

Computerized Attribute-Adaptive Testing: A New Computerized Adaptive Testing Approach Incorporating Cognitive Psychology

Jiawen Zhou, Mark J. Gierl, and Ying Cui
University of Alberta

Presented at the CAT and Cognitive Structure Paper Session, June 7, 2007



2007 GMAC® Conference on Computerized Adaptive Testing

Abstract

Modern computer technology has accelerated computerized adaptive testing implementation. However, few operational computer-based tests consider the underlying cognitive psychology of testing, either in the test development or in the diagnostic feedback provided to students. A concept of computerized attribute-adaptive testing (CA-AT), which integrates computerized adaptive testing with an attribute hierarchy method, is proposed.. The properties and strengths of CA-AT are elaborated, followed by a feasibility study which evaluates CA-AT with respect to attribute reliability.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Copyright © 2007 by the authors.

All rights reserved. Permission is granted for non-commercial use.

Citation

Zhou, J., Gierl, M. J., & Cui, Y. Computerized attribute-adaptive testing: A new computerized adaptive testing approach incorporating cognitive psychology. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Jiawen Zhou, Center for Research in Applied Measurement and Evaluation, Department of Educational Psychology, 6 - 110 Education North, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5, jzhou@ualberta.ca

Computerized Attribute-Adaptive Testing: A New Computerized Adaptive Testing Approach Incorporating Cognitive Psychology

Educational measurement is currently undergoing profound changes, with the developments in cognitive psychology, mathematical statistics, and computer science impacting the testing field. Cognitive psychology is fundamental to educational measurement because most educational tests are based on cognitive problem-solving tasks. Researchers and practitioners are now calling for cognitive diagnostic tests that integrate cognitive psychology and educational measurement to enhance the understanding of the psychology underlying student performance (Gierl, Leighton, & Hunka, 2007; Leighton & Gierl, 2007; Leighton, Gierl, & Hunka, 2004; Mislevy, 2006; Snow & Lohman, 1989).

The advance of computer technology has also accelerated computerized adaptive testing (CAT) implementation. Modern computer science facilitates the application of CAT because it provides technical support that is required for real-time item adaptive administration and ability estimation. Currently, CAT is widely used by many testing programs, such as the Armed Services Vocational Aptitude Battery, the Graduate Record Exam, and the Certified Public Accountants Licensure Exam. The advantages of CAT help explain its growing prevalence. Many benefits in CAT stem from computer-based test administration. These benefits include the potential to increase testing efficiency and decision accuracy, enhanced test security, use of multimedia in the test, and ease in pretesting items.

However, few operational CATs consider the underlying cognitive psychology of testing, either in test development or in the diagnostic feedback provided to students. As CATs began to be widely used in practice, certain practical shortcomings related to cognitive psychology became evident, and CAT incorporating cognitive psychology is required to address these limitations.

A cognitive diagnostic test designed to model examinees' cognitive performances yields specific information not only about the curricular domain, but also about the examinees' problem-solving strengths and weaknesses (Leighton & Gierl, 2007). The attribute hierarchy method (AHM; Leighton et al., 2004) is a psychometric method for classifying examinees' test item responses into a set of structured attribute patterns associated with different components from a cognitive model of task performance. A cognitive attribute is defined as the knowledge or skills required to perform a task involved in a specific domain (Leighton et al., 2004). The AHM helps link cognitive theory and psychometric practice to facilitate the development and analyses of educational and psychological tests.

The purpose of the current study was to integrate CAT and AHM; that is, to propose a concept of *computerized attribute-adaptive testing* (CA-AT). In a CA-AT, the test structure, each item and item adaptive assignment paths are developed based on the attribute hierarchy for a specific domain. CA-AT preserves the appealing efficiency and accuracy characteristics of CAT and also provides diagnostic value because results from this approach yield information on examinees' cognitive strengths and weaknesses. In this paper, the AHM and the fundamental principle of CA-AT are reviewed. The properties and strengths of CA-AT are then elaborated, followed by a feasibility study which evaluates CA-AT with respect to attribute reliability.

