in Figure 2c reveals that item 6 (i.e., column 6) is measuring attribute A7 as well as its prerequisite attributes, attributes A1 and A6. Thus, the input units of the studied neural network are the examinees' responses for items 5 and 6 whereas the output unit of the studied neural network is the probability that an examinee possesses attribute A7. The hidden and output weights for the studied neural network are estimated using exemplars specified by the columns associated with items 5 and 6 of the expected response pattern matrix, and the corresponding column for attribute A7, of the attribute pattern matrix. Using the studied neural network, each examinee's probability on attribute A7 can be calculated.

*Step 3: Defining the matching attributes.* Next, examinees in the focal and the reference group are matched on all attributes independent of the studied attribute by classifying examinees into one of the matching attribute patterns using the matching neural network. The input units of the matching neural network are examinees' responses to items that measure at least one of the matching attributes, but not the studied attribute, whereas the output units are the probabilities that an examinee possesses each of the matching attributes. The exemplars used to train the matching neural network are obtained from the expected response vectors.

For attribute A6, the matching attributes consist of attributes A2 to A5 and A8 to A9. The input units of the matching neural network are an examinee's responses for items 1, 2, 3, 4, 7, and 8 while the output units are the examinee's probabilities of possessing each of the matching attributes (see  $Q_r$  matrix in Figure 2c). The relationship between the input and output units are established by training the matching neural network using exemplars obtained from the expected response patterns. Then, each examinee's probability vector on the matching attributes is calculated by placing his or her responses for items 1, 2, 3, 4, 7, and 8 into the matching

TABLE 3

*Matching Attribute Patterns and Mean Probability Values for Gender ADF on Attribute A6 in Figure 2*

Matching Attribute Pattern	Focal Group		Reference Group	
	Frequency	Mean Probability	Frequency	Mean Probability
000000	42	.71	38	.76
000010	34	.79	25	.72
000011	12	.83	16	.75
001000	32	.78	16	.93
001010	34	.90	29	.89
001011	13	.84	12	.91
001100	15	.86	10	.99
001110	10	.89	29	.79
001111	10	.89	19	.94
100000	66	.74	45	.82
100010	116	.91	67	.83
100011	27	.88	28	.78
101000	79	.88	59	.88
101010	168	.94	151	.94
101011	62	.96	63	.94
101100	33	.87	36	.88
101110	69	.95	98	.96
101111	73	.96	84	.97
110000	23	.95	10	.70
110010	71	.94	57	.92
110011	15	.73	13	.92
111000	13	.84	14	.92
111010	91	.97	76	.99
111011	75	.95	79	.97
111100	20	.84	15	.93
111110	105	.96	125	.98
111111	200	.98	294	.99

neural network. In order to classify each obtained probability vector into one of the expected matching attribute patterns derived from the attribute hierarchy, the least square residual method is used. As shown in Table 3, there are 27 expected matching attribute patterns. Table 3 also contains a summary of the number of examinees who are classified into each expected matching attribute pattern and the mean of the probabilities on the studied attribute for examinees in the focal and reference group, respectively. For example, the first row of Table 3 shows that the number of examinees who are classified as not possessing any of the matching attributes in the focal and reference group are 42 and 38, respectively. The mean probability for attribute A6 with the matching pattern (000000) is .71 and .76 for the examinees in the focal and reference group, respectively.

Similarly, for attributes A7, the matching attributes include attributes A2 to A5 and A8 to A9. For the matching neural network, the input units are examinees' responses for items 1, 2, 3, 4, 7, and 8 whereas the output units are the examinees' probabilities

TABLE 4  
*Matching Attribute Patterns and Mean Probability Values for Gender ADF on Attribute A7 in Figure 2*

Matching Attribute Pattern	Focal Group		Reference Group	
	Frequency	Mean Probability	Frequency	Mean Probability
000000	42	.10	38	.11
000010	34	.15	25	.12
000011	12	.25	16	.19
001000	32	.13	16	.19
001010	34	.26	29	.28
001011	13	.38	12	.09
001100	15	.20	10	.20
001110	10	.20	29	.41
001111	10	.30	19	.32
100000	66	.09	45	.22
100010	116	.22	67	.21
100011	27	.15	28	.39
101000	79	.29	59	.32
101010	168	.29	151	.45
101011	62	.46	63	.46
101100	33	.36	36	.33
101110	69	.46	98	.51
101111	73	.60	84	.63
110000	23	.13	10	.20
110010	71	.30	57	.40
110011	15	.33	13	.54
111000	13	.46	14	.50
111010	91	.43	76	.48
111011	75	.56	79	.55
111100	20	.55	15	.40
111110	105	.53	125	.63
111111	200	.82	294	.79

of possessing each of the matching attributes. Once the hidden and output weights of the matching neural network are estimated using exemplars from the expected response patterns, examinees' probability vectors on the matching attributes are calculated. Using the least square residual method, examinees in the focal and reference group are assigned to the expected matching attribute patterns. For attribute A7, the matching attribute patterns and mean probability values for examinees in the focal and reference group are shown in Table 4.

