Journal of Educational Measurement
Fall 2004, Vol. 41, No. 3, pp. 205–237

The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach

Jacqueline P. Leighton, Mark J. Gierl, and Stephen M. Hunka
Centre for Research in Applied
Measurement and Evaluation, University of Alberta

A cognitive item response theory model called the attribute hierarchy method (AHM) is introduced and illustrated. This method represents a variation of Tatsuoka's rule-space approach. The AHM is designed explicitly to link cognitive theory and psychometric practice to facilitate the development and analyses of educational and psychological tests. The following are described: cognitive properties of the AHM; psychometric properties of the AHM, as well as a demonstration of how the AHM differs from Tatsuoka's rule-space approach; and application of the AHM to the domain of syllogistic reasoning to illustrate how this approach can be used to evaluate the cognitive competencies required in a higher-level thinking task. Future directions for research are also outlined.

Most educational and psychological tests require examinees to engage in some form of cognitive problem solving. On these tests, the knowledge, mental processes, and strategies 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 psychological testing is apparent to many measurement specialists (e. g., Embretson, 1985; Frederiksen, Glaser, Lesgold, & Shafto, 1990; Frederiksen, Mislevy, & Bejar, 1993; Irvine & Kyllonen, 2002; 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 representation of a test is defined by the knowledge, mental processes, and strategies 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, processes, and strategies. 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 competencies 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 item response theory (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 article is to introduce and illustrate a cognitive IRT model called the attribute hierarchy method (AHM). This method represents an important variation of Tatsuoka's rule-space approach (Tatsuoka, 1983, 1984, 1996)—the AHM is used to model cognitive attributes that are hierarchically related. The AHM is similar to Tatsuoka's rule-space approach in so far as observed response patterns are classified or matched against ideal response patterns. In addition, the AHM makes use of Tatsuoka's matrices (i.e., the adjacency, the reachability, the incidence, and the reduced Q matrices) in the process of generating ideal response patterns. However, the fundamental difference between the AHM and Tatsuoka's rule-space approach lies in the assumption made about the cognitive attributes being modeled: In the AHM, the cognitive attributes are assumed to be hierarchically related and therefore dependent. In contrast, Tatsuoka's rule-space approach (Tatsuoka, 1990) indicates that cognitive attributes *need not* share hierarchical relations or dependencies to be modeled, nor do these relations need to be specified in an adjacency matrix. We believe that modeling cognitive attributes necessitates the specification of a hierarchy and its corresponding adjacency matrix. Cognitive research suggests that cognitive skills do not operate in isolation but belong to a network of interrelated processes (e.g., Kuhn, 2001; Vosniadou

& Brewer, 1992). Hence, the AHM, by incorporating the assumption of attribute dependency, represents an important variation to the seminal contribution already provided by Tatsuoka's rule-space approach. In particular, the AHM is designed to link explicitly cognitive theory with psychometric practice to facilitate the development and analyses of educational and psychological tests. In the first section we describe the cognitive properties of the AHM. In the second section we describe the psychometric properties of the AHM and suggest how the AHM differs from Tatsuoka's rule-space approach (Tatsuoka, 1983, 1984, 1996). In the third section we apply the AHM to the domain of syllogistic reasoning. In the fourth section we suggest future directions for research.

Cognitive Component of the Attribute Hierarchy Method

Identifying Cognitive Attributes

The AHM is based on the assumption that test performance depends on a set of hierarchically ordered competencies called attributes.[1] The examinee must possess these attributes to answer test items correctly. Attributes can be viewed as sources of cognitive complexity in test performance (cf. Embretson, 1995a). But, more generally, attributes are those basic cognitive processes or skills required to solve test problems correctly.

Cognitive attributes are discussed at length in Tatsuoka's research on the rule-space approach (e.g., Tatsuoka, 1990, 1991, 1993, 1995; see also M. Tatsuoka, 1986). In particular, the terminology of attributes was presented by Tatsuoka (1990) wherein she describes attributes as "production rules, procedural operations, item types, or, more generally any cognitive tasks" (p. 465). Furthermore, Chipman, Nichols, and Brennan (1995) later argued that "attributes characterize test items, and they may be interpreted as cognitive processes or skills that are required to perform correctly on a particular item" (p. 10). These initial characterizations of attributes are historically important because they introduced a vocabulary with which to describe the competencies needed to perform well on tasks. However, these initial characterizations of attributes lack cognitive psychological detail. For example, should attributes be viewed as dependent entities that organize themselves for the purpose of meeting problem-solving goals?

The importance of correctly identifying the attribute hierarchy cannot be overstated—the first step in making inferences with the AHM depends on accurately identifying the psychological ordering of cognitive competencies required to solve test problems. To be sure, identifying the ordering of attributes required to solve test problems can be challenging because cognitive theories of performance are not always easily applied to assessment purposes. However, identifying an attribute hierarchy of test performance serves a critical function: The hierarchy is a hypothesis of cognitive performance in the domain of interest and is subject to falsification if it fails to successfully classify examinees. In other words, the identification of the attribute hierarchy is the most important input variable for the AHM because it is used to predict the categories of student performance and to infer examinees' cognitive competencies. Leighton, Gierl, and Hunka (1999) provided one description of an attribute:

> 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.

Numerous studies have been conducted to identify various types of attributes needed to perform well on test items and tasks (e.g., Buck & Tatsuoka, 1998; Tatsuoka & Boodoo, 2000). Again, Tatsuoka and her colleagues conducted much of the seminal work on this front. 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 by specialists (e.g., test developers), who are familiar with the content area, test development process, and the way students solve problems, to identify the knowledge and competencies required to solve test items. Examinees can also be asked to think aloud as they solve problems, and protocol analysis (Ericsson & Simon, 1993) can be used to study their problem-solving strategies. Protocol analysis is an effective method for identifying the specific knowledge components and mental processes 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; Hamilton, Nussbaum, & Snow, 1997; Leighton, Rogers, & Maguire, 1999; Magone, Cai, Silver, & Wang, 1994; Norris, 1990).

In the application of the AHM, the hierarchy of attributes required to perform well in a domain must be identified prior to developing a test of the domain. This temporal order of events for applying the AHM is needed because the hierarchical organization of attributes must guide the development of test items. By using the attribute hierarchy to develop test items, the investigator achieves maximum control over the specific attributes each item measures. When test items are developed from a hierarchy, a unique adjacency matrix can be identified for the constructed items. Conversely, when test items have not been developed from a hierarchy and, consequently, the hierarchy must be abstracted from the items (i.e., the reduced Q matrix), it is difficult to identify a unique adjacency matrix for the items.