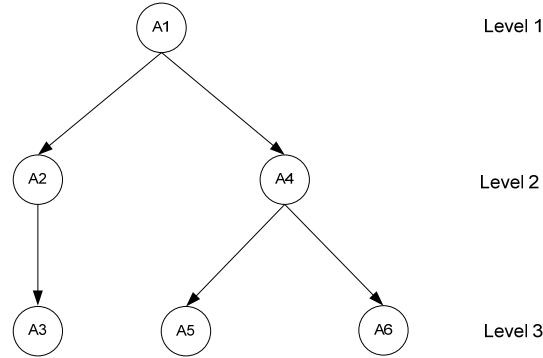
Overview of the Attribute Hierarchy Method

Cognition matters because a psychological perspective determines the nature of every element in an assessment argument and the rationale that orchestrates them. A cognitive model of task performance provides a deep understanding of task performance to yield score inferences specifically targeted to examinees' problem-solving skills. It is a simplified description of human problem solving on standardized tasks at some convenient grain size to facilitate interpretation and prediction of students' performance with respect to their strengths and weaknesses (Leighton & Gierl, 2007). A cognitive model of task performance provides an interpretative framework that can guide item development so test performance can be linked to specific cognitive inferences about examinees' knowledge, processes, and strategies (Snow & Lohman, 1989). With a cognitive model, problem solving is assumed to require the processing of information using relevant sequences of operations.

The AHM (Leighton et al., 2004) is a cognitively-based psychometric procedure for classifying examinees' test item responses into a set of hierarchically structured attribute patterns. An attribute is a description of the procedural or declarative knowledge needed to perform a task in a specific domain. Serving as a cognitive model, the AHM is based on the assumption that cognitive attributes are hierarchically related, which reflects an important characteristics of human cognition because cognitive skills do not operate in isolation (Anderson, 1996; Mislevy, Steinberg, & Almond, 2003). Attributes forming a hierarchy are associated with different components of a cognitive model of task performance. The structure of attributes in a hierarchy also reflects the psychological ordering among the attributes required in the specific domain of interest. The attribute hierarchy represents a specific construct as well as the related knowledge and processing skills underlying test performance for the construct. Consequently, the attribute hierarchy has a foundational role in the AHM, as it represents both the construct as well as the knowledge and processing skills that underlie test performance.

An attribute hierarchy is presented in Figure 1. The hierarchy contains two independent branches which share a common prerequisite — Attribute A1. The attributes span three levels. The first branch includes two additional attributes, A2 and A3, and the second branch includes a self-contained sub-hierarchy which includes Attributes A4 through A6. Two independent branches compose the sub-hierarchy: Attributes A4, A5 and Attributes A4, A6. A formal representation is used where the adjacency, reachability, incidence, reduced-incidence, and expected response matrices are specified (Tatsuoka, 1983, 1990, 1995).

Figure 1. A Hypothetical Hierarchy Containing Six Attributes



A binary adjacency matrix (A) of order (K, K) , where K is the number of attributes, specifies the direct relationships among attributes. Then, a reachability matrix (R) of order (K, K) , where K is the number of attributes, specifies the direct and indirect relationships among attributes. The R matrix is 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 , given A , the adjacency matrix, and I , an identity matrix. The incidence matrix (Q) of order (K, P) where k is the number of attributes and P is the number of potential items, is produced next. The set of potential items is considered a bank or pool of items that probes all combinations of attributes when the attributes are *dependent* and *independent*. The columns of the Q matrix are created by converting the items ranging from 1 to $2^K - 1$ to their binary form. This potential item pool is reduced when the attributes are related in a hierarchical structure because the hierarchy imposes *dependencies* among the attributes resulting in a reduced Q matrix (Q_r). The Q_r matrix is produced by determining which columns of the R matrix are logically included in columns of the Q matrix, using Boolean inclusion. The Q_r matrix represents the cognitive specifications for the test, and it is used as a guide to develop and/or interpret items that measure specific attributes outlined in the hierarchy. 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 imposed by the hierarchy. The (Q_r) matrix for the attribute hierarchy in Figure 1 is given by

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

Finally, the *expected response matrix* (E) is created, again using Boolean inclusion, where the algorithm compares each row of the expected attribute pattern matrix (which is the transpose of the Q_r matrix) to the columns of the Q_r matrix. The E matrix is of order (J, I) , where J is the number of examinees and I is the number of items. The rows of the E matrix are those responses that would be produced by an examinee who possesses the attributes as defined and structured in the attribute hierarchy and presented by the columns of the Q_r matrix. The columns of the E matrix are the items that probe specific attribute combinations. When an examinee's attributes

(Matrix 2)

Attribute reliability refers to the consistency of the decisions made in a diagnostic test about examinees' mastery of specific attributes. In AHM, 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. With the AHM, an item can be designed to measure a combination of attributes. Consequently, for items that measure more than one attribute, each attribute only contributes to a part of the total item-level variance.

