

The Attribute Hierarchy Model for Cognitive Assessment*

Jacqueline P. Leighton
Mark J. Gierl
Stephen M. Hunka

Centre for Research in Applied Measurement and Evaluation
University of Alberta

Paper Presented at the Annual Meeting of the National Council on
Measurement in Education (NCME)

New Orleans, Louisiana

April 2-4, 2002

*This paper can also be downloaded from the CRAME website at <http://www.education.ualberta.ca/educ/psych/crame/>

The Attribute Hierarchy Model for Cognitive Assessment

A. INTRODUCTION

Most educational and psychological tests require examinees to engage in some form of cognitive problem solving. On these tests, the cognitive processes, strategies, and knowledge used by examinees to solve problems should be considered when attempting to validate the inferences made about these examinees (Embretson, 1983, 1994, 1998; Messick, 1989; Snow & Lohman, 1989). The important role that cognitive theory could play in educational and psychology testing is apparent to many measurement specialists (e. g., Embretson, 1985; Frederiksen, Glaser, Lesgold, & Shafto, 1990; Frederiksen, Mislevy, & Bejar, 1993; Nichols, Chipman, & Brennan, 1995; Snow & Lohman, 1989). For example, cognitive analyses could allow researchers to experiment with the internal characteristics of the test, evaluate the assumptions of existing psychometric models, create new psychometric models, and explain the psychology that underlies test performance (Embretson, 1983; Gierl, Leighton, & Hunka, 2000; Hattie, Jaeger, & Bond, 1999; Mislevy, 1996; Nichols, 1994; Nichols & Sugrue, 1999; Royer, Cisero, & Carlo, 1993; Snow & Lohman, 1989).

While cognitive theory can inform psychometric practice in many ways, Embretson (1983), in particular, suggests that cognitive theory can enhance psychometric practice by illuminating the construct representation of a test. The construct or latent trait that underlies test performance is represented by the cognitive processes, strategies, and knowledge used by an examinee to respond to a set of test items. Once these cognitive requirements are sufficiently described, they can be assembled into cognitive models that are then used to develop items that elicit specific knowledge structures and cognitive processes. Test scores anchored to a cognitive model should be more interpretable and, perhaps, more meaningful to a diverse group of users because performance is described using a specific set of cognitive skills in a well defined content area.

Unfortunately, the impact of cognitive theory on test design has been minimal (Embretson, 1998; National Research Council, 2001; Pellegrino, 1988; Pellegrino, Baxter, & Glaser, 1999). Embretson (1994) believes that test developers have been slow to integrate cognitive theory into

psychometric practice because they lack a framework for using cognitive theory to develop tests. Embretson (1998) also argues that cognitive theory is not likely to impact testing practice until its role can be clearly established in test design. To try to overcome this impasse, Embretson (1995a) developed the cognitive design system (CDS). The CDS is a framework where test design and examinee performance are explicitly linked to cognitive theory (also see Embretson, 1994, 1998, 1999). The goal of such a link is to make both the test score and the construct underlying the score interpretable using cognitive theory. Embretson (1999) recently described the CDS as a three-stage process. In the first stage, the goals of measurement are described. In the second stage, construct representation is established. In the third stage, nomothetic span research (i.e., correlating the test score with other well-defined measures) is conducted. The CDS has been used to validate a variety of constructs including verbal reasoning (Embretson, Schneider, & Roth, 1985), abstract reasoning (Embretson, 1998), spatial reasoning (Embretson, 1995a), paragraph comprehension (Embretson & Wetzel, 1987), and mathematical problem solving (Embretson, 1995b).

The appeal of the CDS is the explicit link between the cognitive and psychometric properties of test items. This link is typically achieved using cognitive IRT models. Cognitive IRT models are created when mathematical models containing cognitive variables are combined with IRT models containing the examinees' item responses. This modeling approach yields parameters that represent both the cognitive demands of the items and ability levels of the examinees. Some of these models have proven useful in studying the cognitive factors that influence test performance across diverse tasks, content areas, and age levels (e.g., see reviews in Embretson & Reise, 2000; Nichols, Chipman, & Brennan, 1995; Roussos, 1994; van der Linden & Hambleton, 1997).

The purpose of this paper is to introduce and illustrate a new cognitive IRT model called the attribute hierarchy model (AHM). In Section B of the paper we describe the AHM. The attribute hierarchy, once specified, serves as a cognitive model of test performance. We show that once a hierarchy and the associated test items have been constructed, expected response patterns can

be derived. These expected response patterns, in turn, can be used to classify examinees based on their observed response patterns. In Section C of the paper we apply the AHM to the domain of syllogistic reasoning. In Section D of the paper we outline future directions for research.

B. THE ATTRIBUTE HIERARCHY MODEL: AN OVERVIEW

Cognitive Component of the Attribute Hierarchy Model

Identifying Cognitive Attributes

The AHM is based on the assumption that test performance depends on a set of specific skills or competencies called attributes¹. The examinee must possess these attributes to answer the items correctly. Attributes can be viewed as sources of cognitive complexity in test performance (cf. Embretson, 1995). But, more generally, attributes are those basic cognitive skills required to solve test problems correctly. The importance of correctly identifying the attributes cannot be overstated—the first step in making inferences with the AHM depends on accurately identifying the cognitive skills required to solve test problems. The attributes serve as the most important input variable for the AHM because it provides the basis for making inferences about examinees' cognitive skills. One description of an attribute was provided by Leighton, Gierl, and Hunka (1999):

An attribute is a description of the procedural or declarative knowledge needed to perform a task in a specific domain. Although an attribute is not a strategy, attributes do provide the building blocks for strategies. Furthermore, the set of attributes organized into a strategy serves a momentary problem-solving role, but does not necessarily remain grouped as a strategy. Attributes are dynamic entities. They evolve with a student's increasing competency so that a set of attributes at time 1 may no longer function as useful descriptions of behavior at time 2. Finally, the time periods mentioned are developmentally and/or instructionally dependent, meaning that a student progresses from time 1 to time 2 in response to developmental and/or instructional factors. The attributes for a test can be identified using different methods (e.g., expert opinion, task analysis, written responses from students). However, verbal think-aloud protocols should

be included among the methods used to validate the attribute descriptions using both examinees and test items that are comparable to their target populations.

Attributes can be identified and studied using methods from cognitive psychology. For example, item reviews and protocol analysis can be used to study task requirements. Item reviews are often conducted to identify the knowledge and skills required to solve test items by specialists (e.g., test developers) who are familiar with the content area, test development process, and the way students solve problems. Examinees can also be asked to think aloud as they solve selected problems, and protocol analysis (Ericsson & Simon, 1993) can be used to study their problem-solving skills. Protocol analysis is an effective method for identifying the specific knowledge components and cognitive skills elicited by test items and measurement specialists are using these techniques increasingly to study problem solving on tests (e.g., Baxter & Glaser, 1998; Gierl, 1997; Leighton, Rogers, & Maguire, 1999; Hamilton, Nussbaum, & Snow, 1997; Magone, Cai, Silver, & Wang, 1994; Norris, 1990).

