

Running head: TESTING EXPERT-BASED AND STUDENT-BASED COGNITIVE MODELS

**Testing Expert-Based and Student-Based Cognitive Models: An Application of the Attribute Hierarchy
Method and Hierarchy Consistency Index**

Jacqueline P. Leighton

Ying Cui

M. Ken Cor

Centre for Research in Applied Measurement and Evaluation (CRAME),

University of Alberta

Acknowledgement: Preparation of this article was supported by a grant from the Social Sciences and Humanities Research Council of Canada (Grant No. 410-2003-0210). Grantees undertaking such projects are encouraged to express freely their professional judgment. This article, therefore, does not necessarily represent the positions or the policies of the Canadian government, and no official endorsement should be inferred. We wish to thank Dr. Mark Gierl and Dr. Steve Hunka, as well as Dr. Rebecca Gokiert, and Colleen Heffernan for their help and support with different parts of the study. We also thank the College Board for making this research possible by allowing us to work with their data and test items.

Abstract

The objective of the present investigation was to compare the adequacy of two cognitive models for predicting examinee performance on a sample of algebra I and II items from the March 2005 administration of the SAT™. The two models included one generated from verbal reports provided by 21 examinees as they solved the SAT™ items, and the other generated from the judgment of a content expert. Using the attribute hierarchy method (Leighton, Gierl, & Hunka, 2004) and the hierarchy consistency index (Cui, Leighton, Gierl, & Hunka, 2006), the predictive adequacy of the two cognitive models was tested with three samples of varying size. Results indicated that the student-based model predicted the test performance of examinees of moderately-high ability with excellent accuracy but not the test performance of examinees of moderate or average ability. In contrast, the expert-based model was more consistent, predicting the test performance of both moderately-high ability examinees and moderate ability examinees with reasonable accuracy. Expert-based models may be more predictive of the general response processes used by diverse groups of examinees, whereas student-based models may be preferred when making inferences about examinees at specific ability levels.

Testing Expert-Based and Student-Based Cognitive Models: An Application of the Attribute Hierarchy Method and Hierarchy Consistency Index

Developing cognitive models to predict examinee test performance and validate test-based inferences can be achieved by collecting evidence of examinees' response processes, for example, eye tracking information, reaction times, and verbal report data (AERA, NCME, APA, 1999; Kane, 2006). However, the development of cognitive models based on evidence of examinee response processes can be time-consuming and costly. For example, collecting verbal reports involves recruiting and interviewing students, audio-taping or video-taping their responses, and then transcribing, categorizing, coding, and finally interpreting the contents of the reports so as to design plausible cognitive models of task or test performance (Leighton, 2004). The time associated with the development of cognitive models based on verbal reports is arguably justified if it can be shown that such models are useful for predicting examinee test performance and validating test-based inferences. However, whether verbal reports are useful in the development of cognitive models of task performance is ultimately an empirical question. If it can be shown that verbal reports do indeed lead to cognitive models with greater predictive accuracy of examinee responses in comparison to, for example, cognitive models generated by content experts, then there would be an incentive to pursue this line of model development irrespective of time and cost. The objective of the present paper is to examine this supposition by evaluating two cognitive models of algebra performance—one model developed from content expert analysis and another developed from verbal report analysis.

The paper is divided into four sections. In the first section, we briefly describe the Attribute Hierarchy Method (AHM; Leighton et al., 2004), define a cognitive model of task performance, and describe the use of these models in contemporary diagnostic measurement approaches (i.e., Cognitive Design System, Evidence-Centered Design, and the AHM). In the second section, we describe the hierarchy consistency index (HCI; Cui et al., 2006), and the methods used to develop two cognitive models of algebra performance based on different sources of evidence (content expert analysis versus student verbal report analysis). In the third section, we describe the results of using each cognitive model to predict examinee performance. In the fourth section, we discuss how cognitive models of task performance developed from distinct sources of evidence

compare to each other in terms of predictive accuracy, and how the AHM and HCI can be used to evaluate different cognitive models.

The AHM and Developing Cognitive Models of Task Performance

The Attribute Hierarchy Method (AHM)

The AHM (for a more complete description of this method, see Leighton et al., 2004) is a recently proposed cognitive diagnostic tool designed explicitly to integrate cognitive models (defined in the next section) with a psychometric technique to predict examinee test performance and estimate their mastery of domain knowledge and processing skills. The AHM is based on Tatsuoka's (1983) Rule-Space Methodology (RSM). The AHM and RSM share many similarities. However, one difference is that the cognitive models incorporated in the AHM must have a hierarchical ordering of knowledge and skills. Although this is not a requirement for the RSM (see Tatsuoka, 1983), the AHM is based on the assumption that knowledge and skills function as a network of hierarchical, interrelated processes and do not operate independently. This assumption is based on psychological findings that many cognitive processes are ordered hierarchically in abstract, logical, and mathematical domains (e.g., see Lohman, 2000; Siegler, 1996; Siegler & Chen, 1998). Working from this assumption, then, examinee test performance is described by hierarchically related knowledge and skills called *attributes*.