*Step 4: Testing for ADF.* To evaluate the magnitude and direction of the group difference on the studied attribute,  $\hat{\beta}_{ADF_k}$  is estimated and tested statistically. In our analyses, females comprise the focal group whereas males form the reference group. Therefore,  $\hat{\beta}_{ADF_k}$  value greater than 1.96 indicated the studied attribute favors males while a value smaller than  $-1.96$  indicates the studied attribute favors females.

TABLE 5  
*Gender ADF for Hypotheses 1 and 2*

Hypothesis	Studied Attribute (Prerequisite Attribute)	$\hat{\beta}_{ADF_k}$	$\hat{\sigma}(\hat{\beta}_{ADF_k})$	$\hat{\beta}_{ADF_k}/\hat{\sigma}(\hat{\beta}_{ADF_k})$
1	A6 (A1)	.004	.009	.44
2	A7 (A1, A6)	.042	.016	2.63*

\* $p < .05$ .

The results for ADF hypothesis 1 and 2 are shown in Table 5. Hypothesis 1, which evaluated attribute A6, measures the skills required to solve a linear equation. This attribute showed no gender difference. Hypothesis 2, focused on attribute A7, evaluates the examinees' skills required to solve a quadratic equation. The skills in attribute A7 include attribute A6. Hence, a dependency exists among these attributes. However, A7 includes some additional skills, such as factoring. Attribute A7 favored males. This outcome suggests that while there are no gender differences on A6 skills, males have stronger A7 skills compared to females, at least in our illustrative analysis. Or, said differently, our example allows us to postulate that males and females have comparable basic mathematical skills (i.e., attribute A1) which can be used to manage the operations required to solve linear equations (i.e., attribute A6), but differ in the management of those operations (for example, factoring) required to solve quadratic equations (i.e., attribute A7).

### Summary and Discussion

AE with the AHM relies on two stages. In the model development stage, principled test design procedures are used to identify items that systematically measure each component in the cognitive model. In the pattern recognition stage, the functional relationship between the examinees' expected response patterns and attribute patterns is established so that the attribute probabilities for the examinees' observed response patterns can be estimated. These probabilities are used to establish the studied and matching attribute patterns so that group differences can be evaluated. These methods, in turn, were applied to the study of differential group performance in an attempt to bridge the gap between the substantive and statistical steps commonly applied in DIF detection studies so that group differences can be more easily identified statistically and interpreted substantively. We described and illustrated a four-step procedure required to make inferences about differential performance on tests using the AHM.

### *Directions for Future Research*

The study of ADF using the AHM raises a number of research issues. We outline three key issues that require additional investigation: specifying the ADF hypotheses, establishing the studied and matching attribute patterns, and sample size requirements.

*Specifying ADF hypotheses.* An ADF analysis begins by specifying the hypotheses about the nature of group differences. This initial specification is required because

ADF is guided by the confirmatory logic of hypothesis testing where the cognitive model must first be identified substantively before attribute-level group differences are identified statistically. The approach is advantageous because it produces an interpretable statistical result, given that the substantive explanation is used to generate hypotheses and then a statistical analysis is used to test the hypotheses. Consequently, each analysis provides a test of the proposed hypotheses. But this important advantage has a catch: The researcher or practitioner must have an articulate cognitive model to guide the analysis. In practice, these models are difficult to specify and time consuming to create (Gierl & Leighton, 2007). Because of these substantive requirements, the study of ADF may be slow to develop in testing situations where little is known about the cognitive skills that underlie examinee performance.

Identifying the cognitive model is one part in the substantive step; specifying ADF hypotheses is another part. These hypotheses indicate whether a difference exists between the two groups of interest on an attribute or a combination of attributes in the cognitive model. ADF hypotheses—discerned from existing theory, prior empirical evidence, existing literature, and/or task analyses and content reviews—must guide the study of attribute-level group differences. Unfortunately, these hypotheses are often difficult to specify because our understanding of the cognitive factors that produce differences among diverse groups in a range of testing situations is limited. Hence, more research is needed to identify cognitive models and to evaluate which skill or combination of skills in these models could produce group differences on tests.