Specifying the Attribute Hierarchy to Model Test Performance

Once the attributes are identified, they must be structured within a hierarchy. The hierarchy defines the psychological ordering among the attributes required to solve a test problem. The ordering of the attributes may be derived from empirical considerations (i.e., a series of well-defined, ordered cognitive processes or steps identified via protocol analysis) or theoretical considerations (e.g., a series of developmental sequences suggested by Piaget such as pre-operational, concrete operational, and formal operational). Once specified, the hierarchy containing the attributes serves as the cognitive model of test performance. Consequently, the attribute hierarchy has a foundational role in the AHM because it represents the construct and cognitive processes that underlie test performance.

"Are All Hierarchies Created Equal?": Forms of Hierarchical Structures

Figure 1 contains a range of hierarchies, from structured (i.e., linear) to unstructured. In all hierarchies, attribute A1 (labeled 1 in the Figure) may be considered hypothetical in the sense that it represents all the initial skills that are prerequisite to the attributes that follow (or

alternatively, A1 may be considered a specific attribute). In Figure 1(A) attribute A1 is considered prerequisite to attribute A2; attributes A1 and A2 are prerequisite to attribute A3; attributes A1, A2 and A3 are considered prerequisite to attribute A4. Specifying that attribute A1 is prerequisite to attribute A2 implies that an examinee is not expected to possess attribute A2 unless attribute A1 is also present. In the linear hierarchy the implication is also that if attribute A1 is not present, then all attributes that follow are not expected to be present. If test items are constructed to probe for the attributes in a linear hierarchy, then the expected response pattern of the examinees will be that of a Guttman scale and the total score will relate perfectly to the expected response pattern. The conditions required to obtain this relationship between the expected response pattern and the total score are: (a) the hierarchy is true, (i.e., the hierarchical relationships of attributes is a true model of examinees' cognitive attributes), (b) test items can be written which probe the appropriate attributes, and (c) the examinees respond without error.

Figure 1(D), the unstructured hierarchy, represents the other extreme of possible hierarchical structures. In Figure 1(D), attribute A1 is considered prerequisite for attributes A2 through A6. However, unlike Figure 1(A) where attribute A2 is prerequisite to attribute A3, there is no ordering among attributes A2 through A6 in this hierarchy and there is no unique relationship between the total score and the item response pattern.

Figure 1(B) represents a hierarchy with a convergent branch where two different paths may be traced from A1 to A6. Attribute A2 is prerequisite to A3 and A4, but A3 or A4 are prerequisite to A5. This hierarchy, like Figure 1(A), ends at a single point. This type of hierarchy might be used to describe the cognitive processes leading to a single correct end state such as tasks where the desired outcome is clear (e.g., a multiple-choice test measuring the addition of fractions).

In contrast, Figure 1(C) represents a hierarchy having a divergent branch. This type of hierarchy might be used to describe the cognitive processes leading to an answer consisting of multiple components that can be judged as either correct or incorrect (e.g., a constructed-response item measuring students' knowledge about what social circumstances triggered World War II). This hierarchy might also be used to describe the entire ordering of cognitive processes

required to solve problems successfully in a specific domain. It is important to note that the examples in Figure 1 could be combined to form increasingly complex networks of hierarchies where the complexity varies with the cognitive problem-solving task.

Psychometric Component of the Attribute Hierarchy Model

Formal Representation of a Hierarchy

In order to calculate the expected response patterns for a specific hierarchy, a formal representation of the hierarchy is required². For illustrative purposes the divergent hierarchy of Figure 1(C) is used. The direct relationships among attributes is specified by a binary adjacency matrix (A)³ of order (k, k) where k is the number of attributes. The A matrix for the hierarchy of Figure 1(C) is given below:

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

In the adjacency matrix, a 1 in the position (j, k) indicates that attribute j is directly connected in the form of a prerequisite to attribute k (where j precedes k). For example, the first row indicates that attribute A1 is a prerequisite to attributes A2 and A4. A row of 0s, such as row 3, indicates that attribute A3 is not a prerequisite to any other attributes. Also notice that 1s only appear in the upper triangular portion of the matrix indicating a directional prerequisite relationship (i.e., A1 is prerequisite to A2 and A4, but A2 and A4 are not prerequisite to A1).

To specify the direct and indirect relationships among attributes, a reachability matrix (R) of order (k, k), where k is the number of attributes, is used. The R matrix can be calculated as $R = (A + I)^n$, where n is the integer required for R to reach invariance, $n=1, 2, \dots, k$; A is the adjacency matrix; and I is an identity matrix. Alternatively, R can be formed by a series of Boolean additions of rows of the adjacency matrix. The j^{th} row of the R matrix specifies all the

measure these attributes are specified in the Q_r matrix. In short, cognitive theory has a clearly defined role in test design using the AHM for cognitive assessment.

Generating Expected Response Patterns

Given a hierarchy of attributes, the expected response patterns can be calculated. Associated with these patterns are "expected" examinees. Expected examinees are defined as examinees that invoke attributes consistent with the hierarchy. Moreover, expected examinees do not make "slips" or errors that produce inconsistencies between the observed and expected response pattern (recall, expected response patterns are derived from the attribute hierarchy). For the hierarchy of Figure 1(C) the expected response patterns, total scores, and examinee attributes are shown in Table 1. Row 1 of Table 1 should be interpreted as follows: An examinee who only has attribute A1, [i.e., (10000)] is expected to answer only the first item correctly, producing the expected examinee response pattern (10000000000000). In Table 1, the complete set of expected examinee response patterns forms a matrix. The columns of this matrix may be considered expected item response vectors. For example, the expected item response vector for item 5 is (000011011011011) given by column 5. The expected item response vectors are used to estimate item parameters using a 2-parameter logistic item response theory (IRT) model (discussed in the next section). Also notice that the examinees' total score does not consistently indicate which attributes are present. For example, a score of 2 may be obtained by having attribute patterns (110000) or (100100). If the attribute hierarchy is true, then the only scores that will be observed for the expected examinees are 1, 2, 3, 4, 5, 6, 9, 10, and 15.

Estimating Probabilities of Item Responses

Ideally, the objective of developing a test consistent with a hierarchy is to identify those attributes that are deficient for each examinee. Person-fit indices can be used to evaluate the degree to which an observed response pattern is consistent with the probability of a correct response typically derived from an IRT model. In the AHM, the expected item characteristic curve can be calculated using an IRT model under the assumption that examinees responses are consistent with the attribute hierarchy. For purposes of illustration, we will continue to work with

the attribute hierarchy of Figure 1(C) and will use the two-parameter (2PL) logistic IRT model given by

$$P(u = 1|\Theta) = \frac{1}{1 + e^{-1.7a_i(\Theta - b_i)}},$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, and Θ is the ability parameter. Then, using the 2PL logistic function, the problem becomes one of determining the a- and b-parameters for each item based on the expected item response vectors (given by the columns of the expected response matrix in Table 1).