Attributes are basic knowledge, skills, or cognitive processes, which are ordered sequentially based upon logical and/or psychological properties, that are expected to be required to solve test items (see Gierl, Leighton, & Hunka, 2000). For example, the process of solving categorical syllogisms has been described as requiring the following attributes: understanding the meaning of quantifiers, creating single or multiple representations of the quantified premises, and drawing conclusions from the representations (see Johnson-Laird & Bara, 1984; also Leighton et al., 2004). As shown in Figure 1, the attributes can be ordered hierarchically from basic to complex because it is expected that in order to possess complex attributes such as creating multiple representations of quantified premises (i.e., attributes 4 and 6) one must have mastered basic attributes such as having an understanding of what quantified premises entail and being able to create at least one representation of the quantified premises (i.e., attributes 1 and 2).

The *attribute hierarchy* serves not only to explain and predict examinee test performance but also, by means of the reduced Q-matrix (described in the Method section), to guide the development of test items. In

this way, the attribute hierarchy ideally serves two functions—as an explicit illustration of the construct and also as a test blueprint. In particular, the AHM consists of three stages. In the first stage, an attribute hierarchy is defined to illustrate the construct of interest, including the relevant knowledge and processing skills in a specific target domain. These knowledge and skills are assumed to be used by examinees as they solve problems within the domain. In the second stage, the attribute hierarchy is used as a framework to develop test items to measure the relevant knowledge and skills of interest. An exception might be made at this stage if the AHM is used to summarize student performance on an existing test whose items were developed in the absence of an attribute hierarchy.¹ In this case, the AHM can still be used to classify student test performance but an attribute hierarchy must be specified, post hoc, to illustrate the knowledge and skills measured by the items (see Leighton et al., 2004 for a discussion of this issue). In the third stage, statistical classification procedures are used to categorize each examinee, based on their performance, into one of the knowledge and skill states derived from the attribute hierarchy. It is this third and final stage that permits inferences about examinees' cognitive strengths and weaknesses to be made—such as whether they need remediation with specific attributes. The validity of test-based inferences is premised on an accurate attribute hierarchy—that is, one that realistically describes the knowledge and skills examinees will use to respond to test items so that it not only predicts student performance but also explains the response processes underlying that performance. Given the importance of the attribute hierarchy, we address the issue of its development in the next section.

The Attribute Hierarchy as a Cognitive Model of Task Performance

The attribute hierarchy in the AHM can be viewed, generally, as a cognitive model of task performance, which summarizes the “set of interconnected knowledge and skills that students use to respond correctly to test items within a content domain” (Leighton, 2004, pp. 7-8). It is useful to view the attribute hierarchy as an instance of a cognitive model of task performance because such cognitive models are incorporated in other measurement approaches and can be developed in specific ways. Cognitive models of task performance are incorporated in at least three approaches to designing, summarizing, and interpreting test performance: Embretson's (1998) Cognitive Design System (CDS), Mislevy's (1994) Evidence Centered Design (ECD), and Leighton, Gierl, and Hunka's (2004) Attribute Hierarchy Method (AHM). Embretson and Gorin (2001, p. 351) state that “developing a cognitive model for the designated item type is essential to the

cognitive design system approach...[T]he relevant cognitive processes, strategies, and knowledge structures must be identified and organized into a unified model” (p. 351). Embretson’s (1998) CDS requires the coordination of seven steps: (a) specify goals of measurement, (b) identify design features in the task domain (that will elicit content knowledge and processing skills), (c) develop a cognitive model, (d) generate items, (e) evaluate the cognitive models for generated tests, (f) bank items by cognitive complexity, and (g) validate the test by considering its nomothetic span. Mislevy’s ECD (1994; see also Mislevy, Steinberg, & Almond, 2003) also requires the creation of models. In particular, three models—student, evidence, and task—are created. The *student* model illustrates the knowledge and skills examinees are expected to have mastered in a curriculum of study. The *evidence* model illustrates the types of data that will provide proof that examinees have either mastered or not mastered the knowledge and skills outlined in the student model. The *task* model exemplifies the features of items that should elicit specific data, in the form of behaviors (from examinees), in order to make defensible inferences about their mastery of knowledge and skills. All three models in coordination are relevant to test design and interpretation.

Cognitive models of task performance can be developed in many ways. However, one useful way to think about the demarcation of model development might involve the distinction between a top-down or bottom-up approach (Chi, 1997; Ericsson, 2006; Ericsson & Simon, 1993; Leighton, 2004; Leighton & Gierl, 2007). For example, using a top-down or confirmatory approach, psychologists may conduct a review of the literature to determine whether relevant psychological theories exist to guide the development of a cognitive model in a target domain of interest (see Embretson, 1998, 1999; Embretson & Gorin, 2001; however, see footnote 1). In similar manner, content experts may review learning outcomes and curriculum objectives, anticipate the relevant knowledge and skills with which to describe a construct, and in their design of test items. Content experts are often in a good position to anticipate the knowledge and skills students use to respond correctly to test items since many have worked as teachers and therefore have insights into student thinking and performance. This general top-down approach to developing a cognitive model of task performance is termed *content* or *expert analysis* in the present paper because it often begins with an a priori theory or set of beliefs of the knowledge and skills expected to underlie performance in the target domain. Cognitive models developed using expert analysis gain empirical credibility to the extent that they are corroborated with human studies in the target domain of interest. Corroborating models developed from

expert analysis with empirical data is often done in psychological studies of human information processing (see Ericsson, 2006) but less so in educational testing.