*Establishing the studied and matching attribute patterns.* An ADF analysis requires that the researcher or practitioner identify both the studied and matching attributes. In a DIF analysis, the studied and matching subtests serve as meaningful comparative criteria in the study of group differences. With an ADF analysis, the unit of analysis shifts from item to attribute. The studied attribute is estimated using the studied neural network where the input units are examinees' responses to items that measure the studied attribute (including all prerequisite attributes) while the output unit is the probability that an examinee possesses the studied attribute. The matching attributes are estimated using the matching neural network where the inputs are examinees' responses to items that measure the matching attributes while the output units are the probabilities that an examinee possesses each of the matching attributes. Because of the hierarchical structure of the cognitive model, the studied and matching attributes form independent units. The studied attribute includes the attribute in question as well as all prerequisite attributes that could affect the studied attribute. The matching attributes include only those attributes that have no relationship to the studied attributes. Hence, the matching attributes are purified *by design* in an ADF analysis because they are independent of any effect that could stem from the studied attribute.

Unfortunately, the dependencies created by the hierarchy can also constrain an ADF analysis. Take, for example, the hypothetical hierarchy in Figure 3. This hierarchy contains two independent branches, which share a common prerequisite attribute A1. The first branch includes two attributes, A2 and A3, and the second branch includes a self-contained sub-hierarchy that includes attributes A4 through A9. Three

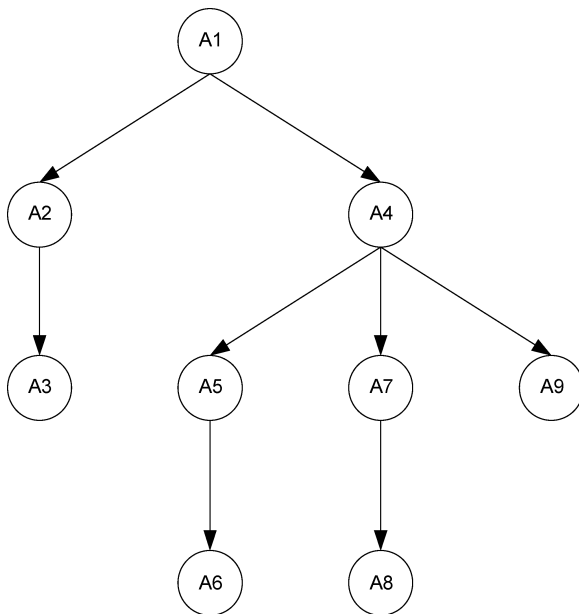


FIGURE 3. A hypothetical hierarchy that would only have a small number of matching attributes.

independent branches form the sub-hierarchy: attributes A4, A5, A6; attributes A4, A7, A8; and attributes A4, A9. One could imagine a situation where a researcher generates hypotheses about group differences that reside in one or more of the attributes in the second branch of the model. However, with our proposed approach, these hypotheses may be difficult to evaluate, particularly if the hypotheses include attribute A4. Because A4 is a key prerequisite for the sub-hierarchies that include A5 and A6, A7 and A8, and A9, any hypotheses about group differences that include A4 mean that the matching attributes are limited to attributes A2 and A3. Simply put, the structure of the cognitive model may dictate the types of ADF analyses that can be conducted. And, in some cases, it may be difficult or even impossible to test specific ADF hypotheses because the structure of the model may yield a studied or, in the case of the Figure 3 example, a matching pattern with an inadequate number of attributes. This problem would be overcome by including a large number of items in the model to measure each attribute. For example, if the model in Figure 3 contained five items per attribute, then ADF hypotheses in the second branch could be evaluated against a three-attribute matching subtest as measured by 15 items. But other matching methods—for example, strategies based on thin versus thick matching (Donoghue & Allen, 1993) at the *attribute level*—should also be investigated.

*Sample size requirements.* Sample size is another consideration when conducting an ADF analysis. Sample size often serves as a key manipulated variable when evaluating the analytic properties of DIF methods because it can impact detection rates. More important to practitioners, perhaps, is the minimum sample size that can

be used for reliably identifying items that produce group differences (e.g., Muniz, Hambleton, & Xingand, 2001; Roussos & Stout, 1996). The sample size used in our SAT example was large, at more than 1,500 examinees per group. The consequences of using a smaller sample for ADF analyses with the AHM are unknown. Therefore, additional studies are needed to evaluate the effect of sample size on ADF detection.