This temporal order of events is distinct from the one typically followed in the rule-space approach where attributes are often identified after the test items have already been constructed (although see Birenbaum, Kelly, & Tatsuoka, 1993; Birenbaum & Shaw, 1985 for examples where task analyses guided item development). This difference in temporal order is a major difference between the AHM and rule-space approach and reflects a different assumption in each method. In the AHM, cognitive attributes are believed to be *organized hierarchically* and any test of the attributes must be sensitive to their organization—this is best achieved by developing a test *after* the organi-

zation of attributes has been identified in a population of students through think-aloud protocol techniques for example (Gierl et al., 2000). In the rule-space approach, cognitive attributes are not assumed to be organized hierarchically; consequently, attributes are often identified by evaluating *existing* test items, which may not be sensitive to a complete hierarchical organization of attributes (e.g., Tatsuoka & Boodoo, 2000).

Specifying the Attribute Hierarchy to Model Test Performance

In the AHM, a hierarchical ordering of attributes is believed to underlie test performance. In the rule-space approach, a hierarchical ordering of attributes is not necessarily assumed because attributes may function independently (Tatsuoka & Boodoo, 2000). In support of this independence assumption, Tatsuoka and Boodoo (2000) analyzed Graduate Record Exam (GRE) quantitative items and found that 14 attributes could explain 78% of the variance in 30 item difficulties. Tatsuoka and Boodoo indicate that these attributes function independently in factor analytic studies and, thus, no hierarchical relationship was detected by applying the classification and regression tree method. However, cognitive attributes must be organized hierarchically for two reasons: First, cognitive research indicates that cognitive skills do not operate in isolation but belong to a network of interrelated competencies (Kuhn, 2001; Vosniadou & Brewer, 1992). Second, the generation of the adjacency and reachability matrices represents a numerical manifestation of the hierarchical organization of attributes. The adjacency matrix represents the *direct* relationships among attributes, whereas the reachability matrix represents the *direct and indirect* relationship among attributes. These matrices, which are sometimes used in the rule-space approach and, always, used in the AHM to generate expected response patterns (or *ideal response patterns* in rule-space terminology), represent cognitive hierarchies. *If cognitive attributes are not organized hierarchically, then it is unclear why the adjacency and reachability matrix are generated—these matrices lose their purpose and importance in modeling cognitive skills.*

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 (e.g., a series of well-defined, ordered cognitive steps identified via protocol analysis) or theoretical considerations (e.g., a series of developmental sequences suggested by Piaget such as preoperational, concrete operational, and formal operational). Once specified, the hierarchy containing the attributes serves as the cognitive model for test *development* and performance. Consequently, the attribute hierarchy has a foundational role in the AHM because it represents the construct and, by extension, the cognitive competencies that underlie test performance. In addition, the hierarchy leads to possible methods of categorizing tests as described in the next section.

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

From the assumption that attributes are hierarchically organized follows the idea that hierarchies of attributes can be combined to form increasingly complex structures of cognitive skills (Kim, 2001). In this section, we will briefly review the forms of hierarchical structures and their possible implications for test development and task construction.

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 competencies that are prerequisite to the attributes that follow or, alternatively, A1 may be considered a specific attribute. In Figure 1A 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 (A5 and A6 follow a similar interpretation). 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 examinee response pattern. The conditions required to obtain this relationship between the expected examinee response pattern and the total score are: (a) the hierarchy is true (i.e., the hierarchical relationships of attributes are a true model of examinees' cognitive attributes), (b) test items can be written that probe the appropriate attributes, and (c) the examinees respond without error. As such, the structure of an attribute hierarchy can be used to categorize tests according to the cognitive model underlying the test.

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

Figure 1B 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 is prerequisite to A5. This hierarchy, like Figure 1A, ends at a single point. This type

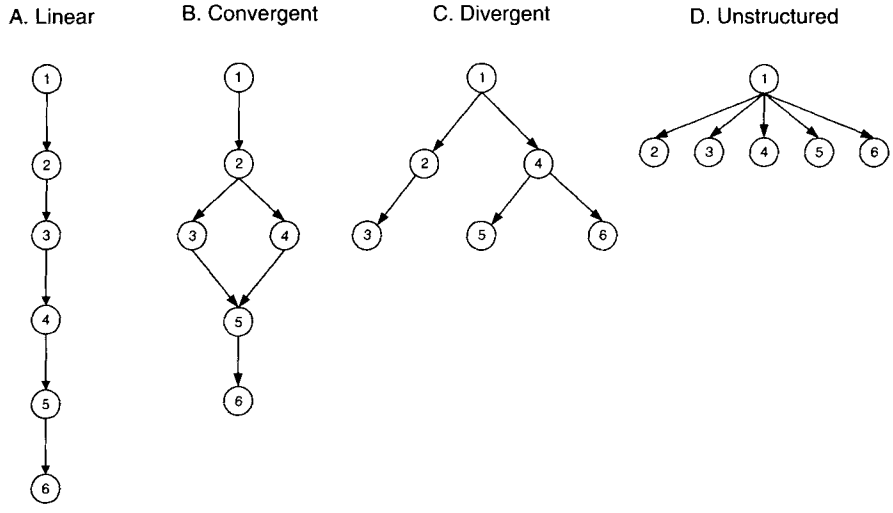

FIGURE 1. *Four hierarchical structures using six attributes.*

of hierarchy might be used to describe the cognitive competencies 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 1C represents a hierarchy having a divergent branch. This type of hierarchy might be used to describe the cognitive competencies 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 competencies 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 (Kim, 2001).

Psychometric Component of the Attribute Hierarchy Method

Formal Representation of a Hierarchy

To calculate the expected examinee response patterns for a specific hierarchy in the AHM, a formal representation of the hierarchy is required.[2] The formal representation of the hierarchy is based on Tatsuoka's rule-space approach and, in particular, the matrices of the rule space, including the adjacency, reachability, incidence, and reduced Q (see Tatsuoka, 1991, 1993, 1995; Tatsuoka & Tatsuoka, 1989). For descriptive purposes the divergent hierarchy of Figure 1C is used to illustrate these matrices (see also Tatsuoka, 1995). The *direct* relationships among attributes are specified by a binary *adjacency matrix (A)* of order (k, k) where k is the number of attributes. The A matrix for the hierarchy of Figure 1C is given below:

$$\begin{bmatrix} 010100 \\ 001000 \\ 000000 \\ 000011 \\ 000000 \\ 000000 \end{bmatrix} \tag{1}$$

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 using $R = (A + I)^n$, where n is the integer required for R to reach invariance and can represent the numbers 1 through 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 jth row of the R matrix specifies all the attributes,

including the jth attribute, for which the jth attribute is a direct or indirect prerequisite. The R matrix for Figure 1C is given below:

$$\begin{bmatrix} 111111 \\ 011000 \\ 001000 \\ 000111 \\ 000010 \\ 000001 \end{bmatrix} \tag{2}$$

For the R matrix, row 1 indicates that A1 is a prerequisite to all attributes; row 2 indicates that A2 is a prerequisite to A2 and A3. The rest of the matrix can be interpreted in the same manner. The R matrix is used to create a subset of items (from a set of potential items) that is conditioned on the structure of the attribute hierarchy.