Cognitive models can also be developed inductively using a bottom-up approach by directly collecting data about examinee response processes and then generating a model of task performance. Think-aloud methods are often used for this purpose (see Chi, 1997, 2006). Using think aloud methods, examinees may be interviewed individually as they attempt to solve items. During the interview, the investigator requests examinees to “think aloud” as they solve an item in order to obtain a record or verbal report of the knowledge and skills used to solve the item. Verbal reports are usually collected from examinees (anywhere from 10 to 25 examinees per item) that are sufficiently at ease communicating orally as to provide a report (see Chi, 1997; Ericsson & Simon, 1993; Taylor & Dionne, 2000). After the reports are collected, the investigator and his or her team attempt to identify common patterns or themes in the knowledge and skills examinees used to solve the items. Common patterns are organized schematically, may be subjected to rater consistency analysis, and used to illustrate the response processes exhibited. This inductive, exploratory approach is called *verbal analysis* (the reader is referred to Chi, 1997, 2006 for a full description of this procedure).

Given the distinct approaches for developing cognitive models of task performance, how does one decide which approach will produce the best model for predicting examinee responses? In an ideal situation, experts would develop models and then collect empirical evidence to corroborate or modify the models (e.g., using verbal analysis). In many cases, however, time and cost will constrain this option. It is often less time-consuming and less expensive to adopt expert analysis (i.e., a top-down approach) alone. For example, in confirmatory factor-analytic studies (e.g., Hamilton, Nussbaum, & Snow, 1997; Leighton, Gokiart, & Cui, 2007), the investigator usually specifies a single cognitive model and builds a case for its plausibility even though other models may exist to fit the data (see Borsboom, 2005; Borsboom, Mellenbergh, & Van Heerden, 2004; Leighton et al., 2007; Luecht, 2007; Rindermann & Neubauer, 2004). Although a cognitive model developed from expert analysis may fit the data well, critics have suggested that better fitting cognitive models could be developed by conducting human studies of examinee response processes directly. In the absence of finding a ready-to-use psychological model from the existing research literature, Leighton (2004; see also footnote 1) has suggested developing cognitive models of task performance directly from student

verbal report data as a way to ensure that models are representative of examinee response processes (Nichols, 1994). The objective of the present paper is to examine this supposition by evaluating simultaneously two cognitive models of algebra performance—one model developed from expert analysis (i.e., a top-down approach) and another developed from verbal analysis (i.e., bottom-up approach).

Method

This section is divided into two parts. First, a brief overview of the hierarchy consistency index (HCI) is presented for testing the accuracy of cognitive models. Second, methods used to generate, test, and compare two cognitive models of algebra performance using Algebra I and II items from the March administration of the SAT™ are presented.

The Hierarchy Consistency Index (HCI)

The hierarchy consistency index (HCI; Cui et al., 2006) is a person-fit statistic designed to evaluate statistically the degree to which an observed examinee response vector is consistent with the attribute hierarchy. The HCI depends on item complexity as illustrated by the attribute hierarchy and the associated reduced Q-matrix, a matrix derived from the attribute hierarchy. In the AHM, the reduced Q-matrix is used to illustrate the attributes (i.e., knowledge and skills) required for examinees to solve each item correctly. The reduced Q-matrix is of order $K \times J$, where K is the number of attributes and J is the number of items. The reduced Q-matrix is therefore an attribute-by-item matrix, in which each column specifies the attributes required by examinees to respond correctly to the item (see Leighton et al., 2004, for a full description of the reduced Q-matrix). As mentioned previously in the introduction to the AHM, the reduced Q-matrix also functions as a test blueprint because it indicates that J items must be designed to measure the hierarchical relationships of the K attributes. For example, consider again the attribute hierarchy presented in Figure 1. The reduced Q-matrix associated with this hierarchy is shown in Table 1. This reduced Q-matrix is of order (7, 15) and illustrates that an item (column 1) must be designed to measure only attribute 1, a second item (column 2) must be designed to measure attributes 1 and 2, and a third item (column 3) must be designed to measure attributes 1, 2, and 3, and the remainder of the columns can be interpreted similarly. Notice also that attribute 1 (row 1) is measured by all 15 items. A zero indicates the item does not measure an attribute. An examinee is considered to have mastered all of the required attributes an item measures when the examinee answers that item correctly. By implication, the examinee is also expected to correctly answer all those items

that require the subset of attributes measured by the correctly-answered item. The HCI is used to assess whether an examinee uses different knowledge and skills (or the same knowledge and skills identified in the attribute hierarchy but in a different hierarchical combination) from what the attribute hierarchy predicts by comparing an observed examinee response vector to the reduced Q-matrix. The HCI for examinee i is given by:

$$HCI_i = 1 - \frac{2 \sum_{j \in S_{correct_i}} \sum_{g \in S_j} X_{i_j} (1 - X_{i_g})}{N_{c_i}},$$

where

$S_{correct_i}$ includes items that are answered correctly by examinee i ,

X_{i_j} is examinee i 's score (1 or 0) to item j ,

S_j includes items that require the subset of attributes measured by item j ,

X_{i_g} is examinee i 's score (1 or 0) to item g where $g \in S_j$, and

N_{c_i} is the total number of comparisons for all the items that are answered correctly by examinee i .

