Computer Adaptive-Attribute Testing

A New Approach to Cognitive Diagnostic Assessment

Mark J. Gierl and Jiawen Zhou

Centre for Research in Applied Measurement and Evaluation, University of Alberta, Canada

Abstract. The influence of interdisciplinary forces stemming from developments in cognitive science, mathematical statistics, educational psychology, and computing science are beginning to appear in educational and psychological assessment. Computer adaptive-attribute testing (CA-AT) is one example. The concepts and procedures in CA-AT can be found at the intersection between computer adaptive testing and cognitive diagnostic assessment. CA-AT allows us to fuse the administrative benefits of computer adaptive testing with the psychological benefits of cognitive diagnostic assessment to produce an innovative psychologically-based adaptive testing approach. We describe the concepts behind CA-AT as well as illustrate how it can be used to promote formative, computer-based, classroom assessment.

Keywords: cognition and assessment, cognitive diagnostic assessment, computer adaptive testing

Educational and psychological assessment is on the brink of profound change. Strong interdisciplinary forces stemming from rapid developments in areas such as cognitive science, mathematical statistics, educational psychology, and computing science are permeating the assessment field. For example, the influence of cognitive psychology on educational measurement, which began almost 20 years ago (Snow & Lohman, 1989), has recently become an important source of new ideas leading to innovations in cognitive diagnostic assessment (e.g., Leighton & Gierl, 2007a). One consequence of these interdisciplinary influences is the emergence of new areas of research. One area – which is the focus of this special issue – is on *adaptive models of psychological testing*. The research in this area combines the merits of adaptive testing with rapid developments in educational and cognitive psychology to produce assessment methods that promote psychologically sophisticated inferences about examinees' knowledge, skills, motivations, and competencies. In our paper, a new assessment method located at the interface between computer adaptive testing and cognitive diagnostic assessment is introduced, where the concepts and procedures are combined across these two areas to affect the development and administration of test items as well as the scoring and reporting of test results so specific diagnostic inferences can be made about examinees' cognitive skills and proficiencies.

Computerized adaptive testing (CAT) is an innovative form of assessment that matches the difficulty of a test item to the ability estimate for an examinee. The matching is accomplished by first presenting an examinee with an item of average difficulty and then, depending on the examinee's response, an item of greater or lesser difficulty is presented until the algorithms controlling item administration meet a specified stopping criterion. The merits of this approach for test administration are well documented (e.g., van der Linden & Glas, 2000). It leads to testing efficiency, as relatively fewer items are required to obtain a higher level of measurement precision and, hence, testing time is reduced, testing on-demand becomes feasible because each exam is uniquely tailored to each examinee, immediate test scoring and reporting becomes possible, and innovative item formats can be introduced into the assessment process because testing is computer based.

Cognitive diagnostic assessment (CDA) is a form of testing that employs a cognitive model to first develop or identify items that measure specific knowledge and skills and then uses this model to direct the psychometric analyses of the examinees' item response patterns to promote specific diagnostic inferences. A cognitive model in educational measurement refers to a "simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills students at different levels of learning have acquired and to facilitate the explanation and prediction of students' performance" (Leighton & Gierl, 2007b, p. 6). Cognitive models are generated by studying the knowledge, processes, and strategies used by examinees as they respond to items. The strength of developing test items and analyzing testing data according to a cognitive model stems from the detailed information that can be obtained about the knowledge structures and processing skills that produce examinees' test score.

In 2004, Leighton, Gierl, and Hunka introduced a procedure for cognitive diagnostic assessment called the *attribute hierarchy method* (AHM). The AHM is a psychometric method used to classify examinees' test item responses into a set of structured attribute patterns associated with different components from a cognitive model of task performance. Attributes include different procedures, skills, and/or processes that an examinee must possess to

DOI 10.1027/0044-3409.216.1.29

solve a test item. These attributes are structured using a hierarchy so the ordering of the cognitive skills is specified. As a result, the attribute hierarchy serves as an explicit construct-centered cognitive model. This model, in turn, provides a framework for designing test items and for linking examinees' test performance to specific inferences about psychological skill acquisition. The purpose of the current study is to integrate CAT and AHM and propose a concept of *computer adaptive-attribute testing* (CA-AT). We begin with an overview of the AHM. Then, we discuss how the CAT concepts of banking, routing, and score reporting can be applied to cognitive diagnostic assessment with the AHM. Finally, we conclude with a summary and we identify two areas requiring additional research.

Overview of the Attribute Hierarchy Method

The attribute hierarchy method (Leighton, Gierl, & Hunka, 2004; Gierl, Cui, & Hunka, 2007), a cognitive diagnostic procedure that evolved from the rule space approach (Tatsuoka, 1983, 1995; see also Gierl, 2007), is used to classify examinees' test item responses into attribute patterns associated with a cognitive model of task performance. The AHM was developed to address specific issues associated with *cognitive model development* and *statistical pattern recognition*. We use these two areas to illustrate our CDA method. But first, we present an example in the domain of high school algebra to help illustrate our approach.

Gierl, Wang, and Zhou (2008a) and Gierl, Leighton et al. (2008) used algebra items from the March 2005 administration of the SAT to develop diagnostic models of algebra performance. The SAT is one of the most widely-used college admissions test in the world. The Mathematics section contains items in the content areas of *number and operations*, *algebra I, II, and functions*; *geometry*; and *statistics*, *probability, and data analysis*. For our analysis, a subset of items in algebra I and II were evaluated. Cognitive models of task performance guide diagnostic inferences because they are specified at a small grain size and they magnify the knowledge and skills that underlie performance. Ideally, a theory of task performance would direct the development of a cognitive model. But, in the absence of such a theory, a cognitive model must still be specified to create the attribute hierarchy. Another starting point is to develop a cognitive model from a task analysis of the items in the domain when a theory or model of task performance is unavailable. In conducting the task analysis of the SAT algebra items we, first, solved each test item and attempted to identify the mathematical concepts, operations, procedures, and strategies used to solve each item. We then categorized these cognitive attributes so they could be ordered in a logical, hierarchical sequence to summarize problem-solving performance. The models were validated and, in some cases, modified, using results from verbal protocol analyses. One of the final attribute hierarchies for the algebra cognitive models is presented in Figure 1. Each attribute is denoted with an A (e.g., A1, A2, etc.).