The *set of potential items* is considered a pool of items that probes all combinations of attributes when the attributes are independent (i.e., when the corresponding R matrix is an identity matrix as it would be for the structure of Figure 1D if the prerequisite attribute were deleted). The size of the set of potential items is $2^k - 1$, where k is the number of attributes. Thus, for even a small value of k, the set of potential items is quite large. The set of potential items is described by the *incidence matrix (Q)* of order (k, i) where k is the number of attributes and i is the number of potential items. In the Q matrix, each item is described by the attributes that are required to respond correctly to the item or, conversely, the attributes required of an examinee for a correct answer. Simply converting the items ranging from 1 to $2^k - 1$ to their binary equivalents can form the columns of the Q matrix. The Q matrix for Figure 1C is shown below:

$$\begin{bmatrix} 101010101010101010101010101010101010101010101010101010101010101 \\ 011001100110011001100110011001100110011001100110011001100110011 \\ 000111100001111000011110000111100001111000011110000111100001111 \\ 000000011111111000000001111111100000000111111110000000011111111 \\ 000000000000000011111111111111110000000000000000111111111111111 \\ 000000000000000000000000000000011111111111111111111111111111111 \end{bmatrix} \tag{3}$$

The Q matrix is of order 6 by 63 (i.e., attributes by items), and each column of the matrix represents one item. For example, column 1 of this matrix represents item 1, and it specifies that only attribute A1 is required to correctly answer this item. Conversely, item 63 (column 63) requires all six attributes for a correct response. The set of potential items can be reduced when the attributes are related in a hierarchical structure. For example, the binary representation of item 2 is (010000) indicating the item probes attribute 2. However, the hierarchy indicates that attribute 2 requires a prerequisite of attribute 1. Thus, item 2 must be represented by (110000). But this representation makes item 2 identical to item 3 with respect to the attributes probed. For this reason item 2 can be removed. The removal of items in this manner produces a *reduced Q matrix* (Q_r).

Determining which columns of the R matrix are logically included in columns of the Q matrix, using Boolean inclusion, forms the Q_r matrix. For example, the last column of the R matrix specifies that any item requiring attribute A6 must also include

at least attributes A1 and A4; if it does not, then the item is removed. The Q_r matrix from Figure 1C is shown below:

$$\begin{bmatrix} 111111111111111 \\ 011011011011011 \\ 001001001001001 \\ 000111111111111 \\ 000000111000111 \\ 000000000111111 \end{bmatrix} \tag{4}$$

The Q_r matrix has a particularly important meaning for test development: It represents the *attribute blueprint* or *cognitive specifications* for test construction. The 15 columns of the Q_r matrix indicate that at least 15 items must be created to reflect the relationships in the hierarchy. For example, as shown in column 1 of the Q_r matrix, an item must be created to measure attribute A1. Similarly, as shown in column 2, an item must be created to measure attributes A1 and A2. The remaining columns are interpreted in the same manner. We have noted the impact of cognitive theory on test design has been limited. With the AHM, this limitation can be overcome when the cognitive requirements are described in the attribute hierarchy and the items required to measure these attributes are specified in the Q_r matrix. As a result, cognitive theory has a clearly defined role in test design using the AHM.

Generating Expected Examinee Response Patterns

Given a hierarchy of attributes, the expected examinee response patterns can be calculated. We use the term "expected examinee response patterns" instead of the original term "ideal examinee response patterns," which Tatsuoka uses (see Tatsuoka, 1995, for a description of the algorithm used to generate ideal response patterns in the rule-space approach; see also Tatsuoka, 1991, 1993, and 1995 for a description of ideal response patterns) because the term expected is a reminder that these response patterns should be observed *if the attribute hierarchy is true as specified in the adjacency or A matrix.* Unlike the term ideal, which simply suggests that these response vectors are desirable or suitable, the term expected conveys the predicted or required nature of these responses, if the A matrix is true.

Associated with these expected 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 examinee response pattern (recall, expected examinee response patterns are derived from the attribute hierarchy). For the hierarchy of Figure 1C the expected examinee response patterns, total scores, and expected examinee attributes are shown in Table 1. The rows and columns of the expected response matrix in Table 1 have distinct interpretations. Row 1 of the expected response matrix should be interpreted as follows: An examinee who only has attribute A1 [i.e., (100000)] is expected to answer only the first item correctly, producing the *expected examinee response pattern* (100000000000000). In contrast, column 1 of the expected response matrix should be interpreted as follows: An item that probes attribute A1 should be answered correctly by all examinees, producing the

TABLE 1
Expected Response Matrix, Total Scores, and Examinee Attributes for a Hypothetical Set of 15 Examinees Based on the Hierarchy in Figure 1C

Examinee	Expected Response Matrix	Total Scores	Examinee Attributes
1	100000000000000	1	100000
2	110000000000000	2	110000
3	111000000000000	3	111000
4	100100000000000	2	100100
5	110110000000000	4	110100
6	111111000000000	6	111100
7	100100100000000	3	100110
8	110110110000000	6	110110
9	111111111000000	9	111110
10	100100000100000	3	100101
11	110110000110000	6	110101
12	111111000111000	9	111101
13	100100100100100	5	100111
14	110110110110110	10	110111
15	111111111111111	15	111111

expected item response pattern (111111111111111). The expected item response patterns are used to estimate item parameters with item response theory (IRT) models (discussed in the next section). Also notice that an examinee's 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* examinee response pattern is consistent with the probability of a correct response derived from an IRT model. In the AHM, the *expected item characteristic curve* can be calculated for each item 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 1C 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

TABLE 2
BILOG Item Parameter Estimates Using the Expected Item Response Patterns 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

expected item response patterns (given by the *columns* of the expected response matrix in Table 1).

To illustrate this procedure, a sample of 1,000 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 patterns 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 calculated 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.[3] 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 with the AHM

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 were 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. When an attribute hierarchy has been identified, from which the Q_r matrix can be derived to guide the construction of test items, the

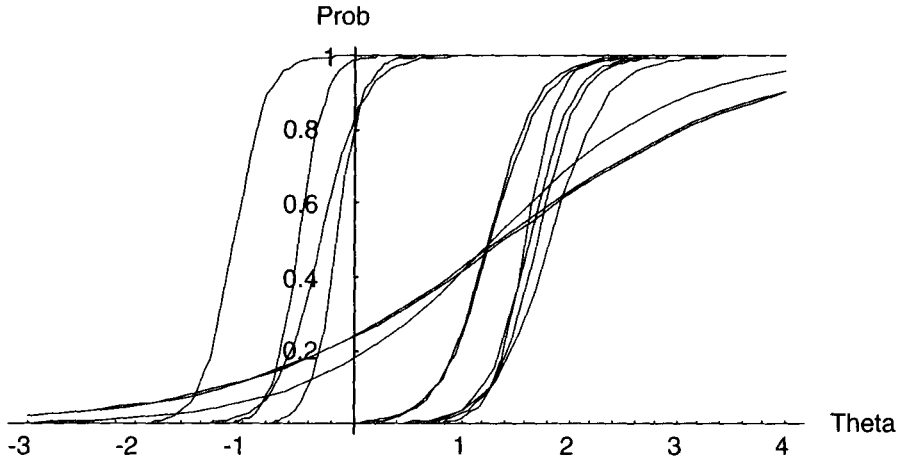

FIGURE 2. *The expected item characteristic curves for the expected item response patterns in Table 1.*

expected response patterns of examinees can be defined. As a result, anomalous observed response patterns can be judged relative to the set of expected examinee response patterns, which are based on the attribute hierarchy.