The term $\sum_{j \in S_{correct_i}} \sum_{g \in S_j} X_{i_j} (1 - X_{i_g})$ in the numerator of the HCI represents the number of misfits between

examinee i 's response vector and the reduced Q-matrix. If examinee i answers item j correctly,

then $X_{i_j} = 1$, and the examinee is also expected to answer item g correctly, which belongs to S_j , namely,

$X_{i_g} = 1$ ($g \in S_j$). If the examinee fails to answer item g correctly, then $X_{i_g} = 0$, $X_{i_j} (1 - X_{i_g}) = 1$, and it

is considered a misfit of the response vector i to the reduced Q-matrix. Thus, $\sum_{j \in S_{correct_i}} \sum_{g \in S_j} X_{i_j} (1 - X_{i_g})$ is

equal to the total number of misfits. The denominator of the HCI, N_{c_i} , contains the total number of

comparisons for items that are answered correctly by examinee i . When the numerator of the HCI is set to

equal the total number of misfits multiplied by 2, the HCI has the property of ranging from -1 to +1, which

makes for ease of interpretation. An HCI value of -1 indicates the maximum misfit of the observed response

vector to the reduced Q-matrix (and attribute hierarchy), while a value of +1 indicates that the observed response vector fits the reduced Q-matrix (and attribute hierarchy) perfectly. Although one is unlikely to obtain HCI values that are precisely -1 or +1, mean HCI values greater than 0.8 suggest an excellent fit, values between 0.6 and 0.8 indicate a moderate fit, and values less than 0.6 suggest a poor fit (Cui, 2007; see also Cui et al., 2006). Although HCI values can vary with the number and structure of attributes in a hierarchy, in practice, higher HCI values are better than lower HCI values when a single attribute hierarchy (and reduced Q-matrix) is fit for comparative purposes to different types of data sets (see Cui, 2007).

To illustrate the calculation of the HCI, note again the reduced Q-matrix shown in Table 1 of order (7, 15). Consider the observed examinee response vector (110001000000000) where items 1, 2, and 6 are answered correctly, namely $S_{correct_i} = \{1, 2, 6\}$. According to the reduced Q-matrix, item 6 (column 6) measures the attributes 1, 2, 4, and 5. Since examinee i answers item 6 correctly, he or she is considered to have mastered the attributes required by this item. Therefore, examinee i is expected to also answer items 1, 2, and 4 correctly, each of which measures a subset of attributes required by item 6. That is, $S_6 = \{1, 2, 4\}$. Therefore, for item 6, there are three comparisons: item 6 vs. items 1, 2, and 4. Since examinee i failed to answer item 4 correctly, $X_{i_6}(1 - X_{i_4}) = 1$, a misfit between the examinee's response vector and the reduced Q-matrix is found. Similarly, the items that measure the subset of attributes required by each of the other correctly answered items (i.e., items 1 and 2) are also identified, namely, $S_2 = \{1\}$ and $S_1 = \{\}$; the latter being an empty set containing no elements. Since S_2 contains item 1, which was answered correctly by examinee i , no misfit is found for item 2. Given that S_1 is an empty set containing no elements, no comparison is made for item 1. Overall, the total number of misfits is 1, and the total number of comparisons is equal to $3 + 1 + 0 = 4$. Hence, the value of the HCI equals to $1 - \frac{2 \times 1}{4} = 0.5$.

Cui (2007) conducted simulation studies to assess the effectiveness of the HCI in evaluating the degree to which an observed response vector fits the attribute hierarchy used in the AHM. Data were generated based on nine different attribute hierarchies by randomly adding 5, 10, and 20 percentages of slips to the expected response vectors associated with each hierarchy for three sample sizes – 500, 1,000, and

1,500. To assess the effectiveness of the HCI, the hypothesis was posited that data sets with a lower percentage of slips should produce higher HCI values than data sets with a higher percentage of slips. Simulation results supported this hypothesis indicating that higher and lower HCI values were produced consistently by data sets with a lower and higher percentage of slips, respectively, across different simulation conditions. Hence, it was concluded that the HCI is effective in determining the degree to which observed response vectors are consistent with the attribute hierarchy.