To illustrate this procedure, a sample of 1000 examinees was generated with the constraint that the total scores associated with each expected examinee response pattern (i.e., the rows of the expected response matrix in Table 1) be approximately normal in distribution. The item parameters were estimated using the expected item response vectors in Table 1 with BILOG 3.11 (Mislevy & Bock, 1990). The default settings in BILOG were used, with the exception of the calibration option that was set to “float” indicating that the means of the priors on the item parameters were estimated using marginal maximum likelihood estimation along with the item parameters, and both the means and the item parameters were updated after each iteration. The estimates are shown in Table 2. A plot of the expected item characteristic curves are shown in Figure 2. Note that only 13 curves are distinguishable because the curves for items (8, 11) and (13, 14) are identical. The expected item characteristic curves display some high slopes indicating that some items are very discriminating. Plots of the expected item and test information functions are shown in Figure 3.

Classification of Observed Response Patterns

Overview. The value of any ability estimate, such as theta or total score, does not indicate, uniquely, those attributes that may be the basis of an examinee’s observed response pattern. It would be more meaningful to the examinees if some indication of the deficient attributes was also reported along with an overall ability score. With this information the examinee and the instructor (e.g., classroom teacher, parent, tutor) could take more specific remedial action. Fundamental to

the classification of examinees based on their item responses is information concerning the probability of each item response and the definition and identification of an atypical set of item responses. When an attribute hierarchy has been defined, from which the Q_r matrix can be derived to guide the construction of test items, the expected response patterns of expected examinees can be defined. As a result, the atypicality of an observed response pattern can be judged relative to the expected response pattern based on the assumption that the attribute hierarchy is true.

By its very definition, an atypical response pattern requires the specification of a typical response pattern. Numerous fit indices have been proposed for the identification of atypical response patterns (e.g., Meijer, 1996; Meijer & Sijtsma, 2001). Most of these indices require the probabilities of the correct responses conditioned on theta (and these probabilities are often produced from an IRT model). However, the indices used to identify "misfitting" item response patterns usually associate the misfit to response behaviors that have little to do with the cognitive skills required to solve test items (e.g., "sleeping behavior", "guessing behavior", cheating, "plodding", alignment errors; see Meijer, 1996, pp. 4-5).

Another approach to the classification of cognitive skills is found in the literature on the rule-space model (Tatsuoka, 1983, 1984, 1996). In rule space, it is assumed that each examinee consistently applies a set of "rules" in answering each item. The rules may be correct or incorrect and, in either case, are represented with a binary response vector. Thus, any rule known to exist on the basis of real data or hypothesized by the test developer through an analysis of the attributes required of an item, correct or incorrect, can be represented by a coordinate in the rule space. In the simplest case, the coordinate can be represented by (Θ, ζ) in which Θ is the ability parameter and ζ is an "atypicality" parameter. Atypical responses in the rule space represent an observed discrepancy in the application of a rule from the coordinate (θ, ζ) associated with the rule. A collection of such atypical patterns is considered to be approximately normal in its distribution. Tatsuoka and Tatsuoka (1989) further indicate that the problem of classifying an examinee's response is solved by existing statistical classification methods and

pattern recognition theory. ζ will have a large numerical value when there are a small number of correct responses for easy items and a large number of correct answers for the difficult items.

Tatsuoka (1996) suggests that the dimensionality of the rule space can be increased to enhance identification of atypical response patterns by calculating values of ζ for subsets of items [e.g., to calculate $(\Theta, \zeta_1, \zeta_2, \zeta_3, \text{etc.})$ where the ζ values are based on different subsets of items].

Applications of the rule-space model to signed addition, addition of fractions, and mathematics items from the Scholastic Assessment Test are illustrated in the research literature (Tatsuoka, 1995, 1996; Tatsuoka & Tatsuoka, 1989). In these examples it is not possible to clearly identify an attribute hierarchy or to associate the administered items to a Q_r matrix derived from the hierarchy. Instead, identification of deficient examinee characteristics are based on a post-hoc analysis of the attributes required to answer each item and the identification of expected response patterns based on the attributes that the examinee likely does not possess. Thus, the use of (Θ, ζ) is not based on the expected response patterns generated from an attribute hierarchy that would allow a description of attribute combinations demonstrated as being available to an examinee.

If an attribute hierarchy has been defined and a Q_r matrix derived to guide the construction of test items, the expected response patterns of expected examinees can be defined. It would seem reasonable, therefore, that the atypicality of an observed response pattern be judged relative to the expected response pattern based on the assumption that the attribute hierarchy is true. This approach would be much simpler than the application of the rule-space model, but would have less generality (i.e., it would not be applicable to a test that is not derived from the Q_r matrix). The procedure we propose does not allow direct identification of incorrect “rules” but it does allow the identification of those attribute combinations that are likely available to the examinee. Two methods are presented to illustrate the classification of observed response patterns in the AHM.

Method A. In this method an observed response pattern is compared against all expected response patterns where slips of the form $0 \rightarrow 1$ and $1 \rightarrow 0$ are identified. The product of the

probabilities of each slip is calculated to give the likelihood that the observed response pattern was generated from an expected response pattern for a given Θ . More formally, let V_j be the j^{th} expected response pattern for n items, and X be an observed response pattern of the same length. Then, $d_j = V_j - X$ produces a vector having elements $(-1, 0, +1)$ corresponding to the type of error that may exist, where $d_j = 0$ (no error), $d_j = -1$ [error of the form $0 \rightarrow 1$ with probability equal to $P_{jk}(\Theta)$], or $d_j = +1$ [error of the form $1 \rightarrow 0$ with probability equal to $1 - P_{jm}(\Theta)$]. In these equations, $P_{jk}(\Theta)$ is the probability of the k^{th} observed correct answer when an incorrect answer was expected (i.e., $0 \rightarrow 1$ error) and $1 - P_{jm}(\Theta)$ is the probability of the m^{th} observed incorrect answer when a correct answer was expected (i.e., $1 \rightarrow 0$ error). The probability of k errors of the form $0 \rightarrow 1$ together with m errors of the form $1 \rightarrow 0$ is given by

$$P_{j_{\text{Expected}}}(\Theta) = \prod_{k=1}^K P_{jk}(\Theta) \prod_{m=1}^M [1 - P_{jm}(\Theta)],$$

where k ranges from 1 to K (i.e., the subset of items with the $0 \rightarrow 1$ error) and m ranges from 1 to M (i.e., the subset of items with the $1 \rightarrow 0$ error). That is, the probabilities of positive slips ($0 \rightarrow 1$) are multiplied by the probabilities of negative slips ($1 \rightarrow 0$), resulting in an estimate of the likelihood that an observed response pattern approximates an expected response pattern at a given Θ . The examinee is classified as having the j^{th} set of attributes when the corresponding

$P_{j_{\text{Expected}}}(\Theta)$ is large.