By its very definition, an anomalous or unusual response pattern requires the specification of an expected or usual response pattern. Numerous fit indices have been proposed for the identification of anomalous 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).

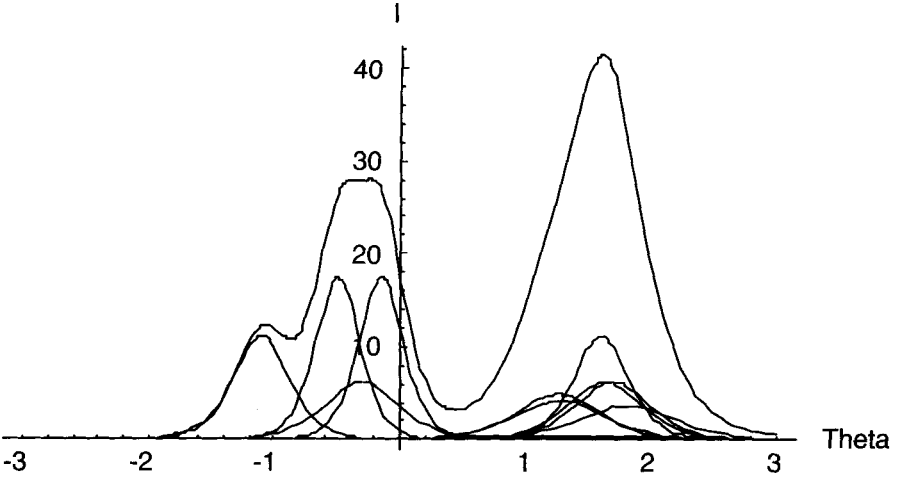

FIGURE 3. *The expected item and test information functions for the expected item response patterns in Table 1.*

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 attributes required to solve test items (e.g., "sleeping behavior," "guessing behavior," "cheating," "plodding," "alignment errors"; see Meijer, 1996, pp. 4–5). The AHM attempts to improve upon this by linking an observed response pattern to a hierarchy of cognitive attributes even when the observed response pattern is unusual in comparison to the expected examinee response patterns.

The rule-space approach. A distinct method to the classification of cognitive competencies is found in the literature on the rule-space approach (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 responses for 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 $(\Theta, \zeta_1, \zeta_2, \zeta_3$, etc.) where the ζ values are based on different subsets of items].

In the rule-space approach, identification of deficient examinee characteristics is based on a posteriori analysis of the attributes or "rules" required to answer each item. Thus, the use of (Θ, ζ) is *not* based on an a priori analysis of the expected examinee response patterns generated from an attribute hierarchy. Applications of the rule-space approach 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 (Tatsuoka, 1991, 1993, 1995, however, suggests that the relationship among attributes—if there is one—can be deduced from an incidence [Q] matrix obtained directly from an analysis of existing test items). Our opinion is that it might be possible to identify the independent attributes from the Q matrix, but it is not possible to identify a hierarchy of attributes from the Q matrix. The Q matrix only shows the attributes that co-occur together and not the generic hierarchical basis of relationships among attributes. In addition, it is also not possible to identify a unique attribute hierarchy from the reduced Q matrix (even though this matrix represents the direct and indirect relationships between attributes).

If an attribute hierarchy has been identified and a Q_r matrix derived to guide the construction of test items, then 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 examinee 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 approach, but would have less generality (i.e., it would not be applicable to a test that is not derived from the Q_t 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: Preliminary classification. In this method an observed response pattern is compared against all expected examinee 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 examinee response pattern for a given Θ. More formally, let V_j be the jth expected examinee response pattern for n items, and X be an observed response pattern of the same length. Then, $d_j = V_j - X$ produces a pattern 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→1 with probability equal to $P_{jk}(\Theta)$], or $d_j = +1$ [error of the form 1→0 with probability equal to $1 - P_{jm}(\Theta)$]. In these equations, $P_{jk}(\Theta)$ is the probability of the kth observed correct answer when an incorrect answer was expected (i.e., 0→1 error) and $1 - P_{jm}(\Theta)$ is the probability of the mth observed incorrect answer when a correct answer was expected (i.e., 1→0 error). The probability of k errors of the form 0→1 together with m errors of the form 1→0 is given by

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

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

For purposes of illustration, consider the classification of an examinee with the observed response pattern (111100000000000). Table 3[4] contains the likelihood of this observed response pattern [i.e., $P_{j_{Expected}}(\Theta)$] given the expected examinee response patterns and the associated ability level for errors of the form 0→1 and 1→0. The results in Table 3 indicate that the observed response pattern approximates the expected examinee response pattern (111000000000000) with the associated attributes (111000) with the likelihood of 0.5000 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 (111100000000000) to the expected examinee response pattern (111000000000000) 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),

TABLE 3
Classification of Observed Response Pattern (11110000000000) Using Method A

Theta	$P_{l_{Expected}}(\Theta)$	Slips	Expected Response Matrix	Examinee Attributes
−1.58	0.0000	4	0000000000000000	000000
−0.85	0.0002	3	1000000000000000	100000
−0.55	0.0406	2	1100000000000000	110000
−0.50	0.5000	1	1110000000000000	111000
−0.42	0.0400	2	1001000000000000	100100
0.20	0.1070	2	1101100000000000	110100
1.02	0.0000	2	1111110000000000	111100
−0.39	0.0371	3	1001001000000000	100110
1.02	0.0000	4	1101101100000000	110110
1.42	0.0000	5	1111111110000000	111110
−0.39	0.0368	3	1001000001000000	100101
1.02	0.0000	4	1101100001100000	110101
1.39	0.0000	5	1111110001110000	111101
−0.19	0.0583	5	1001001001001000	100111
1.54	0.0000	8	1101101101101100	110111
2.37	0.0000	11	1111111111111111	111111

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

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 examinee response patterns. The largest likelihood, 0.0318, for $\Theta = 2.37$ indicates the observed response may have come from the expected examinee response pattern (11111111111111) with one slip. However, the outcome is very unlikely. The conclusion that the observed response pattern (11111111111101) did not likely originate from the expected examinee response pattern (11111111111111) is reasonable given that such an observed response pattern (11111111111101) should not be produced consistently over a large sample of examinees if the attribute hierarchy is true.