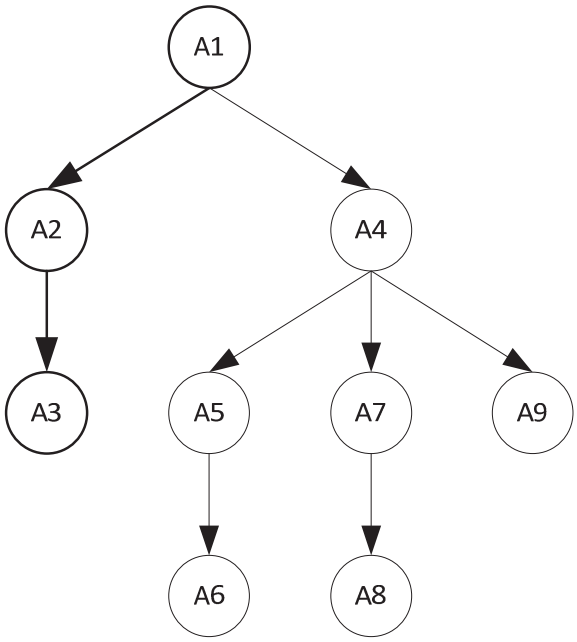

Figure 1. An attribute hierarchy for algebra performance.

This hierarchy represents a cognitive model of task performance for skills in the areas of ratio, factoring, function, and substitution. The hierarchy contains two independent branches which share a common prerequisite, attribute A1. Aside from attribute A1, the first branch includes two additional attributes, A2 and A3, and the second branch includes a self-contained subhierarchy which includes attributes A4 through A9. The order relations among the attributes in the hierarchy are defined psychologically, meaning that the skills required to solve each item are identified and then ordered from the least to most complex. The attributes are also assumed to share dependencies as they function within a much larger network of inter-related processes, competencies, and skills that characterized human information processing (Anderson, 1996; Dawson, 1998, 2004; Fodor, 1983).

For example, attribute A1 includes the most basic arithmetic operation skills, such as addition, subtraction, multiplication, and division of numbers. Attributes A2 and A3 both deal with factors. In attribute A2, the examinee needs to have the basic arithmetic skills (i.e., attribute A1), as well as knowledge about the property of factors. In attribute A3, the examinee not only requires basic arithmetic skills (i.e., attribute A1) and knowledge of factoring (i.e., attribute A2), but also the skills required for the application of factoring. Therefore, attribute A3 is considered a more advanced attribute than A1 and A2.

The self-contained subhierarchy contains six attributes.

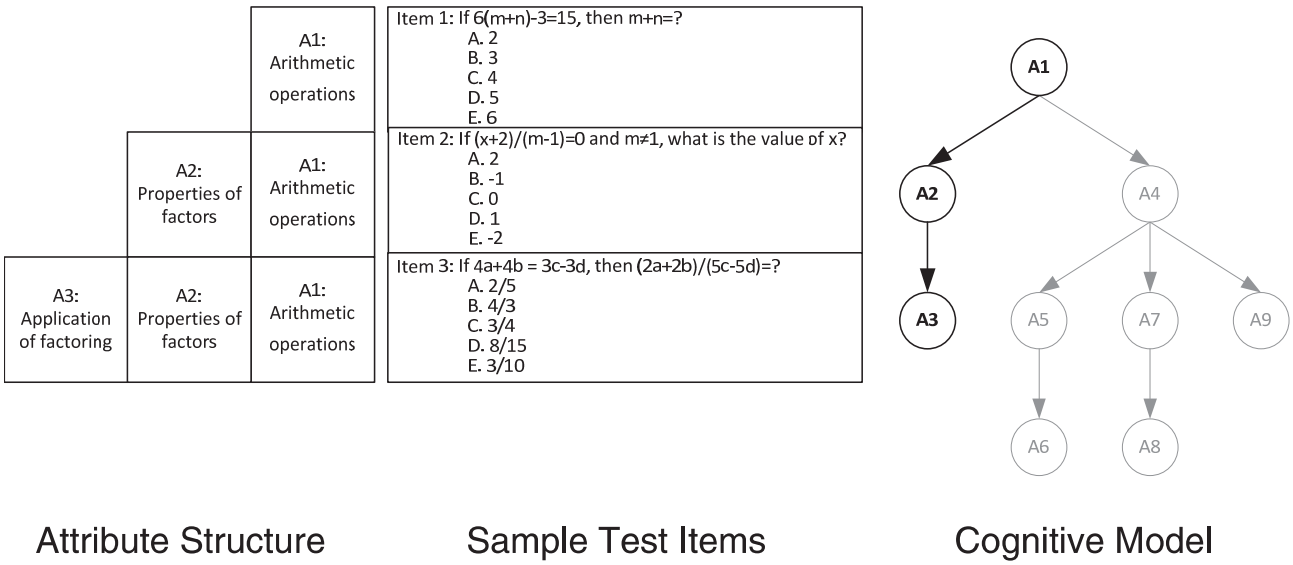

Figure 2. The attribute structure, the associated test items, and the cognitive model of algebra performance.

Among these attributes, attribute A4 is the prerequisites for all other attributes in the subhierarchy. Attribute A4 has attribute A1 as a prerequisite because A4 not only represents basic skills in arithmetic operations (i.e., attribute A1), but it also involves the substitution of values into algebraic expressions. The first branch in the subhierarchy deals, mainly, with functional graph reading. For attribute A5, the examinee must be able to map the graph of a familiar function (e.g., a parabola) with its corresponding function. Attribute A6 deals with the abstract properties of functions, such as recognizing the graphical representation of the relationship between independent and dependent variables. The second branch in the subhierarchy considers the skills associated with advanced substitution. Attribute A7 requires the examinee to substitute numbers into algebraic expressions. The complexity of attribute A7 relative to attribute A1 and A4 lies in the concurrent management of multiple pairs of numbers and multiple equations. Attribute A8 also represents the skills of substitution. However, what makes attribute A8 more difficult than attribute A7 is that algebraic expressions, rather than numbers, need to be substituted into another algebraic expression. The last branch in the subhierarchy contains only one additional attribute, A9, related to skills associated with rule substitution.