For purposes of illustration, consider the classification of an examinee with the observed response pattern (11110000000000). Table 3⁴ contains the likelihood of this observed response pattern [i.e., $P_{j_{\text{Expected}}}(\Theta)$] given the expected response patterns and the associated ability level for errors of the form $0 \rightarrow 1$ and $1 \rightarrow 0$. The results in Table 3 indicate that the observed response pattern with the likelihood of 0.50 approximates the expected response pattern (11100000000000) with the associated attributes (111000) at the ability level of -0.50. One slip

occurred on item 4 of the form 0→1. In other words, if we compare the observed response pattern (11110000000000) to the expected response pattern (11100000000000) we see a 0→1 error for item 4.

To further illustrate classification method A, consider the very unusual observed response pattern (111111111111101) where item 14, which requires attributes (110111), is answered incorrectly. The results are shown in Table 4. For this observed response pattern the likelihood estimates are essentially 0. This outcome suggests a poor fit between the observed and the expected response patterns. The largest likelihood, 0.0318, for $\Theta = 2.37$ indicates the observed response may have come from the expected response pattern (11111111111111) with one slip. However, the outcome is very unlikely.

Method B. A second method for classifying an examinee's observed response pattern, method B, can be obtained by identifying all the expected response patterns that are logically contained within the examinee's observed response pattern. For example, if (111111111111101) is observed for an examinee, then an identification is made of all the expected response patterns that are logically included in this observed pattern [e.g., the expected response pattern (11100000000000) corresponding to the attribute vector (111000) is such a case]. When the expected response pattern is included in the observed response pattern, a "match" is noted and the associated attribute vector is identified as being present. When an expected response pattern is not logically included in the observed response pattern, the likelihood of the slips is computed. For example, the expected response pattern (110110110110110) is not logically included in the observed response pattern (111111111111101), and slips of the form 1→0 are identified and the product of their probabilities calculated. Another interpretation of this approach is that the expected response pattern is used as a mask on the observed response pattern and only those elements in the observed response pattern associated with 1s in the expected response pattern are considered. Thus, all slips for a specific comparison are of the form 1→0.

Table 5 shows the results of this approach for the observed response pattern (111111111111101). The asterisk indicates that the expected response pattern was found in the

observed response pattern (i.e., the expected response pattern is logically included in the observed response pattern). In line 15, there is a discrepancy in the 14th element of the expected response pattern (i.e., 1) and the observed response pattern (i.e., 0) indicating a slip of 1→0 with the probability 0.6835 for the ability estimate $\Theta = 1.54$. In the last line, the same discrepancy is a slip of 1→0 and its probability is 0.0318 for the ability estimate $\Theta = 2.37$. Thus, we can say the examinee possesses the attribute combinations specified by the attribute vectors starting with (100000) of line 2 to (100111) of line 14. The examinee also likely has the attributes specified by line 15 (110111) but definitely not the combination in line 16 (111111).

C. MODELING SYLLOGISTIC REASONING: AN APPLICATION OF THE AHM

Overview of Johnson-Laird's Mental Models Theory

The AHM is based on the assumption that specific cognitive skills called attributes can be organized in a hierarchy to form a cognitive model of test performance. The examinee must possess these skills to answer test items correctly. To illustrate an application of the AHM within the domain of syllogistic reasoning we will use Phil Johnson-Laird's theory of mental models (Johnson-Laird & Byrne, 1991; Johnson-Laird, 1999). Because this theory has received a significant amount of empirical support, it was used in the current study (Evans, Newstead, & Byrne, 1993; Johnson-Laird, 1999; Leighton & Sternberg, in press; Schaeken, De Vooght, Vandierendonck, & d'Ydewalle, 2000).

The theory of mental models can be illustrated with categorical syllogisms, which form a standard task used in psychological reasoning experiments (e.g., Johnson-Laird, 1999; Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991). Categorical syllogisms consist of two quantified premises and a quantified conclusion. The premises reflect an implicit relation between a subject (A) and a predicate (C) via a middle term (B), and the conclusion reflects an explicit relation between the subject (A) and predicate (C). The premises and conclusion below illustrate one form of a categorical syllogism:

All A are B
 All B are C
 ∴ ALL A are C.

Each of the premises and the conclusion contains a quantifier that connects the categories in the syllogism such as—All A are B, Some A are B, Some A are not B, or No A are B. The principle that guides all valid deductions in syllogistic reasoning is that the conclusion is valid only if it is true in every possible interpretation of its premises (Johnson-Laird & Bara, 1984).

According to Johnson-Laird's theory, reasoning is based on the manipulation of information by means of mental models (Johnson-Laird, 1983, 1999). Johnson-Laird (1983) proposed a three-step procedure for drawing logical inferences to syllogistic premises: In the first step, the reasoner constructs an initial "model" or representation that is analogous to the state of affairs (or information) being reasoned about. For example, consider that a reasoner is given two premises and asked to draw a necessary conclusion, if possible, from the premises (taken from Johnson-Laird & Bara, 1984):

PREMISES EXAMPLE #1

Some A are B
 No B are C.

The initial model or representation of the premises the reasoner constructs might be as follows:

INITIAL MODEL

A = B	
?A = B	
	C
	C

In the first line of the initial model, the "=" sign symbolizes that at least one "A" is equal to a "B," which reflects the information in the first premise—Some A are B. The question mark by the "A" in the second line of the model suggests the possibility that all "As" might be "Bs." This possibility needs to be considered by the reasoner because it corresponds to a formal interpretation of the

(as is shown in the initial model). Hence, another conclusion is needed—one that follows from all models of the premises. The conclusion that follows from all three models is Some A are not C. The connective “Some...not” is used to express the possibility that none of the As are C or that some of the As are C. According to Johnson-Laird and Bara (1984), the most difficult syllogisms require constructing three models in order to generate a valid conclusion—if such a conclusion is possible. The easiest syllogisms require the construction of a single model to generate a valid conclusion.

Mental model theory has been used successfully to account for participants' performance on categorical syllogisms (Evans, Handley, Harper, & Johnson-Laird, 1999; Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991). A number of predictions derived from the theory have been tested and observed. For instance, one prediction suggests that participants should be more accurate in deriving conclusions from syllogisms that require the construction of only a single model than from syllogisms that require the construction of multiple models. An example of a single model categorical syllogism is shown below:

PREMISES EXAMPLE #2

All A are B
All B are C.

The model for this single model syllogism is:

A = B = C
A = B = C
?B = C
?C.

A necessary conclusion derived from this model is All A are C. There is no other model that will support a different conclusion. In contrast, a multiple-model syllogism requires that participants construct at least two models of the premises in order to deduce a valid conclusion or determine that a valid conclusion cannot be deduced. Johnson-Laird and Bara (1984) tested the prediction that participants should be more accurate in deriving conclusions from single-model syllogisms than from multiple-model syllogisms by asking 20 untrained volunteers to make an inference from

each of 64 pairs of categorical premises randomly presented. The 64 pairs of premises included single-model and multiple-model problems. An analysis of participants' inferences revealed that valid conclusions declined significantly as the number of models needed to be constructed to derive a conclusion increased (Johnson-Laird & Bara, 1984, Table 6).