Method B: Verification of preliminary classification. A second method for classifying an observed response pattern, method B, can be obtained by identifying all the expected examinee response patterns that are logically contained within the observed response pattern. For example, if (11111111111101) is observed for examinees, then all the expected examinee response patterns that are logically included in this observed response pattern [e.g., the expected examinee response pattern (11100000000000) corresponding to the attribute pattern (111000) is such a case] are identified. When the expected examinee response pattern is included in the observed response pattern, a "match" is noted and the associated attribute pattern is identified as being present for the examinees. When an expected examinee response pattern is not logically included in the observed response pattern, the likelihood of the slips is computed. For example, the expected examinee response pattern (11011011011011) is not logically included in the observed response pattern (11111111111101), 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 examinee response pattern is used

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 Matrix	Examinee Attributes
−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	11011011011010	110111
2.37	0.0318	1	11111111111111	111111

as a mask on the observed response pattern and only those elements in the observed response pattern associated with 1s in the expected examinee 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 (11111111111101). The asterisk indicates that the expected examinee response pattern is logically included in the observed response pattern. In line 15, there is a discrepancy in the 14th element of the expected examinee 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 that examinees possess the attribute combinations specified by the attribute patterns starting with (100000) of line 2 to (100111) of line 14. The examinees also likely have the attributes specified by line 15 (110111) but probably not the combination in line 16 (111111). Notice that the method B classification is more conservative than the method A classification of the observed response pattern (11111111111101). Method B classified the observed response pattern as definitely having the associated attributes (100111) and possibly the attributes (110111) with a theta level of 1.54, whereas method A classified the observed response pattern as having the associated attributes (111111) with a theta level of 2.37, albeit with a likelihood of only 0.0318.

Modeling Syllogistic Reasoning: An Application of the AHM

The AHM is based on the assumption that specific cognitive competencies called attributes are organized in a hierarchy to form a cognitive model of test performance.

TABLE 5
Classification of Observed Response Pattern (11111111111101) Using Method B

Theta	$P_{j_{Expected}}(\Theta)$	Slips	Expected Response Matrix	Examinee Attributes
−1.58	—	—	000000000000000	000000
−0.85	*	0	100000000000000	100000
−0.55	*	0	110000000000000	110000
−0.50	*	0	111000000000000	111000
−0.42	*	0	100100000000000	100100
0.20	*	0	110110000000000	110100
1.02	*	0	111111000000000	111100
−0.39	*	0	100100100000000	100110
1.02	*	0	110110110000000	110110
1.42	*	0	111111111000000	111110
−0.39	*	0	100100000100000	100101
1.02	*	0	110110000110000	110101
1.39	*	0	111111000111000	111101
−0.19	*	0	100100100100100	100111
1.54	0.6835	1	110110110110110	110111
2.37	0.0318	1	111111111111111	111111

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

The examinee must possess these attributes to answer test items correctly. To illustrate an application of the AHM within the domain of syllogistic reasoning we will use Philip Johnson-Laird's theory of *mental models* (Johnson-Laird, 1983; Johnson-Laird & Bara, 1984; Johnson-Laird & Byrne, 1991) because it has received a significant amount of empirical support (Evans, Newstead, & Byrne, 1993; Johnson-Laird, 1999; Leighton & Sternberg, 2003; 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 & 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 and Byrne (1991) 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 connective "Some" (Johnson-Laird & Bara, 1984). The second premise—No B are C—is represented in the third and fourth line of the initial model by separating the "Cs" from the "Bs."

The second step in the procedure involves *drawing a conclusion from the initial model.* For example, from the initial model the reasoner might conclude that *No A are C.* This conclusion follows from the initial model of the premises. However, it is based on a single model or representation of the premises and may not necessarily follow from alternative models of the premises. To verify that this preliminary conclusion does follow necessarily from the premises, the reasoner needs to check whether the conclusion is consistent with alternative models of the premises.

The third step in Johnson-Laird's theory of mental models is to *construct alternative models* of the premises to verify (or falsify) the preliminary conclusion drawn (Johnson-Laird, 1999; Johnson-Laird & Byrne, 1991). This step involves considering other models of the premises to draw a final conclusion that is consistent with all possible models of the premises. The alternative models of the premise set above—Some A are B; No B are C—can be illustrated as follows:

MODEL # 1		MODEL # 2	
A = B		A = B	
?B		?B	
?A	C	?A	C
	C	?A	C

These two alternative models of the premises render false the preliminary conclusion derived from the initial model. The first alternative, model #1, suggests the possibility that there is at least one "A," which is not a "B," that is a "C" (rendering the preliminary conclusion *No A are C* false). The second alternative, model #2, suggests that all "As" that are not "Bs" are "Cs" (also rendering the preliminary conclusion *No A are C* false). These alternative models could lead to a new conclusion—Some A are C. But this new conclusion does not follow from the initial model in which none of the "As" are "Cs." (Because the connective "Some" expresses that at least one A is a C and the possibility that all As are Cs, it fails to include the possibility that none of the As are Cs [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 Cs or that some of the As are Cs. According to Johnson-Laird and Bara (1984), the most difficult syllogisms require constructing three models 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 models 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.* All alternative models of these premises will support this conclusion. In contrast, a multiple-model syllogism requires that participants construct at least two models of the premises 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 that needed to be constructed to derive a conclusion increased (Johnson-Laird & Bara, 1984).

Using Mental Models Theory in the AHM

Identifying the Hierarchy of Attributes

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 course of reasoning (Evans et al., 1993; Johnson-Laird & Byrne, 1991). Figure 4 illustrates one approach of casting mental models theory as a hierarchy of attributes. The first attribute in the hierarchy

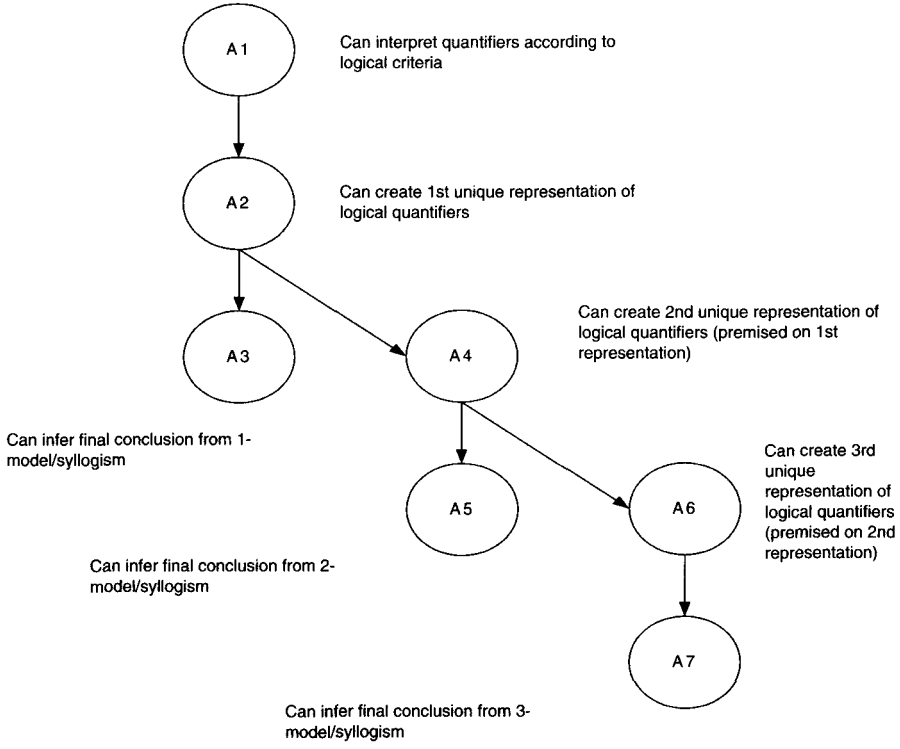

FIGURE 4. *Mental models theory of syllogistic reasoning cast into an attribute hierarchy.*