Stage 1: Development of the Cognitive Model

In Stage 1, the relationships among the attributes in the hierarchy are specified using the adjacency and reachability matrices. The direct relationship among attributes is specified by a binary *adjacency matrix* (A) of order (*k, k*), where *k* is the number of attributes, such that each element in the A matrix represents the absence (i.e., 0) or presence (i.e., 1) of a direct connection between two attributes. To illustrate these concepts, we will focus on one branch in the Figure 1 hierarchy, which represents the relationship among attributes A1 to A3 (the complete set of matrices outlined in Hierarchy 1 are presented in Gierl, Wang et al., 2008b). Three sample items help illustrate the hierarchical nature of A1, A2, and A3.[1] The results are presented in Figure 2. Recall, attribute A1 includes the most basic arithmetic operation skills, such as addition, subtraction, multiplication, and division of numbers, whereas attributes A2 and A3 both deal with factors. The adjacency matrix for the hierarchy in Figure 2 is shown in matrix 1.

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

In matrix 1, attribute A1 is prerequisite to attribute A2 and A3. This hierarchical relationship is expressed in the first and second row by the positions of a 1 in columns two and three. The positions of 0 in row 1 indicate that A1 is neither directly connected to itself nor to A3, but to A2. The 1 in row 2 indicates that A2 is directly connected to A3. The direct and indirect relationships among attributes are specified by the binary *reachability matrix* (R) of order (*k, k*), where *k* is the number of attributes. To obtain the R matrix from the A matrix, Boolean addition and multiplication operations are performed on the adjacency matrix, meaning $R = (A + I)^n$, where *n* is the integer required to reach invari-

[1] These items were not used on an operational SAT and not developed by the College Entrance Examination Board. Rather, we developed these illustrative items for the current study to highlight the key algebraic concepts attributes A1, A2, and A3 are intended to measure.

ance, $n = 1,2. . .m$, and l is the identity matrix. The R matrix for the hierarchy in Figure 2 is shown in matrix 2.

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

The R matrix is interpreted in a similar manner to the A matrix. The first row of the matrix 2 indicates A1 reaches all other attributes, either directly or indirectly, because of the positions of a 1 in columns 1, 2, and 3. Row 2 of this matrix indicates that A2 reaches A3.

The incidence matrix (Q) is generated next. Q includes those items representing all possible combinations of attributes, when the attributes are considered independent of one other. Q is of order (k, p), where k is the number of attributes and p is the number of possible items. This matrix can be reduced, often substantially, to form the *reduced Q matrix* (Q_r) by imposing the constraints of the hierarchy as defined in the R matrix. The Q_r matrix represents only those items that fit the dependencies defined in the attribute hierarchy. The Q_r matrix is formed using Boolean inclusion by determining which columns of the R matrix are logically included in each column of the Q matrix. The Q_r matrix is of order (k, i) where k is the number of attributes and i is the reduced number of items resulting from the constraints in the hierarchy. Q, for our example in Figure 2, is shown in matrix 3. Then, with the constraints imposed by the hierarchy defined in the R matrix, Q_r is produced from Q, as shown in matrix 4.

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

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

Principled test design is used explicitly with the AHM to create items and analyze examinees' observed response patterns. To design test items, the Q_r matrix is used. The Q_r matrix can be interpreted as the *cognitive test specification* because it contains the attribute-by-item specification for each component of the cognitive model of task performance outlined in the A matrix. Hence, the results from the Q_r matrix are used to develop items that measure each specific attribute combination defined in the hierarchy. In the current example, at least three items are required: An item measuring A1; an item measuring A1 and A2; and an item measuring A1, A2, and A3. To increase attribute reliability[2], items measuring each attribute combination must be developed (this concept will be explained and illustrated in the upcoming section on *item banking*).

As a final step in the cognitive model development stage, the expected response patterns are generated. The *expected response matrix* (E) is created, again using Boolean inclusion, where the algorithm compares each row of the attribute pattern matrix (which is the transpose of the Q_r matrix) to the columns of the Q_r matrix. The expected response matrix is of order (j, i), where j is the number of examinees and i is the reduced number of items resulting from the constraints imposed by the hierarchy. The expected response matrix for the hierarchy in Figure 2 is shown below as matrix 5.

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

This result reveals that three unique item response patterns should be produced by examinees who write the cognitively-based items, resulting in 3 possible total scores (i.e., 1, 2, or 3)[3], if the cognitive model is true.

Stage 2: Statistical Pattern Recognition

In Stage 2, an examinee's observed response pattern is judged relative to an expected response pattern under the assumption that the cognitive model is true. Hence, the purpose of the statistical pattern recognition stage is to identify the attribute combinations that the examinee is likely to possess. To estimate the probability that examinees possess specific attributes, both IRT and non-IRT classification methods are available. The IRT methods are described in Leighton et al. (2004), and in Gierl, Leighton, and Hunka (2007). Our focus in this paper is on our more recent classification methods associated with the non-IRT based neural network, which is used to estimate the *attribute probabilities*, as described in Gierl, Cui et al. (2007).

A neural network is a type of parallel-processing architecture that transforms a stimulus received by an input unit to a signal for the output unit through a series of hidden units. To begin, each cell of the input layer receives a value (0 or 1) corresponding to the response values in the exemplar vector. Each input cell then passes the value it receives to every hidden cell. Each hidden cell forms a linearly weighted sum of its input and transforms the sum using the logistic function and passes the result to every output cell. Each output cell, in turn, forms a linearly weighted sum of its inputs from the hidden cells and transforms it using the

[2] Attribute reliability refers to the consistency of the decisions made in a diagnostic test about examinees' mastery of specific attributes. One method for estimating the reliability of an attribute is to calculate the ratio of true score variance to observed score variance on the items that are probing *each attribute*. To isolate the contribution of each attribute to an examinee's item-level performance, the item score is weighted by the subtraction of two conditional probabilities. The first probability is associated with attribute mastery (i.e., an examinee who has mastered the attribute can answer the item correctly) and the second probability is associated with attribute nonmastery (i.e., an examinee who has not mastered the attribute can answer the item correctly). The weighted scores for items that measure the attribute are then used in the reliability calculation (see Gierl, Cui et al., 2007, for details).