Two Cognitive Models of Algebra Performance: Expert Analysis and Verbal Analysis

The attribute hierarchy (i.e., the cognitive model used in the AHM) describes the knowledge and skills underlying test performance and is ideally used a priori to develop test items. Although the attribute hierarchy should be specified prior to developing test items (see stage 2 of AHM discussed in The Attribute Hierarchy Method (AHM) section), it is still possible to use the AHM to classify student test performance on an existing test; that is, a test that has not been developed from an attribute hierarchy. In this case, stage 2 of the AHM is bypassed and an attribute hierarchy is developed, *post hoc*, to account for examinee performance. To develop an attribute hierarchy post hoc, either expert or verbal analysis (see The Attribute Hierarchy as a Cognitive Model of Task Performance section) can still be used. While verbal analysis might be considered a better approach than expert analysis because verbal analysis represents a data-driven approach instead of a theoretical approach (see Leighton, 2004; Nichols, 1994), we are not aware of any studies comparing the adequacy of attribute hierarchies generated from these two approaches. Consequently, for the present study, an attribute hierarchy developed from expert analysis was compared and tested against an attribute hierarchy developed from verbal analysis to determine which approach led to better prediction of examinee performance on a sample of algebra I and II SAT™ items. The AHM and HCI were then used to test the accuracy of these two hierarchies in accounting for examinee performance on the SAT™ items.

Materials: Algebra Test items. Two hierarchies or cognitive models were developed to represent the knowledge and skills underlying performance on eight algebra SAT™ test items. Eight out of 21 algebra I and II items from the March 2005 administration of the SAT™ were chosen because initial expert analysis (see First cognitive model of algebra performance: Expert analysis section) indicated that the knowledge and skills of only these eight items could be represented within a single cognitive model. In order to include all 21 items, multiple cognitive models would have been required to represent the diversity of knowledge and skills.

Multiple models cannot be incorporated in the AHM or any other method premised on the specification of reduced Q-matrices at the present time. Of the eight items, seven were multiple choice (MC) and one was constructed response (CR). Difficulty values were 0.82, 0.43, 0.39, 0.92, 0.90, 0.88, 0.15, and 0.41, respectively. All items were scored dichotomously.

First cognitive model of algebra performance: Expert analysis. An expert in mathematics and statistics developed the first cognitive model of algebra performance. The expert had a bachelor's degree in mathematics and a master's degree in statistics, and was studying to complete a doctorate in measurement and evaluation. The expert originally reviewed the 21 test items to identify the knowledge and skills that an examinee would be expected to possess to solve each SAT™ item correctly. Although models of mathematical cognition exist in the educational and psychological literature, the expert did not use any of these models because none pertained explicitly to the 21 SAT™ items used in the present study (see footnote 1). When reviewing the items, the expert was not instructed or encouraged to develop the model for a particular student ability group. The expert was instructed only to identify the knowledge and skills required by an examinee (*any* examinee) to respond to the items correctly.² The review of the 21 items revealed that a single cognitive model could be created to represent the knowledge and skills an examinee could be expected to use for solving only eight of the 21 items. The remaining 13 items, some of which included *insight* algebra items (i.e., items requiring the sudden perception of an unstated, but key piece of information for solution) reflected a greater diversity of knowledge and skills that required unique models to be specified. After identifying the knowledge and skills for the eight items, the expert grouped them into attributes, and ordered the attributes into a hierarchy based on their logical properties.

The cognitive model of algebra performance developed from expert analysis (*expert-based*) is shown in Figure 2. The expert-based cognitive model consisted of eight attributes and in ascending order of complexity included: (1) prerequisite skills, (2) linear functions, (3) quadratic functions, (4) simple substitutions, (5) complex substitutions, (6) simple exponential computations, (7) complex exponential computations, and (8) representations. The reduced Q-matrix derived from the expert-based cognitive model is shown in Table 2 and is of order (8 by 8). The first column of the reduced Q-matrix is interpreted as showing that examinees need to possess attributes 1, 2, and 6 to solve item 1 correctly. The last column of the reduced Q-matrix shows that examinees need to possess attributes 1, 2, and 8 to respond to item 8 correctly.

Second cognitive model of algebra performance: Verbal analysis. The second cognitive model of algebra performance was developed based on examinee think-aloud verbal reports. In November 2005, 21 students (12 males and 9 females) enrolled in schools in New York City were administered 21 SAT™ items in algebra I and II. The students were sampled from the pool of New York City students who had taken the PSAT in grade 10, had opted-in to the Student Search Service at the College Board, and were likely en route to take the SAT™ in the next year. The following constraints were also used to narrow the pool of potential participants: (a) students must not have required an accommodation for the PSAT, (b) students must have scored between 550 and 650 on the Math portion of the PSAT, and (c) students must have scored between 600 and 800 on the Critical Reading portion of the PSAT. Using these constraints, a statistical analyst at the College Board identified a pool of 75 male and 75 female students of *moderately-high ability* in the database. Students of moderately-high ability were sampled for inclusion in the interview study to ensure a high proportion of correct item responses and fluent verbalizations. Although the 150 students were contacted by mail, only 26 students agreed to participate. Of these 26 students, 21 students attended the interview session held at the College Board in New York City (for a complete description of the procedure, see Gierl, Leighton, Wang, Zhou, Gokiert, & Tan, 2006). Participating students were compensated with 50 dollars and reimbursement of their transportation expenses to and from the College Board.