is the ability to interpret quantified premises 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 representation 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:

$$A_{Mental\ Models} = \begin{bmatrix} 0100000 \\ 0011000 \\ 0000000 \\ 0000110 \\ 0000000 \\ 0000001 \\ 0000000 \end{bmatrix}. \tag{5}$$

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:

$$R_{Mental\ Models} = \begin{bmatrix} 1111111 \\ 0111111 \\ 0010000 \\ 0001111 \\ 0000100 \\ 0000011 \\ 0000001 \end{bmatrix}. \tag{6}$$

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):

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

The Q_r matrix indicates that at least 15 items must be created to probe the attributes in the mental models 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 a 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 competencies?

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

Generating a multiple-choice item that involved evaluating a set 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 the 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 Examinee Response Patterns

The expected examinee response patterns, expected total scores, and expected examinee attributes derived from the mental models 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 examinee response pattern (100000000000000). Likewise, in row 2, an examinee who only has attributes A1 and A2 is expected to answer only the first two items correctly, producing the expected examinee 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* in this article), the probabilities of item responses were calculated from the columns of the expected response matrix in Table 6 (i.e., the *expected item* response patterns) with BILOG 3.11 using the two-parameter (2PL) logistic IRT model. The

TABLE 6
Expected Examinee Response Patterns, Total Scores, and Examinee Attributes for a Hypothetical Set of 15 Examinees Based on the Mental Models Hierarchy in Figure 4

Examinee	Expected Response Matrix	Total Scores	Examinee Attributes
1	100000000000000	1	1000000
2	110000000000000	2	1100000
3	111000000000000	3	1110000
4	110100000000000	3	1101000
5	111110000000000	5	1111000
6	110101000000000	4	1101100
7	111111100000000	7	1111100
8	110100010000000	4	1101010
9	111110011000000	7	1111010
10	110101010100000	6	1101110
11	111111111110000	11	1111110
12	110100010001000	5	1101011
13	111110011001100	9	1111011
14	110101010101010	8	1101111
15	111111111111111	15	1111111

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 attributes in the hierarchy. 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.

TABLE 7
BILOG Item Parameter Estimates Using the Expected Item Response Patterns 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

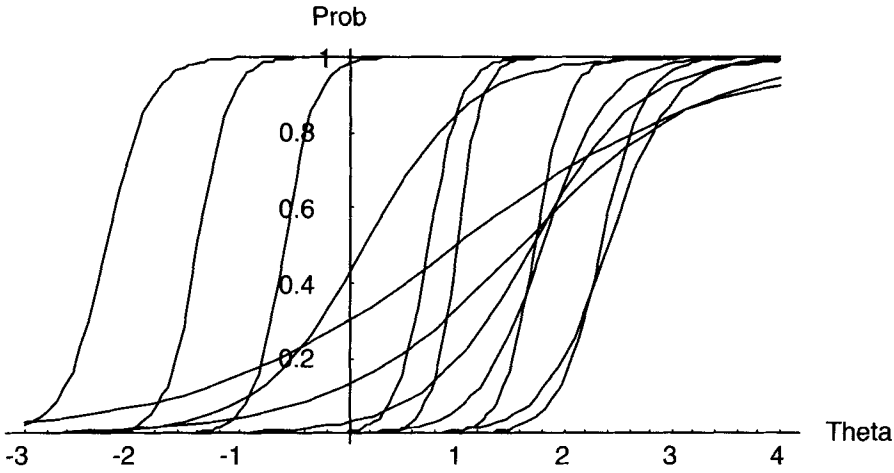

FIGURE 5. *The expected item characteristic curves for the expected item response patterns in Table 6.*

Method A: Preliminary classification. According to Method A, an observed response pattern is compared against all expected examinee 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 examinee response pattern for a given Θ (see *Classification of Observed Response Patterns with the AHM* in this article). As an example, consider the classification of examinees with the observed response pattern (11100000000000).

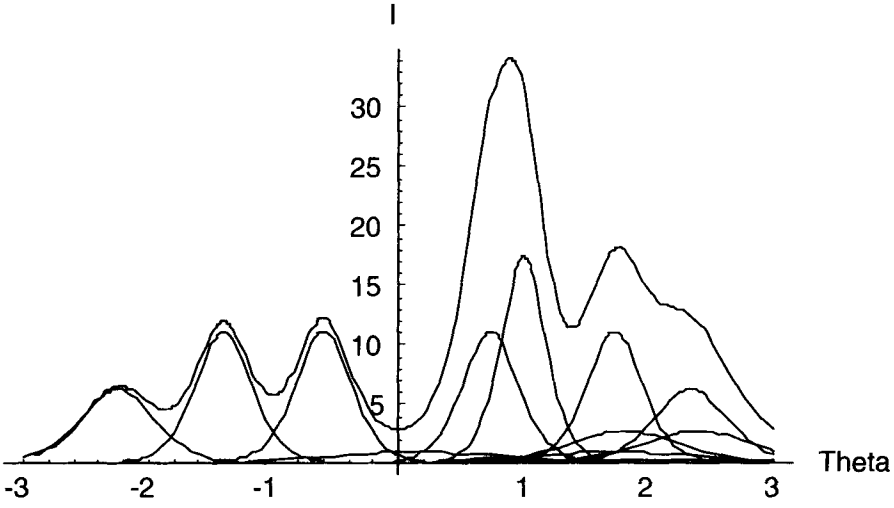

FIGURE 6. *The expected item and test information functions for the expected item response patterns in Table 6.*

TABLE 8
Classification of Observed Response Pattern (111000000000000) Using Method A

Theta	$P_{jExpected}(\Theta)$	Slips	Expected Response Matrix	Examinee Attributes
−3.01	0.0000	3	000000000000000	0000000
−1.82	0.0000	2	100000000000000	1000000
−1.07	0.0000	1	110000000000000	1100000
−0.34	1.0000	0	111000000000000	1110000
−0.34	0.0001	2	100100000000000	1101000
0.83	0.0000	2	110110000000000	1111000
−0.19	0.0001	3	111111000000000	1101100
0.91	0.0000	4	100100100000000	1111100
0.17	0.0001	3	110110110000000	1101010
1.18	0.0000	4	111111111000000	1111010
0.45	0.0000	5	100100000100000	1101110
1.88	0.0000	8	110110000110000	1111110
0.37	0.0000	4	111111000111000	1101011
1.51	0.0000	6	100100100100100	1111011
0.68	0.0000	7	110110110110110	1101111
3.03	0.0000	12	111111111111111	1111111