[3] 0s are often added to row 1 of the E matrix to allow for a possible total score of 0.

logistic function, and outputs the result. The network learns how to make these transformations from input to hidden to output layer using a set of training data. The neural network, therefore, serves as a powerful pattern recognition technique because, given enough training data and practice, it can map any relationship between input and output (see Dawson, 1998, 2004).

The input to train the neural network for a diagnostic analysis using the AHM is the set of expected response vectors presented in matrix 5. Each expected response vector is associated with a specific combination of examinee attributes. The expected response vectors, used as inputs, are meaningful because they are derived from the cognitive model. The output are the specific attribute patterns associated with each expected response pattern. The attribute pattern matrix contains each attribute vector, where examinees' mastery or nonmastery of individual attributes is specified. The attribute pattern matrix, like the expected response matrix, is derived from the cognitive model where examinee responses are explained by the presence or absence of the attributes without any errors or "slips." As a result, these attribute vectors contain only 1s indicating that the examinee has mastered the corresponding attributes and 0s indicating that the examinee has not.

However, in most testing situation, examinees' observed response vectors are not consistent with the expected response vectors because of slips from 1 to 0 or 0 to 1 leading to uncertainty in the value of the attribute probabilities. Slips may occur for different reasons, including: the attributes were not accurately identified; the attribute hierarchy specified is inappropriate; the test items do not measure the attributes in the hierarchy; the test was inappropriate for the student sample; and/or the students produce random responses (e.g., guessing on a multiple-choice item). Therefore, to estimate the attribute probabilities associated with each observed response vector, the relationship between the expected response vectors with their associated attribute vectors is established by presenting each expected response vector to the network repeatedly until it learns each association. The final result is a set of weight matrices that can be used to transform *any response vector* – expected or observed – to its associated attribute vectors. The transformed result can be interpreted as the attribute probability, scaled between 0 to 1, where a higher value indicates that the examinee has a higher probability of possessing a specific attribute (McClelland, 1998). These attribute probabilities provide specific cognitive model-based feedback for each examinee which will be instrumental in the diagnostic reporting process.

To summarize, diagnostic assessment in the AHM proceeds as a two-stage process. In the cognitive model development stage, the model is initially developed and then represented using the A, R, Q, Q_r, and E matrices. The AHM maintains the logical basis outlined in the cognitive model through the A matrix to produce the expected response matrix, thereby allowing the psychometrician to represent complex hierarchical relationships and evaluate a wide range of cognitive models. Items on the test are also developed to specifically measure each attribute combination represented in the Q_r matrix. Then, in the statistical pattern classification stage, the examinee's observed response pattern is judged relative to expected response pattern under the assumption that the cognitive model is true. The purpose of the statistical pattern recognition stage is to identify the attribute combinations that the examinee is likely to possess, and then to report this specific information back to examinees so they can make inferences about their mastery of different cognitive skills.

Overview of Computer Adaptive-Attribute Testing

Computer adaptive-attribute testing combines important characteristics in computer adaptive testing with cognitive diagnostic assessment. CA-AT could be applied in many testing situations. But one of the most promising applications is in the area of education where CA-AT could support learning and instruction in a *formative, computer-based, classroom assessment system*. This type of system has some unique characteristics: It is implemented periodically during the teaching and learning process, often at the discretion of the teacher, meaning the student can be tested at any time and on more than one occasion; the assessment outcomes are intended to guide teaching and learning therefore the content in the assessment should be closely linked to the curriculum which, in turn, is tied to each lesson or instructional unit; the assessment tends to be relatively low stake; the assessment outcomes support decisions that are direct, specific, and immediate, such as deciding on a student's homework assignment or guiding the development of a teacher's next lesson; and the assessment is scored automatically thereby providing immediate feedback for students and teachers.

The testing process for a formative, computer-based, classroom CA-AT could work as follows:

1) The teacher selects the appropriate assessment unit (e.g., ratios and algebra) based on the curriculum standards and the instructional unit of interest;
2) the student logs onto a computer and accesses the CA-AT application through an internet browser in a classroom computer, on a laptop, or in a designated computer lab;
3) the student is presented with items measuring the first attribute in the cognitive model (e.g., ratios and algebra);
4) based on the student's response to the items measuring the initial attribute, the student is either administered items measuring a more complex attribute (e.g., if the student correctly solves the items measuring attribute A1, then the student will receive items measuring attribute A2 and A4 in the Figure 1 hierarchy), or attribute-based item administration is stopped (e.g., if the student

incorrectly solves the items measuring attribute A2, then the student will not receive items measuring attribute A3 in the Figure 1);

5) this process of attribute-based item administration is used to systematically probe every branch at every level in the hierarchy until the stopping rule for each branch is satisfied; and
6) when the test is complete, the system computes the examinee's scores and presents the result in a report on the computer screen.

Two additional considerations must be made in the CA-AT process. In step 3, the number of items administered per attribute must be determined. One method is to establish attribute reliability requirements. For example, if the goal is to achieve an attribute reliability level of 0.70, then the examinee must take the sufficient number of items required to produce this reliability. In step 4, the decision to stop or continue is determined by the examinees' attribute mastery level. Attribute mastery can be determined using a criterion of performance established using a specific outcome level (e.g., an 80% level of correct performance) or using more analytic procedures such as some type of attribute-based standard setting.

In the next section of the paper, we describe the implications of CA-AT formative, computer-based, classroom assessment on banking, routing, and score reporting.