The students were individually interviewed in a quiet conference room at the College Board. The following instructions were used to begin each interview:

Thank you for agreeing to participate in today's study. Please know that your participation is completely voluntary and you are free to go at any time. Now, let me explain what we will be doing today for about 90 minutes.

In this study we are interested in what goes through your mind or what you think about when you find answers to SAT™ questions in math. In order to do this I'm going to ask you to THINK ALOUD as you work on the problems given. What I mean by think aloud is that I want you to tell me EVERYTHING you are thinking from the time you first see the question until you give an answer.

I would like you to talk aloud CONSTANTLY from the time I present each problem until you have given your final answer to the question. I don't want you to try to plan out what you say or try to

explain to me what you are saying. Just act as if you are alone in the room speaking to yourself. It is most important that you keep talking. If you are silent for any long period of time I will remind you to talk. Do you understand what I want you to do?

I will tape record our session because I want to get an accurate record of your think aloud reports. Please know that all the information you share today with me will be kept confidential and anonymous. Do you have any questions?

Students were asked to “think aloud” as they solved the 21 items and their verbalizations were audio-taped. A single 90 minute tape was used for each participating student. Standard probes introduced by Ericsson and Simon (1993) were used to elicit students’ verbalizations if they remained quiet for an interval longer than 20-25 seconds. Following the interviews, two raters were hired to listen to the audio-tapes and create flowcharts of the knowledge and skills each student employed to solve each of the 21 items for a total of 21 students X 21 items = 441 flowcharts.

Following the creation of flowcharts, a pre-service teacher with Bachelor degrees in engineering and education, and enrolled in a Master’s program in measurement and evaluation reviewed the flowcharts corresponding to the eight items used in the expert analysis (see Chi, 1997; and First cognitive model of algebra performance: Expert analysis section). The pre-service teacher had access only to the student flowcharts and was not given access to the algebra SAT™ items; thus, unlike expert analysis, the pre-service teacher was unaware of the specific content or format of the test items. The pre-service teacher reviewed the flowcharts of the 21 students (see Gierl et al., 2006) across the eight items and identified common attributes, inductively, in the reported knowledge and skills. The pre-service teacher then reviewed the flowcharts again and based upon the psychological ordering evident in the student flowcharts ordered the attributes hierarchically. The attribute hierarchy created by the pre-service teacher represented the knowledge and skills of the majority of 21 students.

The cognitive model of algebra performance developed by the pre-service teacher from student verbal reports (*student-based model*) is shown in Figure 3. The student-based cognitive model consisted of the following ten attributes in ascending order of complexity: (1) reading, (2) linear expressions, (3) primary substitution, (4) context independent expression generation for non-word problems, (5) context independent

expression generation for word problems, (6) quadratic expressions, (7) secondary substitution, (8) context dependent expression generation for word problems, (9) exponential expressions, and (10) multiple variable expression manipulation. The reduced Q-matrix for the student-based cognitive model, of order (10 by 8), is shown in Table 3 and is interpreted similarly to the expert-based reduced Q-matrix. The first column of the student-based reduced Q-matrix is interpreted as showing that attributes 1, 2, and 4 are required for examinees to solve item 1 correctly. The last column of this reduced Q-matrix shows that correct responses to item 8 require examinees to possess attributes 1, 5, 8, and 10.

Observed item response vectors. The AHM and the HCI were used to evaluate statistically the accuracy of the expert-based and student-based models (and associated reduced Q-matrices) for predicting examinee performance on the eight SAT™ items. Three data sets were used for this purpose. The first data set consisted of the sample of 21 examinees who had been interviewed in New York City, and whose verbal reports had been analyzed to create the student-based cognitive model. The second data set consisted of a random sample of 5000 examinees whose item responses were extracted from the original population of respondents from the March 2005 administration of the SAT™. The third data set consisted of a random sample of 100 *moderately-high ability* examinees drawn from the sample of 5000 (second data set) based on a similar frequency of total scores as the sample of 21 examinees (first data set).

Results

As described in the Method section, two cognitive models (attribute hierarchies) based on expert and verbal analysis, were developed (see Figures 2 and 3). The reduced Q-matrices (i.e., expert-based and student-based) associated with each of these two models were compared to the observed response vectors of examinees from three data sets ($n=21$, $n=5000$, $n=100$). A mean HCI value was computed to indicate the fit between each cognitive model (expert-based versus student-based) with each corresponding data set. The means and standard deviations of HCI values for the expert-based and student-based models across the three data sets are summarized in Table 4.

As shown in Table 4, the highest mean HCI value was obtained when the student-based cognitive model (and associated reduced Q-matrix) was fit to the observed response vectors of the 21 students interviewed in New York City. A mean HCI value of 0.95 indicates a very strong fit between the student-based cognitive model and the response data, indicating that the hierarchical arrangement of attributes induced

from student verbal reports predicted the response patterns of these 21 examinees very well (see Hierarchy Consistency Index [HCI] section). In comparison, the student-based model did not predict very well the observed response vectors of the 5000 examinees randomly sampled from the population of SAT™ test-takers (mean HCI=0.48). However, when 100 examinees of moderately-high ability, similar to the 21 examinees interviewed, were sampled from the group of 5000 test-takers, the student-based model again predicted the observed response vectors of these 100 students very well (mean HCI=0.93).