Using Mental Models Theory in the AHM

Identifying Attributes and Specifying the Hierarchy

The first two steps in mental models theory—the construction of an initial model and the generation of a conclusion—involve primarily comprehension processes. The third step, the search for alternative models, defines the process of syllogistic reasoning (Evans et al., 1993; Johnson-Laird & Byrne, 1991). Figure 4 illustrates one approach of casting mental model theory as a hierarchy of attributes. The first attribute in the hierarchy is the ability to interpret premises containing quantifiers according to formal logical criteria; in other words, the ability to interpret “Some A” as pertaining to at least one A and possibly all As. The logical interpretation conflicts with everyday or informal interpretations of “Some A,” the latter of which suggests that “Some” pertains to at least one A but not all As (Grice, 1975). For this reason, attribute A1 is present in the hierarchy; interpreting quantifiers according to formal criteria is fundamental to normative syllogistic reasoning. A2 involves the ability to create an initial model or representation of the premises; that is, combining the representation of the first and second premises into a whole representation. A3 involves the ability to draw a conclusion from the initial model created from the premises. A4 involves the ability to generate a second unique model of the premises; that is, to generate another interpretation of the premises. A5 involves the ability to generate a conclusion that is consistent with the initial model and this second model of the premises. A6 involves the ability to generate a third unique model of the premises, and, finally, A7 involves the ability to generate a conclusion that takes into account all three models of the premises.

The A matrix for this attribute hierarchy is:

$$A_{Mental\ Models} = \begin{pmatrix} 0100000 \\ 0011000 \\ 0000000 \\ 0000110 \\ 0000000 \\ 0000001 \\ 0000000 \end{pmatrix}. \quad (5)$$

The A matrix of order (k, k), where k is the number of attributes, indicates all the direct connections among attributes. For example, the first row of the matrix indicates that attribute A1 is directly connected only to attribute A2 as illustrated by the position of a 1 in the second column (0100000).

The R matrix, which is derived from the A matrix, indicates the direct and indirect relationships among attributes:

$$R_{Mental\ Models} = \begin{pmatrix} 1111111 \\ 0111111 \\ 0010000 \\ 0001111 \\ 0000100 \\ 0000011 \\ 0000001 \end{pmatrix}. \quad (6)$$

The first row of the R matrix of order (k, k), where k is the number of attributes, indicates that the first attribute is either directly or indirectly connected to all attributes in the hierarchy as indicated by the position of a 1 in all columns.

The Q matrix resulting from the hierarchy of attributes is of order (k, i), where k is the number of attributes and i is the number of items. The Q matrix for the mental models hierarchy of syllogistic reasoning is (7, 127)—that is, 127 items are possible given the independence of the 7 attributes. However, given that the 7 attributes are not independent but ordered in a hierarchy, the Q_r matrix is of order (7, 15):

$$\begin{array}{r}
 111111111111111 \\
 011111111111111 \\
 001010101010101 \\
 Q_{r_{Mental\ Models}} = 000111111111111 \\
 000001100110011 \\
 000000011111111 \\
 000000000001111
 \end{array} \quad (7)$$

The Q_r matrix indicates that 15 items must be created to probe the attributes in the mental model hierarchy. For example, column 4 or item 4 of the Q_r matrix probes attributes A1, A2, and A4 as indicated by the position of the 1s under column 4 in rows 1, 2, and 4. It is worthwhile noting that the adequacy of the hierarchy can be evaluated, in part, by the feasibility of creating items that probe specific combination of attributes. For example, can item 4 be created to probe attributes 1, 2, and 4? In other words, can an item be created that probes the following abilities:

- (1) Interprets premises containing quantifiers according to formal logical criteria,
- (2) Creates an initial model or representation of the premises, and
- (3) Generate a second model of the premises.

Generating a multiple-choice item that involved evaluating sets of dual models to a syllogism without drawing a conclusion could be one way of operationalizing item 4. That is, students would not be required to generate a conclusion to the syllogism, but would only be required to evaluate possible models of the syllogism. In this way, the attributes of interpreting quantifiers logically, and creating/evaluating possible representations or models of the premises could be assessed. A constructed-response item could also be used where students would be asked to draw possible representations of the syllogisms without drawing a conclusion.

Generating Expected Response Patterns

The expected response patterns, expected total scores, and expected examinee attribute vectors derived from the mental model hierarchy are shown in Table 6. Row 1 of Table 6 should be interpreted as follows: An examinee who only has attribute A1, [i.e., (1000000)] is expected to answer only the first item correctly, producing the expected response pattern

(100000000000000). Likewise, an examinee who only has attributes A1 and A2 is expected to answer only the first two items correctly, producing the expected response pattern (110000000000000). The total scores expected if the attribute hierarchy is a true description of how examinees solve categorical syllogisms are 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, and 15.

Estimating Probabilities of Item Responses

Employing the approach illustrated earlier (see Estimating Probabilities of Item Responses on p.11), the probabilities of item responses were calculated from the columns of the expected response matrix (i.e., the expected item response vectors) in Table 6 with BILOG 3.11 using the two-parameter (2PL) logistic IRT model. The purpose of this analysis is to generate estimates of the item parameters. The estimates are shown in Table 7. Some of the expected item characteristic curves display high slopes indicating that some items are very discriminating (see Figure 5). Also, as expected, the items probing preliminary cognitive attributes are less difficult than the items probing later or “final stage” attributes. Note that only 13 curves are distinguishable because the curves for items (3, 5) and (10, 12) are identical. The expected item and test information functions are shown in Figure 6.

Classification of Observed Response Patterns

Method A. According to method A, an observed response pattern is compared against all expected response patterns where slips of the form 0→1 and 1→0 are identified. The product of the probabilities of each slip is calculated to give the likelihood that the observed response pattern was generated from an expected response pattern for a given Θ (see section Classification of Observed Response Patterns on p.12). As an example, consider the classification of an examinee with the observed response pattern (1110000000000). Row 4 of Table 8 shows the likelihood that this observed response pattern originated from the expected response pattern (111000000000000) is 1.0000, whereas the likelihood that that this observed response pattern originated from any other expected response vector is essentially 0.

Consider now the classification of an examinee with an anomalous observed response pattern (101000000000000) (i.e., a response pattern not matching any of the expected response

patterns). As column 2 of Table 9 illustrates, the probability that this observed response pattern originated from any of the expected response patterns is essentially 0. This observed response pattern nearly matches the expected response pattern (11100000000000) in row 4. However, the likelihood that it approximates this expected response pattern is still very low because it would be highly unusual to answer item 3 correctly without also answering item 2 correctly.

Method B. Recall that method B involves identifying all the expected response patterns that are logically contained within the observed response pattern. For the observed response pattern (1010000000000000), Table 10 indicates that the only expected response pattern logically included in this observed response pattern is (1000000000000000).