Item Banking

CA-AT, like any adaptive testing system, makes heavy demands on the items in the bank. Fortunately, one key benefit of using the AHM for diagnostic assessment lies in its facility to guide item development. The Q_r matrix is of order (*k, i*) where *k* is the number of attributes and *i* is the reduced number of items resulting from the constraints in the hierarchy. In other words, Q_r serves as the cognitive test specifications. Hence, items can be developed systematically to measure each attribute in the cognitive model. For example, three items are required to measure A1 to A3 in Figure 2. Item 1 measures attribute A1, which includes basic arithmetic operation skills; Item 2 measures attribute A2, which includes knowledge about the properties of factors in addition to basic arithmetic operation skills (i.e., A1); Item 3 measures attribute A3, which includes the application of factoring in addition to the properties of factors (i.e., A2) and basic arithmetic operation skills (i.e., A1). That is, items are developed according to specific hierarchical ordering of attributes, as outlined in the Q_r matrix, which are designed to measure cognitive processes of increasing complexity (see matrix 4).

In addition to creating each attribute-specific item, multiple instances of these items must also be created to produce a large functional bank for continuous testing. Research on item modeling and automated item generation can support this process (e.g., LaDuca, Staples, Templeton, & Holzmann, 1986; Bejar, 1996; Bejar, Lawless, Morley, Wagner, Bennett et al., 2003). Traditional item development using manual processes can be inefficient, largely because items are treated as isolated entities that are individually developed, formatted, and evaluated. Item modeling is an alternative method for developing *classes of isomorphic items* from shells (Haladyna & Shindoll, 1989) or templates (Case & Swanson, 2002) that are intended to elicit similar, if not identical, responses from examinees. An item model produces large numbers of *instances* or items of that model and these instances are designed to function similarly on tests. Item modeling also promotes efficiency because each model can automate many aspects of the item development process. For example, each model specifies a set of variables and constraints that, when manipulated, produce multiple isomorphic instances of the model. Hence, test developers focus on creating item models and specifying the variables and constraints for these models rather than creating single, unique and, often, noninterchangeable items to satisfy the requirements for the exam. Developing an item model, therefore, serves as the first step in providing a detailed description of each item class, particularly when the Q_r matrix can be used to specify the cognitive structure for the required items. The outcomes from these activities will provide developers with the foundation and initial tools required to begin to implement principled test design processes.

An example of an item model that measures attribute A1 is presented in Figure 3. The model contains five sections. Section 1 (Item Sample for Attribute A1) includes an exemplary instance of an operational item. Section 2 (Stem) presents the item stem as a model where the manipulated variables (i.e., V_1, V_2, V_3) are specified. Section 3 (Variables) outlines the content or value variations from the manipulated variables in the stem. The content for each variable can be an incidental or radical item characteristic. An *incidental* or surface characteristics of the item includes the features that are not expected to affect the psychometric characteristics of the item, such as the difficulty level or dimensionality. The *radical* or deep characteristics of the item includes the features that could change the psychometric properties of the item. Because the goal is to generate a large item bank to measure each attribute in the hierarchy, isomorphic instances for each attribute are required. Therefore, each manipulated variable should act as an incidental item characteristic. Section 4 (Constraints) presents the algorithmic method for producing the correct answer and incorrect alternatives. Section 5 (Key) identifies the correct answer. In the current example, 192 (3 V_1 * 4 V_2 * 16 V_3) unique items could be generated using the attribute A1 model.

As a final point, the number of items in the bank should support the administrative demands prescribed by the structure of the attribute hierarchy. Because CA-AT requires many isomorphic items to measure each attribute, the items in the bank must be organized according to the structure of the hierarchy. Provisions must also be made for

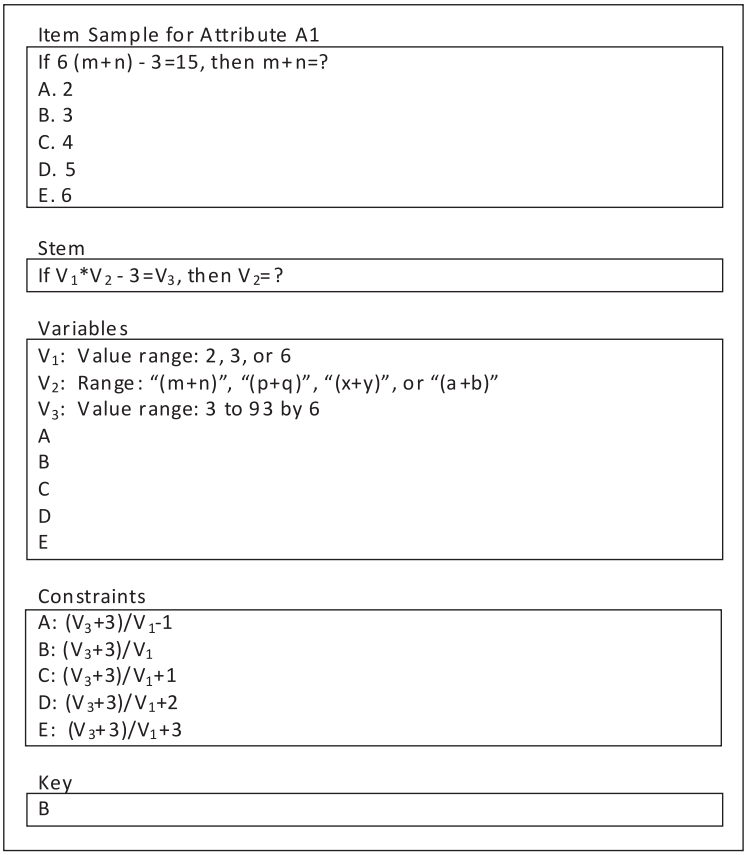

Figure 3. An item model for measuring Attribute A1.

the starting rules in the adaptive test. For example, if A1 is the starting attribute, then more isomorphic items measuring A1 are required in the bank because this attribute will be administered and, thus, exposed frequently. Similarly, if A3 is a difficult cognitive skill, then fewer isomorphic items measuring A3 are required in the bank because items measuring this attribute will be administered and exposed comparatively less than those for an easier cognitive skill or a starting attribute such as A1.