Examinees producing this pattern are assumed to have mastered the first three attributes in the mental models hierarchy; that is, interpreting quantifiers according to logical criteria (A1), creating an initial representation of quantifiers (A2), and inferring a conclusion from this representation (A3), which are the attributes associated with correctly answering one-model problems. Row 4 of Table 8 shows the likelihood that this

TABLE 9
Classification of Observed Response Pattern (101000000000000) Using Method A

Theta	$P_{jExpected}(\Theta)$	Slips	Expected Response Matrix	Examinee Attributes
−3.01	0.0000	2	000000000000000	0000000
−1.82	0.0000	1	100000000000000	1000000
−1.07	0.0000	2	110000000000000	1100000
−0.34	0.0008	1	111000000000000	1110000
−0.34	0.0000	3	100100000000000	1101000
0.83	0.0000	3	110110000000000	1111000
−0.19	0.0000	4	111111000000000	1101100
0.91	0.0000	5	100100100000000	1111100
0.17	0.0000	4	110110110000000	1101010
1.18	0.0000	5	111111111000000	1111010
0.45	0.0000	6	100100000100000	1101110
1.88	0.0000	9	110110000110000	1111110
0.37	0.0000	5	111111000111000	1101011
1.51	0.0000	7	100100100100100	1111011
0.68	0.0000	8	110110110110110	1101111
3.03	0.0000	13	111111111111111	1111111

observed response pattern originated from the expected examinee response pattern (11100000000000) is 1.0000, whereas the likelihood that that this observed response pattern originated from any other expected examinee response pattern is essentially 0.

Consider now the classification of an anomalous observed response pattern (10100000000000) (i.e., a response pattern not matching any of the expected examinee response patterns). This response pattern is inconsistent with the mental models hierarchy of syllogistic reasoning, and is likely due to a response slip or data coding error. Anomalous patterns, like this example, can still be classified. As Table 9 illustrates, this observed response pattern nearly matches the expected examinee response pattern (11100000000000) in row 4. However, the likelihood that it approximates this expected examinee response pattern is still very low (i.e., 0.0008) because it would be highly unusual to answer item 3 correctly without also answering item 2 correctly according to the mental models hierarchy. The probability that this observed response pattern originated from any other expected examinee response pattern in Table 9 is essentially 0.

Method B: Verification of preliminary classification. Recall that Method B involves identifying all the expected examinee response patterns that are logically contained within the observed response pattern. For the observed response pattern (10100000000000), Table 10 indicates that the only expected examinee response pattern logically included in this observed response pattern is (10000000000000). Again, we observe that Method B's classification is more stringent. According to Method B, examinees producing the observed response pattern (10100000000000) are classified as having the attribute pattern (1000000) and a theta of −1.82, whereas according to Method A, examinees producing this observed pattern are classified as having attribute pattern (1110000) and a theta of −0.34.

TABLE 10
Classification of Observed Response Pattern (10100000000000) Using Method B

Theta	$P_{jExpected}(\Theta)$	Slips	Expected Response Matrix	Examinee Attributes
−3.01	—	—	00000000000000	0000000
−1.82	*	0	10000000000000	1000000
−1.07	0.0941	1	11000000000000	1100000
−0.34	0.0008	1	11100000000000	1110000
−0.34	0.0001	2	10010000000000	1101000
0.83	0.0000	3	11011000000000	1111000
−0.19	0.0000	3	11111100000000	1101100
0.91	0.0000	5	10010010000000	1111100
0.17	0.0000	3	11011011000000	1101010
1.18	0.0000	5	11111111100000	1111010
0.45	0.0000	5	10010000100000	1101110
1.88	0.0000	9	11011000110000	1111110
0.37	0.0000	4	11111100111000	1101011
1.51	0.0000	7	10010010010100	1111011
0.68	0.0000	7	11011011011010	1101111
3.03	0.0000	13	11111111111111	1111111

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

From a cognitive perspective, what can we say about examinees who exhibit the observed response pattern (101000000000000) according to the mental models hierarchy? If the hierarchy proposed is accurate in describing how examinees reason about categorical syllogisms, we can certainly say that the examinees possess attribute A1 (interprets quantifiers logically) because this attribute is associated with an expected examinee response pattern that is logically included within the observed response pattern. The examinees, however, do not likely possess (likelihood of 0.0941) attribute A2 (can create an initial representation of the quantifiers), and it is highly unlikely that they possess (likelihood of 0.0008) attribute A3 (can generate logical conclusions from the initial mental representation). In other words, the examinees may have some surface knowledge or understanding of logical quantifiers, which allow them to respond correctly to items involving definitions of quantifiers. But the examinees may not have a substantive understanding of logical quantifiers to draw a valid conclusion.

Conclusions and Discussion

The purpose of this article was to introduce the AHM for cognitive assessment. The AHM illustrates a variation of Tatsuoka's rule-space approach (Tatsuoka, 1983, 1984, 1995) because it is designed to represent test performance in domains where a hierarchical ordering of attributes is fundamental to predicting and classifying examinee responses. We began by providing an overview of the AHM, focusing on the cognitive and psychometric components. Specifically, we described how an attribute hierarchy is formally represented, including the matrices that are derived from the hierarchy. Next, we described how expected examinee response patterns were generated from the Q_r matrix and the methods by which examinees' observed response patterns were 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 highlight limitations of the AHM, and suggest at least four lines of future research related to both the AHM and its application. The first line of future research involves the classification of observed response patterns. Methods A and B are used to classify an observed response pattern over a large sample of examinees; that is, the classification assumes that many examinees have produced the observed pattern in question. It is for this reason that very low likelihood estimates might be produced when an anomalous observed response pattern needs to be classified. Consider the observed response pattern (101000000000000) under the mental models hierarchy. As mentioned in the previous section, this observed pattern fails to match the expected examinee response pattern (11100000000000) at item 2 due to a single slip between the two patterns. Classifying the observed response pattern (101000000000000) into the expected examinee response pattern (111000000000000) might seem reasonable because the two patterns differ at only a single point. However, under Method A, the likelihood of numerous examinees producing the observed response pattern (101000000000000) from an expected response pattern (111000000000000) is highly unlikely if the mental models hierarchy is true.

As a result, we are currently investigating the classification of an observed response pattern for a single examinee using the AHM within an artificial neural network

architecture. One benefit of using neural networks is that they do not rely on strong assumptions about the distribution of examinees. Rather, they can be used to classify one examinee's observed response pattern into a single examinee attribute pattern by minimizing the error associated with the classification.