What, then, can we say about an examinee who exhibits the observed response pattern (1010000000000000) according to the mental model hierarchy? If the hierarchy proposed is accurate in describing how examinees reason about categorical syllogisms, we can say that the examinee possesses attribute A1 (interprets quantifiers logically). The examinee, however, does not likely possess attribute A2 (can create an initial representation of the quantifiers) and it is highly unlikely that he or she possesses attribute A3 (can generate logical conclusions from the initial mental representations). In other words, the examinee may have some surface knowledge or understanding of logical quantifiers, which allow him or her to respond correctly to items involving definitions of quantifiers. But the examinee may not have a substantive understanding of logical quantifiers to draw a valid conclusion.

D. CONCLUSIONS AND DISCUSSION

The purpose of this paper was to introduce the attribute hierarchy model (AHM) for cognitive assessment. We began by providing an overview of the AHM, focusing on the cognitive and psychometric components. Specifically, we describe how an attribute hierarchy is formally represented, including the matrices that are derived from the hierarchy. Next, we described how expected item response patterns are generated from the Q_r matrix and the methods by which examinees' observed response patterns are classified. Finally, we applied the AHM to syllogistic

reasoning to demonstrate how this approach can be used to evaluate the cognitive processes required in a higher-level thinking task.

The results of this study provide a new method for evaluating examinee's cognitive skills. The results also suggest at least three different lines of future research. The first and most general application of this study is related to future applications of the AHM itself. With the AHM, test developers can liberate the single test score. Clearly, items can measure different attributes. The AHM offers a new method for assessing and reporting these attributes to examinees. This new method of assessment—where examinees receive measures of their cognitive skills rather than a single test score—should be applied and evaluated in a real testing situation. To-date, this has not been done. As a result, the authors intend to use this approach to study the cognitive attributes associated with syllogistic reasoning with first year university students. By using the AHM in an applied setting, we can evaluate its strengths and weaknesses and we can identify new issues that must be addressed when using this method for cognitive assessment.

A second line of future research, closely related to the first, should focus on attribute format issues. The attributes used in the AHM should promote the classification of cognitive skills. Scriven (1991) emphasized that a diagnosis always involves classifying the condition in terms of an accepted typology of afflictions and malfunctions. In other words, attribute descriptions should evolve into an accepted typology for talking about cognitive strengths and weakness. The typology or classification process should also provide results that are meaningful and useful to students and teachers as well as provide an interface between the test developer and test user. Presently, no such typology or agreed upon language exists.

To address this issue, a much better understanding of how students solve problems on tests is required. In addition, more research is needed to understand how teachers think about test problems (Frederiksen, Mislevy, & Bejar, 1993). In both cases, think aloud protocols may be useful. Perhaps instructional effects could be illuminated if patterns in students' problem-solving approach were apparent in the teachers' approach. By matching and evaluating these types of protocols, researchers may come to understand how and why students' solve problems on tests

in a specific way. Protocol analyses of students and teachers' problem-solving approaches may also provide a valuable link for the study of how instruction transfers to test performance.

Regardless, much more work is needed to understand how examinees' attributes should be formatted and presented to promote cognitive diagnosis and remediation as well as guide future instruction.

A third line of future research should focus on mining the distracters as well as the keyed option for information about the examinees' cognitive skills. Currently, attributes are associated with the cognitive skills used by examinee to solve test items correctly. However, the incorrect solutions may also yield meaningful information about attributes and cognitive skills. The AHM could profit from using the response patterns for the distracters and the keyed option to diagnose students' cognitive proficiencies. This outcome could be achieved by expanding each item in the reduced Q_r matrix to include responses to each distracter. This approach will only prove useful, however, if the distracters are created with the attributes in mind.

- Ericsson, K. A., & Simon, H. A. (1993). Protocol analyses: Verbal reports as data (rev. ed.). Cambridge, MA: MIT Press.
- Evans, J. St. B. T., Handley, S. J., Harper, C. N. J., & Johnson-Laird, P. N. (1999). Reasoning about necessity and possibility: A test of the mental model theory of deduction. Journal of Experimental Psychology: Learning, Memory, and Cognition, *25*, 1495-1513.
- Evans, St. B. T. J., Newstead, S. E., & Byrne, R. M. (1993). Human reasoning: The psychology of deduction. Hillsdale: Lawrence Erlbaum.
- Frederiksen, N., Glaser, R. L., Lesgold, A. M., & Shafto, M. G. (1990). Diagnostic monitoring of skills and knowledge acquisition. Hillsdale, NJ: Erlbaum.
- Frederiksen, N., Mislavy, R. J., Bejar, I. I. (1993). Test theory for a new generation of tests. Hillsdale, NJ: Erlbaum.
- Gierl, M. J. (1997). Comparing the cognitive representations of test developers and students on a mathematics achievement test using Bloom's taxonomy. Journal of Educational Research, *91*, 26-32.
- 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.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), Syntax and semantics Volume 3: Speech acts (pp. 41-58). London: Academic Press.
- Hamilton, L. S., Nussbaum, M., Snow, R. E. (1997). Interview procedures for validating science assessments. Applied Measurement in Education, *10*, 181-200.
- Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. Review of Research in Education, *24*, 393-446.
- Johnson-Laird, P. N. (1983). Mental models. Towards a cognitive science of language, inference, and consciousness. Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. N. (1999). Deductive reasoning. Annual Review of Psychology, *50*, 109-135.
- Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. Cognition, *16*, 1-61.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). Deduction. Hillsdale: Lawrence Erlbaum.

Leighton, J., Gierl, M. J., & Hunka, S. (1999, April). Attributes in Tatsuoka's rule-space model. Poster presented at the annual meeting of the National Council on Measurement in Education, Montréal, Quebec, Canada.

Leighton, J. P., Rogers, W. T., & Maguire, T. O. (1999). Assessment of student problem-solving on ill-defined tasks. Alberta Journal of Educational Research, *45*, 409-427.

Leighton, J. P., & Sternberg, R. J. (in press.). Reasoning and problem solving. In A. F. Healy & R. W. Proctor (Eds.), Experimental Psychology (pp. 000-000). Volume 4 in I. B. Weiner (Editor-in-Chief) Handbook of psychology. New York: Wiley.

Magone, M., Cai, J., Silver, E. A., & Wang, N. (1994). Validating the cognitive complexity and content quality of mathematics performance assessment. International Journal of Educational Research, *21*, 317-340.

Meijer, R. R. (1996). Person-fit research: An introduction. Applied Measurement in Education, *9*, 3-8.

Meijer, R. R. & Sijtsma, K. (2001). Methodological review: Evaluating person fit. Applied Psychological Measurement, *25*, 107-135.

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

Mislevy, R. J. (1996). Test theory reconceived. Journal of Educational Measurement, *33*, 379-416.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic test models [Computer Program]. Mooreville, IN: Scientific Software.

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.