Item Routing

With the item bank in-place, routing rules for adaptive administration must be established next. As with item development, routing in a CA-AT is governed by the structure of the attribute hierarchy. For a conventional CAT, routing is determined in real-time by considering the examinees' response to the previous item, given the examinee's provisional ability estimate, the psychometric properties for the remaining items in the bank, and other content constraints (e.g., item exposure rules). For a CA-AT, routing is determined by the structure (i.e., paths and levels) in the attribute hierarchy. The algebra hierarchy in Figure 1 helps illustrate the CA-AT routing concept. If the starting point is A1, then examinees are administered items measuring this attribute. Examinees are administered items measuring the next attribute or attributes, only when they answer the items measuring A1 correctly. For example, when the examinee correctly answers items measuring attribute A1 (basic arithmetic operation skills), items measuring both A2 (knowledge about the properties of factors) and A4 (the substitution of values into algebraic expressions) will be administrated. If an examinee answer items measuring A2 incorrectly, then the administration on the left branch stops and the examinee is not permitted to answer items measuring A3 (applying the rules of factoring) because the examinee does not possess the prerequisite skills A2. Similarly, if the examinee correctly answers items measuring attribute A4, then items measuring A5, A7, and A9 can be administered. This routing approach, where each branch and layer of the hierarchy are systematically probed, provides a specific guide for item administration.

Zhou, Gierl, and Cui (2007) recently conducted a feasibility study to determine the number of items to administer on a CA-AT under different testing conditions to achieve satisfactory attribute reliability. The nine-attribute hierarchy presented in Figure 1 was used. Two variables were manipulated in their simulation study: the number of items measuring each attribute and the slip level. The number of items measuring each attribute was two, three, or four in each condition. Item responses of 5000 examinees to the three testlets were simulated using the 3-PL multidimensional item response theory model. Also, response slips

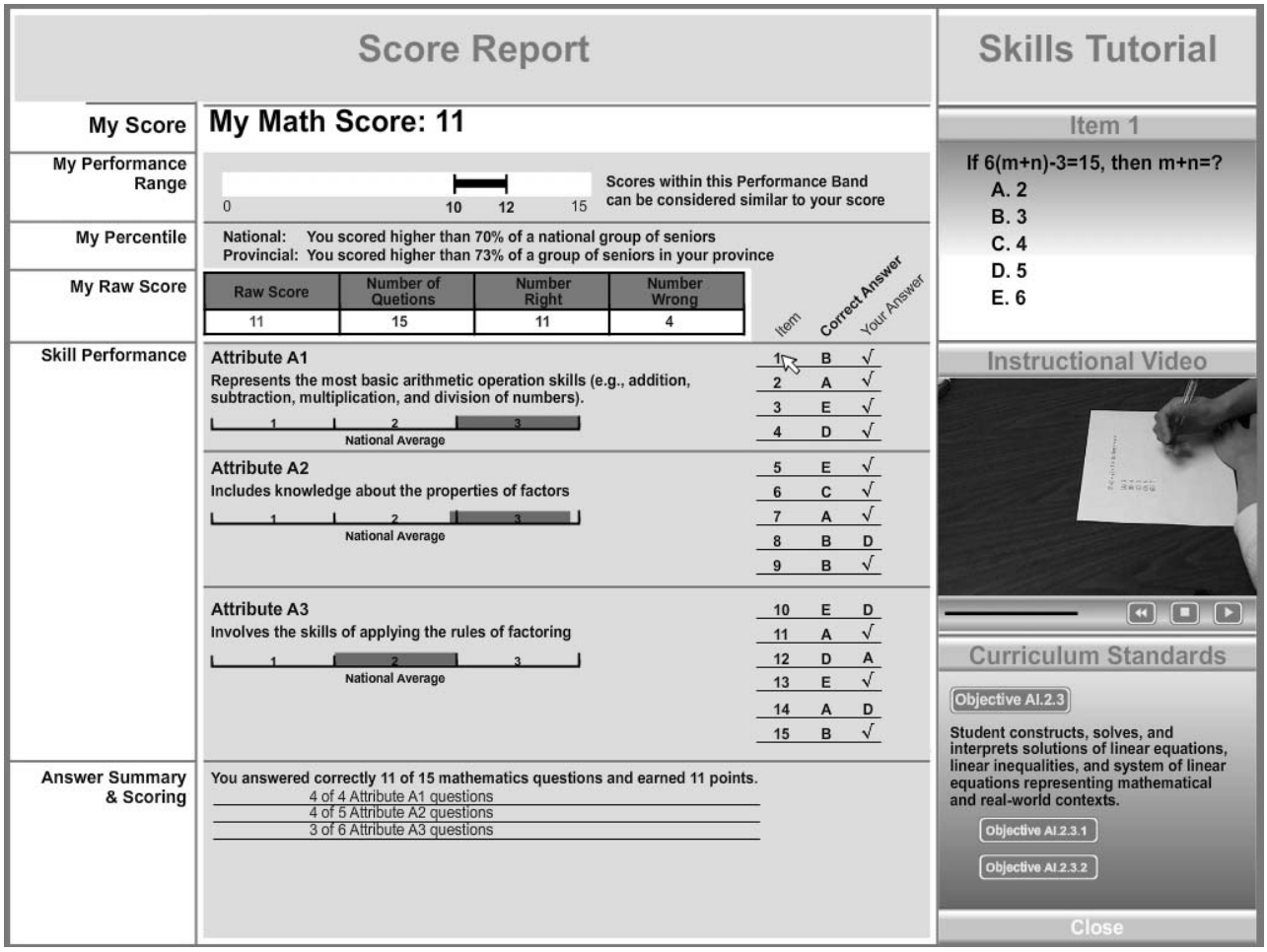

Figure 4. The screenshot for a computer-based cognitive diagnostic score report.

were generated and placed within the item response data to simulate a real testing situation. A slip is the discrepancy between the observed response pattern and the expected response pattern (e.g., a 1 was expected, but a 0 was observed or a 0 was expected, but a 1 was observed) based on the structure of the hierarchy. Three percentages of slips, 5%, 10%, and 15%, were added to the expected response patterns. For example, in the 5% condition, one slip in the form of either 0 to 1 or 1 to 0 was added to each of the 250 randomly selected observed response patterns.