Turning to the expert-based model and its associated reduced Q-matrix, the expert-based model provided a better fit of the observed response vectors of the 5000 SAT™ examinees (mean HCI=0.71) than the student-based model (mean HCI=0.48). The expert-based model also provided a reasonably good fit of the observed response vectors associated with the 100 moderately-high ability SAT™ examinees (mean HCI=0.74), and the 21 moderately-high ability students interviewed for the present study (mean HCI=0.82). In general, whereas the expert-based model provided a better fit of the observed response vectors of the 5000 randomly sampled examinees³, the student-based model provided a better fit of the observed response vectors of the 21 and 100 moderately-high ability examinees. These results are elaborated in the Discussion section.

Discussion

The objective of the present paper was to compare two cognitive models of algebra performance, one developed from expert analysis and the other developed from verbal analysis, for predicting examinee knowledge and skills on eight algebra I and II items from the March 2005 administration of the SAT™. The AHM and HCI results suggested that both expert analysis and verbal analysis were useful for generating cognitive models to account for examinee response data. Although the highest mean HCI values were obtained when the student-based model was used to predict the eight item responses of the 21 and 100 moderately-high ability examinees (HCI = 0.95, 0.93, respectively), the lowest mean HCI value was obtained when this same model was used to predict the responses of the 5000 examinees randomly sampled from the March 2005 administration of the SAT™ (HCI = 0.48). HCI values can range between -1 and +1. Although the differences in HCI values in Table 4 are not large relative to their standard deviations⁴, an average value of 0.48 indicates a comparatively weak fit between the student-based model and the sequence of observed responses of the 5000 randomly sampled examinees compared to the fit provided by the expert-based model

(HCI=0.71). However, the expert-based model did not fit the responses of the 21 and 100 moderately-high ability examinees as well as the student-based model (expert-based, HCI=0.82, 0.74 versus student-based, HCI=0.95, 0.93). Why would the student-based cognitive model provide a relatively excellent fit of the observed response vectors associated with examinees of moderately-high ability but not to the examinees of moderate ability? We consider this question in the next two sections by discussing the student-based model first because it led to the highest mean HCI values and then, second, the expert-based model.

Verbal Analysis: Student-Based Model of Algebra Performance

The poor fit between the student-based model and the responses of the sample of 5000 examinees might be attributable to the characteristics of the examinee sample used to generate the student-based model. The 21 examinees who provided verbal reports to develop the student-based cognitive model were selected to represent examinees of moderately-high ability en route to take the SAT™ (see Method section). As described previously, this sample was selected for moderately-high ability because students needed to be able to articulate their response processes, solve many of the algebra items successfully so that a cognitive model of (correct) algebra performance could be developed, but not solve the items automatically as to impede the collection of verbal reports (see Ericsson & Simon, 1993; Leighton, 2004). As a result, the verbal reports collected from these examinees provided information about the response processes that *moderately-high ability* examinees could be expected to use to solve the algebra test items. These response processes did not generalize well to the randomly drawn sample of 5000 examinees, which included low ability as well moderate ability examinees. However, when a random sample of 100 moderately-high ability examinees was drawn from the sample of 5000 based on a similar frequency of total scores as the sample of 21 examinees, the fit between the student-based model and this sample of 100 was strong (mean HCI=0.93). Thus, the verbal reports provided by examinees of moderately-high ability informed the development of a cognitive model of task performance about a specific type of student—*moderately-high ability students*.

That the response processes of moderately-high ability examinees did not generalize to the response processes of low- to moderate ability examinees is understandable. Research investigations of expert and novice problem solvers (Chi, Glaser, & Farr, 1988; Ericsson, 2006; Glaser, Lesgold, & Lajoie, 1987; Leighton & Dawson, 2001; Leighton & Sternberg, 2003; Mislevy, 2006; Sternberg & Pretz, 2005) suggest that experts not only have more knowledge and skills within a target domain than novices but these knowledge and skills are

often organized differently, usually more efficiently. In the development of expertise not only is there an accumulation of knowledge and skills, but also a restructuring of these knowledge and skills (e.g., Chi et al., 1988; Ericsson, 2006). This restructuring may be due to different processes. In his discussion of the implications of expertise research for assessment, Mislevy (2006, p. 283) states “developing expertise in a domain is marked by the increase of experiential and automatized domain-relevant cognition, and also by increasing of metacognitive capabilities.” Although the moderately-high ability examinees in the present study cannot be viewed as experts in the traditional sense of the term since they are not mathematicians (see Ericsson, 2006), they arguably have more knowledge and skills in mathematics than low- to moderate ability examinees. Relatively speaking, then, moderately-high ability examinees within a target domain may represent emerging experts (Sternberg, 1999) who maybe beginning to show the restructuring of their knowledge and skills within the domain.