The second line of future research should focus on extracting information about cognitive competencies from the distracters as well as the keyed option (Thissen, Steinberg, & Fitzpatrick, 1989). Currently, attributes are associated with the cognitive competencies used by examinees to solve test items correctly (although K. Tatsuoka has conducted research to identify "incorrect rules" that lead to incorrect answers). However, the incorrect solutions may also yield meaningful information about attributes and cognitive competencies. 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.

A third line of future research should focus on attribute format issues. The attributes used in the AHM should promote the classification of cognitive competencies and tests. Scriven (1991) emphasized that a diagnosis always involves classifying the condition in terms of an accepted typology. In other words, attribute descriptions should evolve into an accepted typology for describing cognitive strengths and weaknesses and forms of diagnostic tests. 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, although attribute hierarchies may provide such a language.

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 strategies were apparent from the teachers' instructional 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 (Pellegrino et al., 1999). Much more work is needed to understand how examinees' attributes should be organized and presented to promote cognitive diagnosis and remediation, as well as guide future instruction (van Essen, 2001).

A fourth line of 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 method for assessing and reporting these attributes to examinees. This method of assessment—where examinees receive measures of their cognitive competencies 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, we intend to use this method to study 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 with this new method for cognitive assessment.

Notes

The authors thank Patrick Kyllonen, Robert Mislevy, W. Todd Rogers, and three anonymous reviewers for their comments on an earlier version of this article. This research was presented at the 2002 annual meeting of the National Council on Measurement in Education in New Orleans, LA.

This 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.

[2]A complete *Mathematica* (Wolfram, 1996) library with the algorithms for the procedures described in this article is available from the third author. The algorithms have been produced in *Mathematica* for research purposes only using information in the public domain. Readers, who are interested in a commercial package of similar algorithms, are encouraged to contact Tanar Software (Varadi & Tatsuoka, 1989, 1992) in the United States.

[3]To keep the item parameter estimates within a reasonable range, the BILOG parameters were adjusted to minimize the error sum of squares in fitting the expected item responses. Our adjusted item parameter estimates provided a better fit to the expected response matrix compared to the initial BILOG item parameter estimates. The ability parameters were then reestimated after adjusting the item parameters.

[4]The 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 item response pattern for item 1 (i.e., the first column of the expected response matrix in Table 3) is (011111111111111). Third, the ability estimates for the first and second rows of the expected response matrix 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.*

References

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17,* 37–45.

Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1993). Diagnosing knowledge states in algebra using the rule-space model. *Journal for Research in Mathematics Education, 24,* 442–459.

Birenbaum, M., & Shaw, D. J. (1985). Task specification chart: A key to a better understanding of test results. *Journal of Educational Measurement, 22,* 219–230.

Buck, G., & Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15,* 119–157.

Chipman, S. F., Nichols, P. D., & Brennan, R. L. (1995). Introduction. P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 1–18). Hillsdale, NJ: Erlbaum.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics.* Orlando, FL: Academic Press.

Embretson, S. E. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107–135). New York: Plenum Press.

Embretson, S. E. (1995a). Developments toward a cognitive design system for psychological testing. In D. Lupinsky & R. Dawis (Eds.), *Assessing individual differences in human behavior* (pp. 17–48). Palo Alto, CA: Davies-Black Publishing.

Embretson, S. E. (1995b). A measurement model for linking individual learning to processes and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement, 32,* 277–294.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380–396.

Embretson, S. E. (1999). Cognitive psychology applied to testing. In F. T. Durso, R. S. Nickerson, R. W., Schvaneveldt, S. T. Dumais, D. S. Lindsay, & M. T. H. Chi (Eds.), *Handbook of Applied Cognition* (pp. 629–660). New York: Wiley.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Hillsdale, NJ: Erlbaum.

Embretson, S. E., Schneider, L. M., & Roth, D. L. (1985). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement, 23,* 13–32.

Embretson, S. E., & Wetzel, D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11,* 175–193.

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, J. St. B. T., Newstead, S. E., & Byrne, R. M. (1993). *Human reasoning: The psychology of deduction.* Hillsdale, NJ: 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., Mislevy, 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.

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development.* Mahwah, NJ: Erlbaum.

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, NJ: Lawrence Erlbaum.

Kim, S. (2001). *Towards a statistical foundation in combining structures of decomposable graphical models* (Research Report No. 01-2). Yusong gu, Taejon: Korea Advanced Institute of Science and Technology, Division of Applied Mathematics.

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

Leighton, J., 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. (2003). Reasoning and problem solving. In A. F. Healy & R. W. Proctor (Eds.), *Experimental Psychology* (pp. 623–648). 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 Education, Macmillan.

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].* Moorseville, 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,* 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 Education, Macmillan.

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. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Fredrickson, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1991) *Boolean algebra applied to determination of universal set of knowledge states* (Tech. Rep. No. RR-91-44-ONR). Princeton, NJ: Educational Testing Service.

Tatsuoka, K. K. (1993). Item construction and psychometric models appropriate for constructed responses. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 107–133). Hillsdale, NJ: Erlbaum.

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, K. K., & Boodoo, G. (2000). Subgroup differences on the GRE quantitative test based on the underlying cognitive processes and knowledge. In D. Lesh & W. E. Kelly (Eds.), *Research design and methodologies for mathematics and science. Hillsdale, NJ: Erlbaum.*

Tatsuoka, M. M. (1986). Graph theory and its applications in educational research: A review and integration. *Review of Educational Research, 56,* 291–329.

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.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26,* 161–176.

van der Linden, W., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory.* New York: Springer.

van Essen, T. (2001, April). Developing and presenting enhanced skill descriptors for the PSAT/NMSQT. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Varadi, F., & Tatsuoka, K. K. (1989). *BUGLIB [Computer program].* Trenton, NJ: Tanar Software.

Varadi, F., & Tatsuoka, K. K. (1992). *BUGLIB Modified version* [Computer program]. Trenton, NJ: Tanar Software.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24,* 535–585.

Wolfram, S. (1996). *The mathematica book* (3rd ed.). Cambridge: Cambridge University Press.

Authors

JACQUELINE P. LEIGHTON is Assistant Professor, the Centre for Research in Applied Measurement and Evaluation, University of Alberta, 6-110 Education Centre North, Faculty of Education, Edmonton, Alberta, Canada, T6G 2G5; jacqueline.leighton@ualberta.ca. Her areas of specialization include cognitive psychology and educational assessment.

MARK J. GIERL is Associate Professor, the Centre for Research in Applied Measurement and Evaluation, University of Alberta, 6-110 Education Centre North, Faculty of Education, Edmonton, Alberta, Canada, T6G 2G5; mark.gierl@ualberta.ca. His areas of specialization include unidimensional and multidimensional IRT, differential item functioning, and cognitive assessment.

STEPHEN M. HUNKA is Professor Emeritus, the Centre for Research in Applied Measurement and Evaluation, University of Alberta, 6-110 Education Centre North, Faculty of Education, Edmonton, Alberta, Canada, T6G 2G5; steve.hunka@ualberta.ca. His areas of specialization include statistical models and computation.