Nichols, P. (1994). A framework of developing cognitively diagnostic assessments. Review of Educational Research, 64, 575-603.

Nichols, P. D., Chipman, S. F., Brennan, R. L. (1995). Cognitively diagnostic assessment. Hillsdale, NJ: Erlbaum.

Nichols, P., & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. Educational Measurement: Issues and Practice, 18 (2), 18-29.

Norris, S. P. (1990). Effect of eliciting verbal reports of thinking on critical thinking test performance. Journal of Educational Measurement, 27, 41-58.

Pellegrino, J. W. (1988). Mental models and mental tests. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 49-60). Hillsdale, NJ: Erlbaum.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practices. In A. Iran-Nejad & P. D. Pearson (Eds.), Review of Research in Education (pp. 307-353). Washington, DC: American Educational Research Association.

Roussos, L. (1994). Summary and review of cognitive diagnosis models. Unpublished manuscript, University of Illinois, Urbana-Champaign, The Statistical Laboratory for Educational and Psychological Measurement.

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

Schaeken, W., De Vooght, G., Vandierendonck, A., & d'Ydewalle, G. (Eds.). (2000). Deductive reasoning and strategies. Mahwah, NJ: Lawrence Erlbaum Associates.

Scriven, M. (1991). Evaluation thesaurus (4th ed.). Newbury Park, CA: Sage.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 263-331). New York: American Council on Educational, Macmillian.

- Sternberg, R. J. (1977). Intelligence, information processing, and analogical reasoning. Hillsdale, NJ: Erlbaum.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, *20*, 345-354.
- Tatsuoka, K. K. (1984). Caution indices based on item response theory. Psychometrika, *49*, 95-110.
- Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), Cognitively diagnostic assessment (pp. 327-359). Hillsdale, NJ: Erlbaum.
- Tatsuoka, K. K. (1996). Use of generalized person-fit indexes, Zetas for statistical pattern classification. Applied Measurement in Education, *9*, 65-75.
- Tatsuoka, M. M., & Tatsuoka, K. K. (1989). Rule space. In S. Kotz & N. L. Johnson (Eds.), Encyclopedia of statistical sciences (pp. 217-220). New York: Wiley.
- van der Linden, W., & Hambleton, R. K. (Eds.). (1997). Handbook of modern item response theory. New York: Springer.

Author Notes

Jacqueline P. Leighton, Centre for Research in Applied Measurement and Evaluation, University of Alberta, 6-110 Education Centre North, Faculty of Education, Edmonton, Alberta, Canada, T6G 2G5
Email: mark.gierl@ualberta.ca

Mark J. Gierl, Centre for Research in Applied Measurement and Evaluation, University of Alberta, 6-110 Education Centre North, Faculty of Education, Edmonton, Alberta, Canada, T6G 2G5 Email: mark.gierl@ualberta.ca

Stephen M. Hunka, Centre for Research in Applied Measurement and Evaluation, University of Alberta, 6-110 Education Centre North, Faculty of Education, Edmonton, Alberta, Canada, T6G 2G5
Email: steve.hunka@ualberta.ca

Footnotes

¹This assumption is similar to the information-processing metaphor used by some cognitive psychologists (e.g., Sternberg, 1977). According to the information-processing metaphor, human cognitive performance can be described as following from the application of ordered mental processes.

²A complete Mathematica (Wolfram, 1996) library with the algorithms for the procedures described in this paper is available from the authors.

³The adjacency, reachability, incidence, and reduced incidence matrices are also used in Kumi Tatsuoka's rule-space model (Tatsuoka, 1983, 1995; Tatsuoka & Tatsuoka, 1989). An overview of rule space is presented in Gierl, Leighton, and Hunka (2000).

⁴The results in Table 3 also help illustrate why the discrimination parameters are high for some items. Consider three points: First, the a-parameter for item 1 is 4.00. Second, the expected examinee response pattern for item 1 (i.e., the first column of the expected item response vectors in Table 3) is (0111111111111111). Third, the ability estimates for the first and second rows of the expected item response vectors in Table 3 are -1.58 and -0.85, respectively. Taken together, these results reveal that item 1 will effectively separate examinees with a given expected item response pattern within a relatively narrow ability range when the hierarchy is true.

Table 1

Expected Response Patterns, Total Scores, and Examinee Attributes for a Hypothetical Set of

Table 2

BLOG Item Parameter Estimates Using the Expected Item Response Vectors in Table 1

Item	a-parameter	b-parameter
1	4.00	-1.10
2	3.00	-0.30
3	0.70	1.30
4	5.00	-0.50
5	5.00	-0.15
6	2.47	1.28
7	0.50	1.40
8	2.66	1.26
9	3.00	1.63
10	0.50	1.36
11	2.66	1.26
12	2.30	1.82
13	3.00	1.70
14	3.00	1.70
15	4.00	1.60

Table 3

Classification of Observed Response Pattern (11110000000000) Using Method A

Theta	$P_{j_{Expected}}(\Theta)$	Slips	Expected Response Patterns	Attribute
-1.58	0.0000	4	00000000000000	000000
-0.85	0.0002	3	10000000000000	100000
-0.55	0.0406	2	11000000000000	110000
-0.50	0.5000	1	11100000000000	111000
-0.42	0.0400	2	10010000000000	100100
0.20	0.1070	2	11011000000000	110100
1.02	0.0000	2	11111100000000	111100
-0.39	0.0371	3	10010010000000	100110
1.02	0.0000	4	11011011000000	110110
1.42	0.0000	5	11111111100000	111110
-0.39	0.0368	3	10010000010000	100101
1.02	0.0000	4	11011000011000	110101
1.39	0.0000	5	11111100011100	111101
-0.19	0.0583	5	10010010010010	100111
1.54	0.0000	8	11011011011011	110111
2.37	0.0000	11	11111111111111	111111

Note. A slip is an inconsistency between the observed and expected response pattern.

Table 4

Classification of Observed Response Pattern (11111111111101) Using Method A With A Large Number of Slips

Theta	$P_{j_{Expected}}(\Theta)$	Slips	Expected Response Patterns	Attribute
-1.58	0.0000	14	00000000000000	000000
-0.85	0.0000	13	10000000000000	100000
-0.55	0.0000	12	11000000000000	110000
-0.50	0.0000	11	11100000000000	111000
-0.42	0.0000	12	10010000000000	100100
0.20	0.0000	10	11011000000000	110100
1.02	0.0000	8	11111100000000	111100
-0.39	0.0000	11	10010010000000	100110
1.02	0.0000	8	11011011000000	110110
1.42	0.0028	5	11111111100000	111110
-0.39	0.0000	11	10010000010000	100101
1.02	0.0000	8	11011000011000	110101
1.39	0.0024	5	11111100011100	111101
-0.19	0.0000	9	10010010010010	100111
1.54	0.0126	6	11011011011011	110111
2.37	0.0318	1	11111111111111	111111

Table 5

Classification of Observed Response Pattern (11111111111101) Using Method B

Theta	$P_{j_{Expected}}(\Theta)$	Slips	Expected Response Patterns	Attribute
-1.58	-	-	00000000000000	000000
-0.85	*	0	10000000000000	100000
-0.55	*	0	11000000000000	110000
-0.50	*	0	11100000000000	111000
-0.42	*	0	10010000000000	100100
0.20	*	0	11011000000000	110100
1.02	*	0	11111100000000	111100
-0.39	*	0	10010010000000	100110
1.02	*	0	11011011000000	110110
1.42	*	0	11111111100000	111110
-0.39	*	0	10010000010000	100101
1.02	*	0	11011000011000	110101
1.39	*	0	11111100011100	111101
-0.19	*	0	10010010010010	100111
1.54	0.6835	1	110110110110110	110111
2.37	0.0318	1	111111111111111	111111

Note. Each attribute vector with an asterisk is identified as being available to the examinee.