Let W_{ik} denote the weight for item i in the calculation of attribute k . A W_{ik} value of 1 indicates that performance on item i is completely determined by attribute k . Hence, the variance of the responses on item i should be used in the calculation of the reliability of attribute k . Conversely, if W_{ik} has a value of 0, indicating that the mastery of attribute k could not increase the probability of solving item i correctly, then item i should not be used to calculate the reliability of attribute k . W_{ik} can be calculated as

$$W_{tk} = p(X_t = 1 | A_k = 1) - p(X_t = 1 | A_k = 0), \quad (1)$$

where $p(X_i = 1|A_k = 1)$ is the conditional probability that an examinee who has mastered attribute can answer item i correctly, and $p(X_i = 1|A_k = 0)$ is the conditional probability that an examinee who has not mastered attribute k can answer item i correctly.

The term $p(X_i = 1|A_k = 1)$ is calculated as

$$p(X_i = 1|A_k = 1) = \frac{p(A_k = 1, X_i = 1)}{p(A_k = 1)}, \quad (2)$$

where $p(A_k = 1, X_i = 1)$ is the joint probability that an examinee has attribute k and correctly answers item i , and $p(A_k = 1)$ is the marginal probability that an examinee has attribute k . To obtain $p(A_k = 1, X_i = 1)$ and $p(A_k = 1)$, the attribute patterns, the expected response patterns, and the population probabilities associated with each of the expected response patterns must be specified. Alternatively, the term $p(X_i = 1|A_k = 0)$ should be 0 because examinees are not expected to answer item i correctly since they lack attribute k required by item i . However, in an actual testing situation, it is possible that examinees can still answer the item correctly by guessing or by applying partial knowledge to reach their solution, particularly when the multiple-choice item format is used. Therefore, $p(X_i = 1|A_k = 0)$ can also be fixed at a specific value (e.g., 0.20) that reflects a “pseudo-guessing” parameter.

Once the W_{ik} s are specified, the weighted scores can be used to calculate attribute reliability by adapting Cronbach’s alpha for the AHM framework. The derived formula is given by

$$\alpha_{AHM_k} = \frac{n_k}{n_k - 1} \left[1 - \frac{\sum_{i \in S_k} W_{ik}^2 \sigma_{X_i}^2}{\sigma_{\sum_{i \in S_k} W_{ik} X_i}^2} \right], \quad (3)$$

where α_{AHM_k} is the reliability for attribute k , n_k is the number of items that are probing attribute k in the Q_r (i.e., the number of elements in S_k), $\sigma_{X_i}^2$ is the variance of the observed scores on item i , $\sum_{i \in S_k} W_{ik}^2 \sigma_{X_i}^2$ is the weighted observed total score on the items that are measuring attribute k , and $\sigma_{\sum_{i \in S_k} W_{ik} X_i}^2$ is the variance of the weighted observed total scores.

With attribute reliability specified by adapting Cronbach’s alpha for the AHM framework, the Spearman-Brown formula can also be used to estimate the effect of changes to test length. The Spearman-Brown formula is specified as

$$\alpha_{AHM-SB_k} = \frac{n_k \alpha_{AHM_k}}{1 + (n_k - 1) \alpha_{AHM_k}}, \quad (4)$$

where α_{AHM-SB_k} is the Spearman-Brown reliability of attribute k if n_k additional items sets that are parallel to items measuring attribute k are added to the test. This formula can be used to evaluate the effect of adding parallel items to the reduced-incidence matrix.

Overview of Computerized Attribute-Adaptive Testing

CAT is a test administration method in which examinees are administered items based on their performance on previous items. Items to be administered in a structures CAT can be arranged on the basis of their difficulty and discrimination levels. CAT offers the advantages of increased testing efficiency and decision accuracy compared with conventional paper-and-pencil

testing. However, the cognitive aspect of examinees' test performance is not considered in a CAT, making it difficult to link test performance to specific cognitive inferences about examinees' knowledge, processes, and strategies.

CA-AT, guided by an attribute hierarchy in a specific domain, differs from CAT in four ways: cognitive model integration, item bank, item routing, and results reporting. First, items involved in a CA-AT are designed on the basis of an attribute hierarchy while only the content of a specific domain is considered in CAT. Second, the item bank of CA-AT is organized according to the structure of an attribute hierarchy for a specific domain whereas CAT classifies items in terms of calibrated item parameters with equal representation all along the ability scale. Third, the item administration routing of a CA-AT is linked to branches or cognitive nodes from a specific attribute hierarchy. In CAT, items are administered according to item difficulty levels along a linear score scale. Fourth, strong inferences about examinees' attribute-based strengths and weakness can be made in CA-AT while only a global inference based on a total test score is available in CAT. CA-AT possesses these four distinct advantages compared with CAT while it retains the efficiency and accuracy of CAT.