### Acknowledgments

We would like to thank Dr. Steve Hunka for his comments on our manuscript. The research reported in this study was conducted with funds provided to the first author by the College Board and by the Social Sciences and Humanities Research Council of Canada. However, the authors are solely responsible for the methods, procedures, and interpretations expressed in this study, as our views do not necessarily reflect those of the College Board or Social Sciences and Humanities Research Council of Canada. A tutorial, programmed in Mathematica 5.0, to accompany this study is available at the CRAME website under the “Research” page at <http://www.education.valberta.ca/educ/psych/crame/>.

### Note

<sup>1</sup>Again, a row of 0s is added to the examinee attribute pattern matrix because it is possible that an examinee may not possess any of the attributes. This row of 0s corresponds to the first row in the expected response matrix given in matrix 4. In the reduced Q matrix in matrix 3, however, only items that measure at least one attribute are included—hence, a row of 0s is *not* added to this matrix.

### References

- Angoff, W. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Lawrence Erlbaum.
- Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.
- Cui, Y., Leighton, J. P., Gierl, M. J., & Hunka, S. (2006). *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.
- Donoghue, J. R., & Allen, N. L. (1993). Thin versus thick matching in the Mantel-Haenszel procedures for detecting DIF. *Journal of Educational Statistics, 18*, 131–154.
- Engelhard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3*, 347–360.
- Gierl, M. J. (2005). Using a dimensionality-based DIF analysis paradigm to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice, 24*, 3–14.
- Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule space model and attribute hierarchy method. *Journal of Educational Measurement, 44*, 325–340.

- Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the DIF analysis paradigm. *Journal of Educational Measurement, 40*, 281–306.
- 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.
- Gierl, M. J., & Leighton, J. P. (2007). Part C: Future challenges in psychometrics: Linking cognitively-based models and psychometric methods. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics*, Vol. 26 (pp. 1103–1106). Amsterdam: Elsevier North Holland.
- 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 in education: Theory and applications* (pp. 242–274). Cambridge, UK: Cambridge University Press.
- Gierl, M. J., Leighton, J. P., Wang, C., Zhou, J., Gokiart, R., & Tan, A. (2007). *Developing and evaluating cognitive models of algebra performance on the SAT<sup>®</sup>* (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<sup>®</sup>* (Research Report). New York: The College Board.
- Gierl, M. J., & Zhou, J. (in press). Computer adaptive-attribute testing: A new approach to cognitive diagnostic assessment. To appear in the Special Issue of *Zeitschrift für Psychologie—Journal of Psychology*, Adaptive Models of Psychological Testing, Wim J. van der Linden (Guest Editor).
- Haynie, K. C., Haertel, G. D., Lash, A. A., Quellmalz, E. S., & DeBarger, A. H. (2006). *Reverse engineering the NAER floating pencil task using the PADIF design system* (PADI Technical Report 16). Menlo Park, CA: SRI International.
- Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*, 3–16.
- Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement, 41*, 205–236.
- Luecht, R. M. (2006a, May). *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. (2006b, September). *Assessment engineering: An emerging discipline*. Paper presented in the Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB, Canada.
- Luecht, R. M., Gierl, M. J., Tan, X., & Huff, K. (2006, April). *Scalability and the development of useful diagnostic scales*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- McClelland, J. L. (1998). Connectionist models and Bayesian inference. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 21–53). Oxford: Oxford University Press.
- Muniz, J., Hambleton, R. K., & Xingand, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*, 115–136.
- O'Neill, K. A., & McPeck, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255–276). Hillsdale, NJ: Lawrence Erlbaum.

- Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215–230.
- 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.
- Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159–194.
- Standards for Educational and Psychological Testing* (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.
- Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*, 485–518.
- Sudweeks, R. R., & Tolman, R. R. (1993). Empirical versus subjective procedures for identifying gender differences in science test items. *Journal of Research in Science Teaching, 30*, 3–19.

### Authors

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. 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; and psychometric methods for evaluating test translation and adaptation.

YINGGAN ZHENG is Statistical Analyst, Canadian VIGOUR Centre, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5; yinggan.zheng@ualberta.ca. His primary research interests include examining the validity and reliability of different measurement tools, analyzing the economic impact on health and livelihoods, and epidemiology.

YING CUI is Assistant Professor of Educational Psychology, Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5; yc@ualberta.ca. Her primary research interests include investigating cognitive factors underlying student test performance, using cognitive models to interpret student test performance, and evaluating the construct validity of achievement tests.