The results of the simulation study reveal that attribute reliability is affected both by the number of items measuring each attribute and by the slip percentage. Not surprisingly, attributes measured by only two items yield a lower reliability estimate than attributes measured by three or four items. Also, when the same number of items were used to measure each attribute, the reliability decreases as the percentage level of slip increases. For the hierarchy in Figure 1, attribute reliability was acceptable (a value of 0.70 or higher was adopted to define an acceptable reliability level) when three or more items were used to measure each attribute as long as the slip percentage was no greater than 10%. From these results we conclude that at least three items should be administered for each attribute and that the model-data fit should be relatively strong (i.e., slip percentage should not exceed 10%) for the cognitive model of interest.

Score Reporting

With an operational item bank and appropriate routing strategy, examinees can be administered items in the CA-AT system. Once examinees complete the test, their data are scored and their reports are generated. Because CA-AT is a mode of assessment guided by a cognitive model of task performance, items directly measure specific cognitive attributes of increasing complexity. As a result, both a total test score and diagnostic subscores can be produced for each examinee. The potential benefits of diagnostic score reporting also increase, as we will illustrate, with computer-based administration. One example is presented in Figure 4. This sample report is limited to descriptions for our three attribute example – the complete report to accompany the cognitive model in Figure 1 would contain nine attributes. This reporting scheme is also computer

based. But, for the purposes of this paper, only a screen shot is presented.

Our sample report contains two sections. Section 1, on the left side of the page, titled "Score Report" contains the examinees' total score (i.e., My Score; My Performance Range; My Percentile; My Raw Score)[4] and their diagnostic subscores (i.e., Skill Performance; Answer Summary & Scoring). The total score is the composite score of all items administered in the test. The examinee's performance range represents the confidence interval of the examinee's total score. The examinee's percentile positions the student's performance relative to a variety of standards, such as national or provincial outcomes. In addition to these traditional test score results, the report also provides information about examinees' mastery of each attribute. Examinees can identify their strengths and weaknesses for each attribute measured on the test. The location of each attribute bar is determined by the attribute probability (see description in Stage 2: Statistical Pattern Recognition) while the length of each bar is determined by the attribute reliability. For ease of interpretation, attribute probabilities are only reported in three categories (1–0% to 33%; 2–34% to 66%; 3–67% to 100%)–other reporting conventions could also be used. Examinees also receive a summary of their item-by-attribute performance at the bottom of the page.

The unique information that helps link assessment with instruction begins in the middle of the score report. Each examinee is presented with an interactive attribute-by-item summary. That is, the "Item," "Correct Answer," and "Your Answer" columns in the middle of the report can be used to activate different content in Section 2, "Skills Tutorial." To begin, the examinee uses the cursor on the computer screen to select an item. For the current example, the examinee has selected Item 1 (i.e., the white pointer arrow is on Item 1 in Figure 4). Once the item is selected, Section 2 of the score report becomes active. In this section, the item used to measure the attribute is presented to the examinee, along with an instructional video where a "model student" describes how the item is solved and a link to the curriculum standard for the item. The instructional videos and curriculum standards allow the examinee as well as teachers, parents, tutors, or other stakeholders to immediately see how the item can be solved and which objective in the curriculum is being evaluated.

The diagnostic feedback provided in this report is intended to promote learning and instruction in two ways. First, the report supports strong inferences about examinees' cognitive skills because the attribute grain size for reporting is small, thus illuminating the knowledge and skills required to perform competently on each task. The hierarchy also facilitates these diagnostic inferences because the attributes and their organization are based on a cognitive model of task performance and items are explicitly designed to measure the attributes. Second, scores are directly linked to items, instructional videos, and curricular standards. The link is established in real time due to the computer-based administration thereby allowing the reporting resources to be dynamic, modifiable, and easily updated. These resources also provide a way to link cognition, learning, and assessment for students, teachers, and parents because the test items are closely aligned to cognitive skills, instructional resources, and curricular outcomes in a single, succinct, dynamic, on-line report. Once testing is complete, the examinee could also access the report at any time, at any location, and on multiple occasions.

Summary

The purpose of the study was to introduce a new type of adaptive model for psychological testing. Computer adaptive-attribute testing (CA-AT) combines the administrative benefits of computer adaptive testing with the psychological benefits of cognitive diagnostic assessment to produce a new psychologically-based adaptive testing procedure. The *attribute testing component* in CA-AT reside with the attribute hierarchy method, which is a psychometric procedure for classifying examinees' test item responses into a set of attribute patterns associated with a cognitive model of task performance. An attribute is a description of the procedural or declarative knowledge needed to perform a task in a specific domain, and these attributes form a hierarchy that define the psychological ordering of problem-solving skills. The model is foundational in our approach to cognitive diagnostic assessment because it provides an interpretative framework that can guide the development of items and the analyses of examinees' item responses so test performance can be linked to specific cognitive inferences about examinees' knowledge and skills. The *computer adaptive component* supports the cognitive diagnostic model by allowing for new and expanding on existing test administration options, including attribute-adaptive item selection, continuous testing, immediate scoring, and dynamic computer-based reporting.

Directions for Future Research

Adaptive psychological testing, as a new interdisciplinary area of research in educational and psychological assessment, may come to rely on other emerging concepts and ideas, most notably assessment engineering. *Assessment engineering* (Luecht, 2006a, 2006b; Luecht, Gierl, Tan, & Huff, 2006) is a research area where principled test design concepts are used to direct the development of tests as well as the analysis, scoring, and reporting of test results. With

[4] We use the raw score in this example to make the scoring and reporting process transparent. In an operational adaptive test, the raw score would be transformed to a scale score because examinees receive different sets of test items.

this approach, the assessment specialist begins by defining the construct of interest using specific, empirically-derived cognitive models of task performance. Next, assessment task templates are created from cognitive models to produce replicable assessment tasks. Finally, confirmatory psychometric models are applied to the examinee response data collected using the templates to produce scores that are both replicable and interpretable. Assessment engineering relies on a cognitive model, of some type, to develop items and analyze examinee item response data, generate scores, and guide score interpretations. In fact, when the goal is to produce specific diagnostic inferences about examinees, some form of assessment engineering is required because the assessment must be designed to identify and evaluate the examinees' cognitive skills. Often, this type of specificity cannot be obtained using a *posthoc* or *retrofitting* approach to test design (e.g., coding existing items for cognitive attributes) because items with these specific cognitive characteristics are unlikely to exist on a test developed without a cognitive model. Moreover, when a cognitive analysis of an existing test using retrofitting procedures is conducted, the fit between the cognitive model and the test data will invariably be tenuous leading to weak and, possibly, inaccurate diagnostic inferences.