Incorporated Cognitive Model

A psychological perspective determines the nature of each element in an assessment argument. To make specific inferences about problem solving, cognitive models of task performance that provide a deep understanding of problem-solving are required. A *cognitive model* in educational measurement refers to a simplified description of human problem solving on standardized tasks at some convenient grain size or level of detail in order to facilitate explanation and prediction of students' performance, including their strengths and weaknesses (Leighton & Gierl, 2007). A cognitive model of task performance is specified at a small grain size because it magnifies the cognitive processes underlying test performance.

The AHM is a psychometric method for classifying examinees' test item responses into a set of structured attribute patterns associated with different components from a *cognitive model of task performance*. These models provide an interpretive framework that can guide item development so test performance can be linked to specific cognitive inferences about examinees' knowledge, processes, and strategies. Consequently, the attribute hierarchy has a foundational role in the AHM, as it represents both the construct as well as the knowledge and processing skills that underlie test performance. The AHM helps link cognitive theory and psychometric practice to facilitate the development and analyses of educational and psychological tests. Results from AHM analyses yield information on examinees' cognitive strengths and weaknesses.

Integrated with the attribute hierarchy method, a CA-AT contains items directly measuring specific content with cognitive processes of increasing complexity. Test performance is, therefore, closely linked to specific inferences about examinees' cognitive strengths and weakness as items are developed to measure different components and processes in the model. The total score and the detailed attribute-based diagnostic feedback illuminate the examinees' cognitive problem-solving strengths and weaknesses. The attribute hierarchy method, therefore, helps link cognitive theory to psychometric practice to facilitate CA-AT test construction as well as diagnostic feedback and test score reporting. The incorporation of the attribute hierarchy in CA-AT overcomes the cognitive limitation in the traditional computer-based testing by administering items based on their underlying cognitive model of task performance.

Item Bank Construction

Item bank construction includes item development and item bank organization. The item bank associated with a CA-AT is different from a conventional CAT because of the central role of the attribute hierarchy. For each attribute hierarchy, there is an associated item bank.

One strength of the AHM lies in its facility to guide item and test development. In AHM, a Q_r matrix of order (K, I) indicates I items are required to measure the K attributes given the restrictions imposed by the attribute hierarchy (Cui, Leighton, Gierl, & Hunka, 2006). The structure of an attribute hierarchy is represented by a Q_r matrix. For example, in Matrix 1, the first item measures attribute A1 only ([100000]) while the second item measures attributes A1 and A2 ([110000]), indicating that attribute A1 is the prerequisite of attribute A2. The number of items to measure each attribute included in a test can be determined based on the attribute reliability indices (Gierl, Cui, & Hunka, 2007).

Items written for the CA-AT item bank are linked directly to each attribute involved in the hierarchy. The items created for one attribute are assumed to measure the same attribute, therefore having the same difficulty level. Creating items using an attribute hierarchy, a test developer can orchestrate which attributes are measured by which items, so that test performance can be linked to information about students' cognitive strengths and weaknesses. Assessments based on an attribute hierarchy should be developed so that test items directly measure specific cognitive processes of increasing complexity in the understanding of a domain.

A test developed with a particular attributes' hierarchical order in mind measures cognitive processes of increasing complexity in the understanding of a specific construct. That is, items developed for a CA-AT warrant the representation of cognitive processes underlying mastery of each attribute. A large number of items are designed to measure each attribute in a CA-AT item bank to facilitate the item selection in an adaptive test as well as to avoid item over-exposure.

Consequently, the rule of organizing items for a CA-AT item bank is also affected by the incorporation of a cognitive model of task performance. A CA-AT item bank represents the structure of the corresponding attribute hierarchy. Each item is designed to target one specific attribute. Items are therefore associated with the hierarchical organization of the attributes in the hierarchy.