Table 6

Expected Response Patterns, Total Scores, and Examinee Attributes for a Hypothetical Set of Fifteen Examinees Based on the Mental Models Hierarchy in Figure 4

Examinee	Expected Response Patterns	Total Scores	Examinee Attributes
1	10000000000000	1	1000000
2	11000000000000	2	1100000
3	11100000000000	3	1110000
4	11010000000000	3	1101000
5	11111000000000	5	1111000
6	11010100000000	4	1101100
7	11111110000000	7	1111100
8	11010001000000	4	1101010
9	11111001100000	7	1111010
10	11010101010000	6	1101110
11	11111111111000	11	1111110
12	11010001000100	5	1101011
13	11111001100110	9	1111011
14	11010101010101	8	1101111
15	11111111111111	15	1111111

Table 7

BLOG Item Parameter Estimates Using the Expected Item Response Vectors in Table 6

Item	a-parameter	b-parameter
1	3.00	-2.25
2	4.00	-1.44
3	4.00	0.74
4	4.00	-0.60
5	4.00	0.74
6	0.50	1.00
7	1.20	1.70
8	1.20	0.15
9	5.00	1.00
10	0.70	1.60
11	4.00	1.73
12	0.70	1.60
13	2.00	1.80
14	2.00	2.40
15	3.00	2.34

Table 8

Classification of Observed Response Pattern (11100000000000) Using Method A

Theta	$P_{j_{Expected}}(\Theta)$	Slips	Expected Response Patterns	Attribute
-3.01	0.0000	3	0000000000000000	0000000
-1.82	0.0000	2	1000000000000000	1000000
-1.07	0.0000	1	1100000000000000	1100000
-0.34	1.0000	0	1110000000000000	1110000
-0.34	0.0001	2	1001000000000000	1101000
0.83	0.0000	2	1101100000000000	1111000
-0.19	0.0001	3	1111110000000000	1101100
0.91	0.0000	4	1001001000000000	1111100
0.17	0.0001	3	1101101100000000	1101010
1.18	0.0000	4	1111111110000000	1111010
0.45	0.0000	5	1001000001000000	1101110
1.88	0.0000	8	1101100001100000	1111110
0.37	0.0000	4	1111110001110000	1101011
1.51	0.0000	6	1001001001001000	1111011
0.68	0.0000	7	1101101101101100	1101111
3.03	0.0000	12	1111111111111111	1111111

Table 9

Classification of Observed Response Pattern (10100000000000) Using Method A

Theta	$P_{j_{Expected}}(\Theta)$	Slips	Expected Response Patterns	Attribute
-3.01	0.0000	2	0000000000000000	0000000
-1.82	0.0000	1	1000000000000000	1000000
-1.07	0.0000	2	1100000000000000	1100000
-0.34	0.0008	1	1110000000000000	1110000
-0.34	0.0000	3	1001000000000000	1101000
0.83	0.0000	3	1101100000000000	1111000
-0.19	0.0000	4	1111110000000000	1101100
0.91	0.0000	5	1001001000000000	1111100
0.17	0.0000	4	1101101100000000	1101010
1.18	0.0000	5	1111111110000000	1111010
0.45	0.0000	6	1001000001000000	1101110
1.88	0.0000	9	1101100001100000	1111110
0.37	0.0000	5	1111110001110000	1101011
1.51	0.0000	7	1001001001001000	1111011
0.68	0.0000	8	1101101101101100	1101111
3.03	0.0000	13	1111111111111111	1111111

Table 10

Classification of Observed Response Pattern (10100000000000) Using Method B

Theta	$P_{j_{Expected}}(\Theta)$	Slips	Expected Response Patterns	Attribute
-3.01	-	-	0000000000000000	0000000
-1.82	*	0	1000000000000000	1000000
-1.07	0.0941	1	1100000000000000	1100000
-0.34	0.0008	1	1110000000000000	1110000
-0.34	0.0001	2	1001000000000000	1101000
0.83	0.0000	3	1101100000000000	1111000
-0.19	0.0000	3	1111110000000000	1101100
0.91	0.0000	5	1001001000000000	1111100
0.17	0.0000	3	1101101100000000	1101010
1.18	0.0000	5	1111111110000000	1111010
0.45	0.0000	5	1001000001000000	1101110
1.88	0.0000	9	1101100001100000	1111110
0.37	0.0000	4	1111110001110000	1101011
1.51	0.0000	7	1001001001001000	1111011
0.68	0.0000	7	1101101101101100	1101111
3.03	0.0000	13	1111111111111111	1111111

Note. Each attribute vector with an asterisk is identified as being available to the examinee.

Figure Captions

Figure 1. Four hierarchical structures using six attributes, A1 to A6.

Figure 2. The expected item characteristic curves for the expected item response vectors in Table 1.

Figure 3. The expected item and test information functions for the expected item response vectors in Table 1.

Figure 4. Mental model theory of syllogistic reasoning cast into an attribute hierarchy.

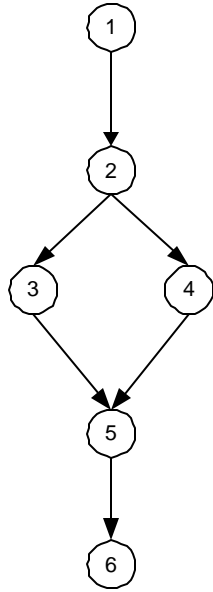
Figure 5. The expected item characteristic curves for the expected item response vectors in Table 6.

Figure 6. The expected item and test information functions for the expected item response vectors in Table 6.

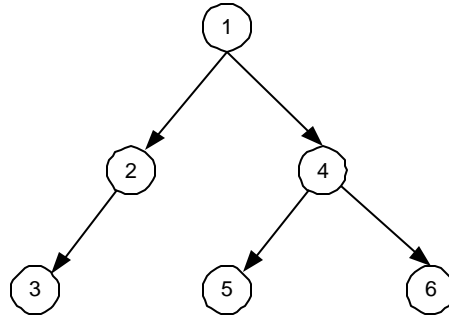
A. Linear



B. Convergent



C. Divergent



D. Unstructured