Thus, to overcome the limitations with retrofitting, a more principled approach to test design and analysis is required. CA-AT requires that items be developed systematically to measure each component in the cognitive model using the Q_r matrix. Then, attribute-based item models are used to create multiple isomorphic instances to populate the bank. This approach to test development and item banking is essential for ensuring that cognitive principles are closely aligned with test design practices, thereby providing a direct connection between cognitive theory and educational measurement. Unfortunately, assessment engineering is currently not a common method for designing educational or psychological assessments and it has never been used, to our knowledge, in an operational testing situation, adaptive or otherwise.

CA-AT also relies on cognitive models to promote inferences about examinees' problem-solving skills. These models provide the framework necessary to guide item development and direct the psychometric analyses. Hence, cognitive models must be developed for each CA-AT content area and grade level. These models must also be validated to ensure they support specific diagnostic inferences, given the strong AHM assumption that the underlying cognitive model accurately describes examinees' observed item response patterns. Currently, we lack a collection of cognitive models, specifically, and psychological theories, more generally, on aptitudes and achievements that can serve educational and psychological assessment in a broad and useful manner. While some models and theories exist to guide diagnostic inferences (see examples presented in Mislevy, 2006; Yang & Embretson, 2007), research is required on a range of educational and psychological constructs in different domains and at different grade levels before models to support CA-AT applications become more widely available.

Acknowledgments

The research reported in this study was conducted, in part, with funds provided to the first author by the College Entrance Examination Board. We would like to thank the College Entrance Examination Board for their support. However, the authors are solely responsible for the ideas, methods, procedures, and interpretations expressed in this study. Our views do not necessarily reflect those of the College Entrance Examination Board. We would also like to thank Drs. Steve Hunka and Wim J. van der Linden for their comments on an earlier version of this manuscript.

References

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

Bejar, I.I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium* (ETS Research Report 96–13). Princeton, NJ: Educational Testing Service.

Bejar, I.I., Lawless, R., Morley, M.E., Wagner, M.E., Bennett, R.E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment, (2)3*. Retrieved on July 8, 2006, from http://www.jtla.org

Case, S.M., & Swanson, D.B. (2002). *Constructing written test questions for the basic and clinical sciences* (3rd ed., revised). Philadelphia, PA: National Board of Medical Examiners.

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

Dawson, M.R.W. (2004). *Minds and machines: Connectionism and psychological modeling.* Malden, MA: Blackwell.

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

Gierl, M.J. (2007). Making diagnostic inferences about cognitive attributes using the rule space model and attribute hierarchy method. *Journal of Educational Measurement, 44,* 325–340.

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

Gierl, M.J., Leighton, J.P., & Hunka, S. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J.P. Leighton, & M.J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 242–274). Cambridge, UK: Cambridge University Press.

Gierl, M.J., Leighton, J.P., Wang, C., Zhou, J., Gokiert, R., & Tan, A. (2008). *Developing and validating cognitive models of algebra performance on the SAT®* (Research report). New York: The College Board.

Gierl, M.J., Wang, C., & Zhou, J. (2008a). Using the attribute hierarchy method to make diagnostic inferences about exami-

nees' cognitive skills in algebra on the SAT®. *Journal of Technology, Learning, and Assessment, 6*(6).

Gierl, M.J., Wang, C., & Zhou, J. (2008b). *Using the attribute hierarchy method to develop cognitive models and evaluate problem-solving skills in algebra on the SAT®* (Research Report). New York: The College Board.

Haladyna, T., & Shindoll, R. (1989). Items shells: A method for writing effective multiple-choice test items. *Evaluation and the Health Professions, 12,* 97–106.

LaDuca, A., Staples, W.I., Templeton, B., & Holzman, G.B. (1986). Item modeling procedure for constructing content-equivalent multiple-choice questions. *Medical Education, 20*, 53–56.

Leighton, J.P., & Gierl, M.J. (Eds.) (2007a). *Cognitive diagnostic assessment for education: Theory and practices*. Cambridge, UK: Cambridge University Press.

Leighton, J.P., & Gierl, M.J. (2007b). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26,* 3–16.

Leighton, J.P., Gierl, M.J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement, 41*, 205–236.

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

Luecht, R.M. (2006b, September). *Assessment engineering: An emerging discipline*. Paper presented in the Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, Canada.

Luecht, R.M., Gierl, M.J., Tan, X., & Huff, K. (2006, April). *Scalability and the development of useful diagnostic scales.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

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

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

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, Macmillian.

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. (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.

Yang, X, & Embretson, S.E. (2007). Construct validity and cognitive diagnostic assessment. In J.P. Leighton, & M.J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and practices* (pp. 119–145). Cambridge, UK: Cambridge University Press.

van der Linden, W.J., & Glas, C.A.W. (2000). *Computer adaptive testing: Theory and practice.* Dordrecht, The Netherlands: Kluwer.

Zhou, J., Gierl, M.J., & Cui, Y. (2007, June). *Computerized attribute-adaptive testing: A new computerized adaptive testing approach incorporating cognitive psychology.* Paper presented at Graduate Management Admission Council (GMAC®) Conference on Computerized Adaptive Testing, Minneapolis, MN.

Mark J. Gierl

Professor of Educational Psychology and Canada Research Chair in Educational Measurement
Centre for Research in Applied Measurement and Evaluation
Department of Educational Psychology
6-110 Education North
University of Alberty
Edmonton, Alberta T6G 2G5
Canada
Tel. +1 780 492-2396
Fax +1 780 492-0001
E-mail mark.gierl@ualberta.ca