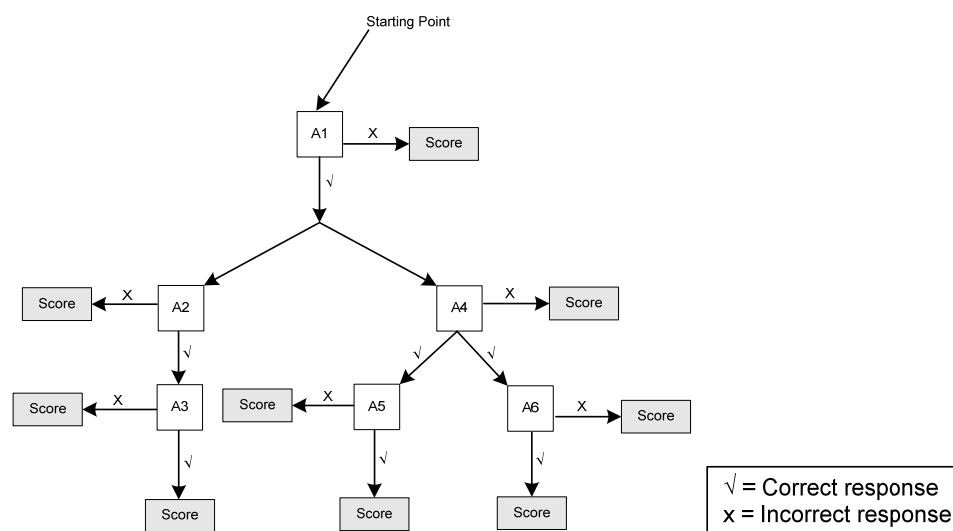
Item Routing

The rule of item routing for a CA-AT is also influenced by the attribute hierarchy. The fundamental routing rule is to distribute items based on the structure of the hierarchy. Compared to conventional CAT, in which items for administration to a test taker are chosen in real-time based on the test taker's responses to previous items, CA-AT provides more specific information about examinees' underlying cognitive skills.

Item adaptive assignment paths are developed based on the attribute hierarchy model for a specific domain. Often, items measuring the first attribute in one attribute hierarchy are administered to examinees as the starting point of a CA-AT, as presented in Figure 2. Examinees are given items measuring the next attribute only when they answer the items on the previous attribute correctly. For example, as indicated in Figure 2, when the examinee correctly answers items measuring attribute A1, items measuring attribute A2 and attribute A4 can then be administered. If an examinee fails to answer items measuring attribute A2 correctly, the item administration on the left branch stops. The examinee has no chance to solve items measuring

attribute A3 because the prerequisite skill for attribute A3 (i.e., attribute A2) has not been mastered. Similarly, when the examinee correctly answers items measuring attribute A4, he/she has the chance to write items measuring attributes A5 and A6. That is, if an examinee is administered a CA-AT based on an attribute hierarchy, he/she has the chance to answer the items measuring one attribute only if he/she correctly answers items associated with the prerequisite attribute. Such item administration routing of a CA-AT links to the layout of a specific attribute hierarchy. The attribute hierarchy model thereby has a control over item routing in CA-AT.

Figure 2. CA-AT Item Routing for the Algebra Hierarchy Presented in Figure 1



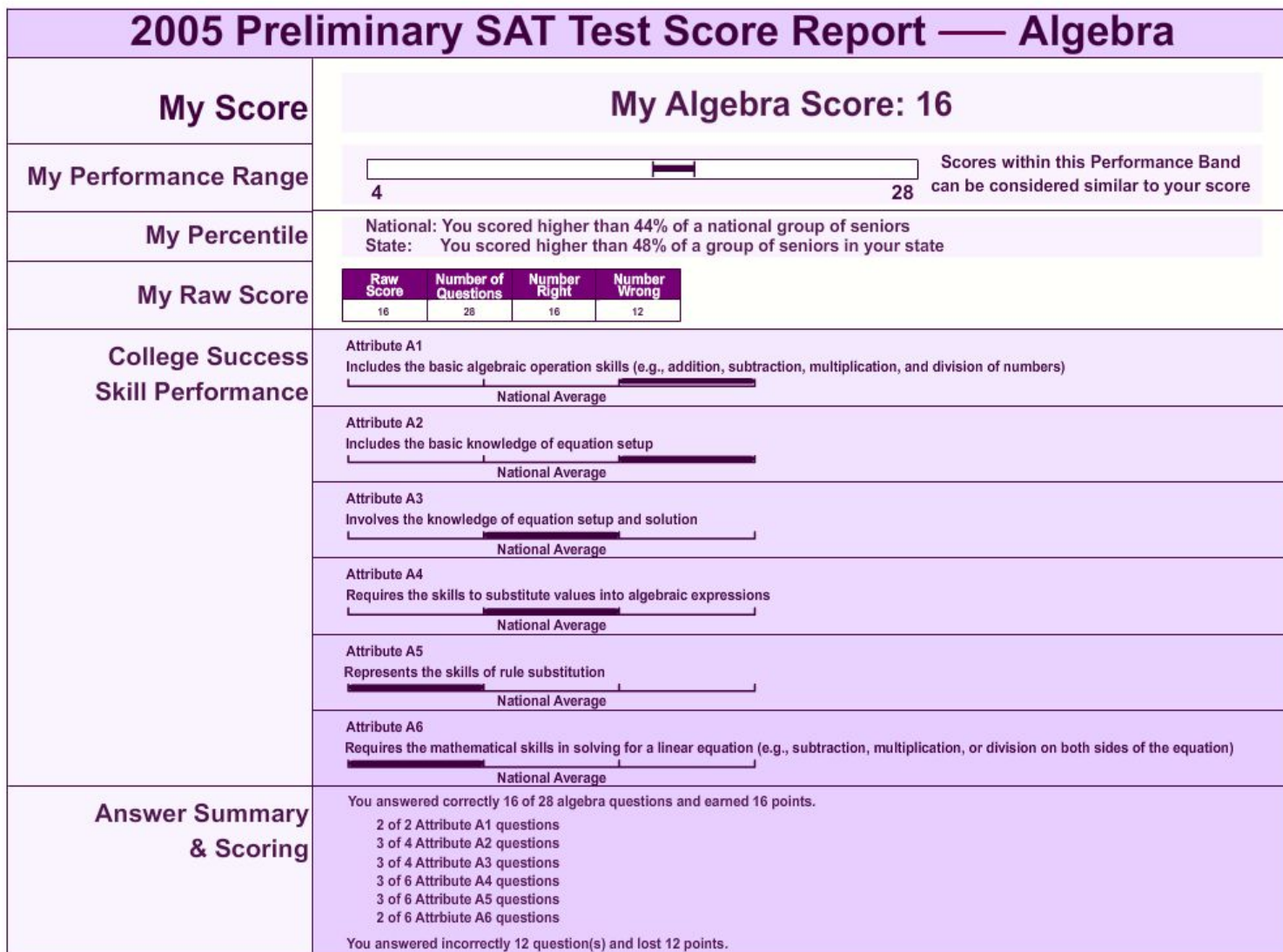
Score Reporting

CA-AT is an assessment format based on cognitive models of task performance. Test items administered in a CA-AT directly measure specific cognitive processes of increasing complexity. As a result, not only the total test score, but also diagnostic feedback can be provided with CA-AT. For the purpose of demonstration, a sample PSAT algebra test report is presented in Figure 3. The PSAT is a standardized test designed to provide firsthand practice for the SAT Reasoning Test™. In the report, examinees' score, examinees' performance range, and examinees' percentiles are reported. In addition to this traditional test report content, the report also provides information about examinees' mastery and non-mastery of each attribute assessed in the test, which is a characteristic of a diagnostic test. For example, an examinee who gets 16 of 28 items correct on the PSAT algebra test was informed that his/her performance is in the middle level compared to the whole nation. Also, the examinee could find his/her strengths and weaknesses regarding the attributes assessed in the PSAT algebra test. That is, he/she has mastered attribute A1 to attribute A4 while he/she has not mastered the knowledge and skills associated with attribute A5 and attribute A6, which measure the skills of rule substitution in equation solving and mathematical operations (e.g., subtraction, multiplication, or division on both sides of the equation) in solving a linear equation, respectively.

Diagnostic feedback on a preliminary test like the PSAT helps students compensate and overcome their deficiencies. The hierarchy facilitates the diagnostic property of CA-AT by

offering the basic cognitive theory underlying the domain. Specific diagnostic inferences can be generated when items are developed to measure different components and processes in the model. Using the hierarchy to create items to measure the cognitive components described by the model, test developers can orchestrate which attributes are measured by which items. Using the attribute hierarchy to decide item routing, examinees are assessed according to their cognitive proficiency. Using the attribute hierarchy to interpret test performance, test users gain control over the scores and the inferences about processes and skills associated with test performance. Strong inferences about examinees' cognitive skills can be made because the small grain size in these models help illuminate the knowledge and skills required to perform competently on testing tasks.

Figure 3. A Sample PSAT Algebra Test Score Report



Summary

Developed and administered based on the cognitive principle of an attribute hierarchy in a specific domain, CA-AT shows distinct advantages compared to traditional adaptive tests. With the integration of a cognitive model, the items of a CA-AT link directly to each attribute involved in a hierarchy, providing manifest cognitive inferences about examinees' knowledge strengths and weaknesses. The organization of the test bank of a CA-AT represents the attribute components of a hierarchy. CA-AT test items are administered according to the structure of a specific hierarchy, based on the assumption of AHM that cognitive attributes are hierarchically related. An examinee has to show that he/she possesses knowledge and skills representing one attribute to have the opportunity of answering the items that measure the next, more complex attribute. Strong inferences about examinees' cognitive strengths and weakness can therefore be made in CA-AT.

A feasibility study was conducted to evaluate the attribute reliability of a CA-AT. Attribute reliability refers to the consistency of the decisions made in a diagnostic test about examinees' mastery of a specific attribute (Gierl, Cui, & Hunka, 2007). In a CA-AT, the higher the reliability of each attribute, the more consistent are the decisions about examinees' mastery of specific attributes in the corresponding hierarchy. Attribute reliability was calculated with different numbers of items for each attribute in a hierarchy.

Feasibility Study

CA-AT is a testing mode that not only possesses efficiency and accuracy characteristics of computerized adaptive testing but also provides diagnostic value by incorporating the properties of the attribute hierarchy into test development and analysis. Prior to CA-AT administration, it is critical to determine the number of items required to measure each attribute to achieve satisfactory attribute reliability. A simulation study was conducted to explore the required number of items for each attribute in a specific hierarchy. The attribute hierarchy presented in Figure 1 was used in the simulation.

Method

Two variables were manipulated in this study: the number of items measuring each attribute and the *slip level*. A slip is the discrepancy of the observed response pattern compared to the expected response pattern based on the hierarchy. The attribute reliability of each attribute for two sets of items was calculated. In total, fourteen items were used in the first set of items while twenty-eight items were included in the second set of items. In the 1-2-3 set, *one* item measured attribute A1, *two* items measured attribute A2 and A4, respectively, and *three* items measured attribute A3, A5, and A6. In the 2-4-6 set, the number of items that measured each attribute was doubled; that is, there were *two* items measuring attribute A1, *four* items measuring each of the attributes A2 and A4, and *six* items measuring each of the attributes A3, A5, and A6.

The six items that measured attributes A1 to A6, as presented in the Figure 1 model, were selected from the PSAT 2006 Mathematics administration. On the basis of the calibrated item parameters, item responses of 5,000 examinees to the fourteen-item testlet and twenty eight-item testlet were simulated using the three-parameter logistic (3-PL) item response theory (IRT) model. The 3-PL IRT model can be expressed by the following formula

$$P(\theta) = c_i + (1 - c_i) \frac{1}{1 + e^{-Da_i(\theta - b_i)}}, \quad (5)$$

where c_i is the guessing parameter, b_i is the item difficulty parameter, a_i is the discrimination parameter, and $D = 1.7$.

To simulate a real testing situation, response slips were generated and placed within the response data. The percentage of slips was manipulated in the simulation study. Two types of slips were generated. First, slips were created for the items that examinees were expected to incorrectly answer relative to the attribute hierarchy. For example, an examinee was assumed to incorrectly answer items measuring attribute A3 when he/she cannot correctly answer items measuring attribute A2 based on the attribute hierarchy presented in Figure 1. To generate a slip in this situation, a response pattern with an incorrect answer for items measuring attribute A2 but with a correct answer for items measuring attribute A3 was produced. Second, slips were created for the items that examinees were expected to correctly answer relative to the attribute hierarchy. For example, an examinee was assumed to correctly answer items measuring attribute A1 when he/she correctly answered items measuring attribute A2, because attribute A1 is the prerequisite of attribute A2 in the PSAT algebra hierarchy. To generate a slip in this situation, a response pattern with an incorrect answer for items measuring attribute A1 but with a correct answer for items measuring attribute A2 was produced. Three percentages of slips, 0%, 5%, and 10%, were added to the total number of response patterns.

Results

Six attribute reliability indices were calculated, as illustrated in Table 1. Across all conditions, attribute A1 had the highest attribute reliability, which is reasonable because attribute A1 is the prerequisite of attributes A2 to A6. In other words, an examinee who can correctly answer items measuring attributes A2 to A6 was assumed to be able to correctly answer items measuring the attribute A1.

Table 1. Attribute Reliability of Six Attributes in the PSAT Algebra Hierarchy

Number of items measuring the attribute	Level	Attribute	Slips		
			0%	5%	10%
1-2-3 set	1	A1	0.82	0.81	0.79
		A2	0.69	0.69	0.67
	2	A4	0.69	0.68	0.67
		A3	0.66	0.67	0.64
	3	A5	0.70	0.71	0.69
		A6	0.71	0.71	0.68
2-4-6 set	1	A1	0.90	0.90	0.89
		A2	0.82	0.81	0.81
	2	A4	0.83	0.83	0.83
		A3	0.79	0.78	0.78
	3	A5	0.80	0.80	0.80
		A6	0.70	0.69	0.69

The reliability for the attributes in the 1-2-3 set was lower than 0.70, which may be viewed as unacceptable. The reliability for the attributes in the 2-4-6 set was acceptable, as they were approximately 0.70 or higher. In the 2-4-6 set, the attribute reliability for prerequisite attribute was consistently greater than that for the next attribute. For example, when the percentage of slips was zero, the attribute reliability for attribute A1 was 0.90, which was greater than that (0.82) for attribute A2,. However, in the 1-2-3 set, the attribute reliability for attribute A4 was consistently smaller than that for attribute A5 and A6, across all slip conditions. For example, when the percentage of slips was 5%, the attribute reliability for attribute A4, 0.68, was smaller than that for attributes A5 and A6, both of which were 0.71.

The attribute reliabilities of the 2-4-6 set for all attributes A1 to A6 were greater than those in the 1-2-3 set, with only a few exceptions. For example, with zero slip attribute reliabilities for the six attributes were 0.90, 0.82, 0.79, 0.83, 0.80, and 0.70 in the 2-4-6 set. The attribute reliabilities were 0.82, 0.69, 0.66, 0.69, 0.70, and 0.71 in the 1-2-3 set. Only the attribute reliability of attribute A6 in 2-4-6 set was slightly smaller than that in 1-2-3 set.

Attribute reliability for the six attributes changed little with varying slip percentages in both the 1-2-3 and 2-4-6 sets. For example, in the 2-4-6 set, attribute reliabilities for the six attributes were 0.90, 0.82, 0.79, 0.83, 0.80, and 0.70 when there was no slip generated in the data simulation. The attribute reliabilities were 0.90, 0.81, 0.78, 0.83, 0.80, and 0.69 when there were 5% slips. In other words, the two sets of attribute reliabilities were quite similar to one other.

Conclusions

A feasibility study that evaluates the attribute reliability of a CA-AT was conducted. The variables *number of items measuring each attribute* and *the percentage of slips* were manipulated. For a specific hierarchy, as presented in Figure 1, the number of items measuring each attribute affected attribute reliability while the slip level had little affect.

When there were *two* items measuring attribute A1, *four* items measuring each of the attribute A2 and A4, and *six* items measuring each of attributes A3, A5, and A6, the reliability for all six attributes was acceptable. However, when *one* item measured attribute A1, *two* items measured attribute A2 and A4, respectively, and *three* items measured attribute A3, A5, and A6, respectively, most of the attribute reliabilities were lower than 0.70. Therefore, a hierarchy with three levels requires at least two items measuring the attribute in level 1, four items measuring the attribute in level 2, and six items measuring the attribute of level 3 to obtain acceptable attribute reliabilities across the whole model.

The varying percentage levels of slips affected attribute reliability little in the current study. Across three slip percentage conditions, attribute reliability changed little in both 1-2-3 set and 2-4-6 set. This result indicates that CA-AT has the potential of maintaining acceptable attribute reliabilities even with relatively inconsistent observed response patterns relative to the expected response patterns.

Summary and Future Directions

A new test concept, computerized attribute-adaptive testing, was proposed in the current study. CA-AT, incorporating the attribute hierarchy method, links cognitive theory and psychometric practice to facilitate the development and analyses of educational and psychological tests. CA-AT preserves the appealing characteristics of efficiency and accuracy in computerized adaptive testing but also provides diagnostic value because results from this approach yield information on examinees' cognitive strengths and weaknesses. A feasibility study was also conducted to evaluate the required number of items measuring each attribute.

In the current study, only simulated data were analyzed, which might limit the generalizability of the results. The discrepancy among items measuring the same attribute was also ignored in the simulation study, which might lead to limited results and conclusions because the variability among items measuring the same attribute might deviate examinees' observed response patterns. Hence, further research using real data is still needed.

To realize the CA-AT test format, much more research is required. A well developed attribute hierarchy that represents the cognitive relationship among attributes is fundamental to CA-AT. The development of a hierarchy model needs efforts from content experts and cognitive psychologists. The routing strategy in CA-AT is another topic that deserves further exploration. Optional item routing strategies include multi-stage item routing, the percentage routing strategy that administers items measuring the next attribute based on the percentage of correctly answered items measuring the pervious attribute, and the number-correct item routing strategy. These procedures should be evaluated and compared in future studies.

References

- Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, 51, 355-365.
- Cui, Y., Leighton, J. P., Gierl, M. J., & Hunka, S. (2006, April). *A person-fit statistic for the attribute hierarchy method: The hierarchy consistency index*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Gierl, M. J., Cui, Y., & Hunka, S. (2007). The attribute hierarchy method for cognitive assessment: Technical developments. *Applied Measurement in Education*. Manuscript submitted for publication.
- 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 practices*. Cambridge University Press.
- Leighton, J. P., & Gierl, M. J. (Eds.) (2007). *Cognitive diagnostic assessment for education: Theory and practices*. Cambridge University Press.
- Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement*, 41, 205-236.
- Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257-306). Washington, DC: American Council on Education.
- Mislevy, R. J., Steinberg, L. S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-62.
- 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). American Council on Education, Macmillian, New York.
- 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. (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. (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.