The relative lack of fit between the student-based cognitive model and the observed responses of the 5000 examinees indicates a potential weakness in the AHM and other cognitive diagnostic psychometric techniques, which rely on a single cognitive model to represent examinee response processes. The use of a single cognitive model (attribute hierarchy) in the AHM reflects the assumption that one model can describe the response processes of *all* examinees, and that the performance of high ability and low ability examinees can be distinguished simply by the mere presence or absence of attributes within the single model. This is a handy assumption. However, the results from the present study suggest that this assumption may be unwarranted.

As mentioned previously, while examinees of moderately-high ability may possess more knowledge and skills (i.e., attributes) for responding correctly to test items in comparison to low- to moderate ability examinees, they may also have these knowledge and skills organized differently and more efficiently than low- to moderate ability examinees (Chi et al., 1988; Ericsson, 2006; Glaser et al., 1987; Leighton & Dawson, 2001; Leighton & Sternberg, 2003; Mislevy, 2006; Sternberg & Pretz, 2005). The upshot of these different organizations of knowledge and skills is that they may require specifying different cognitive models (attribute hierarchies) for examinees at different points along the ability scale (see Luecht, 2007). The logistics of specifying *ability-specific* cognitive models, however, may be quite challenging when one considers that low-ability participants are likely to answer many test items incorrectly. Incorrect or buggy answers could in

principle originate from a vaster array of knowledge and skills than correct answers, thus leading to many more cognitive models to represent the response processes of low ability examinees. Moreover, many low ability students may experience difficulties and even frustration when expressing their thoughts about items they do not understand (in the event of collecting verbal reports) and thus could justify why they do not respond correctly. Such justifications are unlikely to produce accurate models of task performance (see Ericsson & Simon, 1993). Notwithstanding these complexities, it is imperative to find solutions to the question of how to best identify the response processes of low-ability students lest we are left with the paradoxical outcome of not being able to provide diagnostic information to those very students who need it most urgently.

Expert Analysis: Expert-Based Model of Algebra Performance

The expert-based model did not fit the responses of the 21 and 100 examinees of moderately-high ability as well as the student-based model. However, the expert-model was shown to provide the better fit, relative to the student-based model, of the responses of the 5000 examinees (mean HCI=0.71 vs. 0.48). This result suggests that models derived from content expert analysis may be better calibrated than heretofore assumed (see Leighton, 2004; Lohman & Nichols, 2006).⁵ Recall that the content expert in our study did not have access to student verbal reports when generating the expert-based model, but only to the SAT™ test items. Thus, the expert created the cognitive model of algebra performance by only reviewing the test items and anticipating the knowledge and skills students would be expected to use to respond correctly. That the expert-based model provided a better account of the 5000 observed response vectors than the student-based model suggests that content experts may be able to accurately anticipate the response processes used by examinees on test items, at least in some domains (Lohman & Nichols, 2006; Nichols, 1994; Poggio et al., 2005). Although it is prudent to exercise caution when making assumptions about the response processes examinees use to respond to test items, developing cognitive models directly from student verbal reports may not always lead to better accounts of task performance.

The question that needs to be addressed at this stage is why the expert-based model provided a better fit of the observed responses of the 5000 examinees than the student-based model. Although a definitive answer to this question can be provided only with additional study, preliminary answers are explored presently. First, the expert developed a cognitive model of algebra performance using expert-

analysis based on the knowledge and skills *an examinee would be expected to possess* to respond to the items correctly. The expert was not instructed or encouraged to think of a specific type of student or ability level when designing the model. Because the expert was not instructed to develop a model for students at a specific ability level, it is possible that the knowledge and skills the expert used to develop the expert-based model were those expected of a student of average or moderate ability. A brief examination of the two cognitive models in Figures 2 and 3 suggest differences in the detail associated with the hierarchies of attributes. Although both the student-based and expert-based models have four “branches” of attributes relative to the prerequisite attribute, one of the branches in the student-based model holds up to four attributes, indicating a deeper structure than the expert-based model. None of the branches in the expert-based model holds more than two attributes. Consequently, the expert-based model may have included more basic level attributes than the student-based model, thus permitting a better account of the knowledge and skills that students of average ability used to respond correctly to algebra items. Second, the pre-service teacher who developed a cognitive model of algebra performance directly from student verbal reports was constrained by the reports provided by the students of moderately-high ability. Thus, implicitly, the teacher was developing a model targeted to moderately-high ability students. Future studies should be conducted to determine whether expert-based cognitive models targeted at specific levels of ability can produce mean HCI values comparable to those obtained from student-based models developed from verbal analysis. Instructing content experts to develop models representing students of specific ability levels may also be desirable because models with specific knowledge and skill attributes may be more useful for diagnostic inferences than those with overly basic or general knowledge and skills. Consider that the evidentiary support of diagnostic inferences is bound to be stronger from a cognitive model with a mean HCI value of 0.95 than from one with a mean HCI value of 0.82 if only because these values indicate relative model-data fit. Test developers, administrators, and practitioners are likely to prefer models with higher mean HCI values relative to models with lower mean HCI values. This preference is reasonable given that diagnostic inferences, by their very nature, require accurate and specific information about examinee knowledge and skills (Nichols, 1994). It is noted, however, that an area for future research is to explore the differences between mean HCI values and the extent of HCI variations due to chance alone.

Before concluding, several limitations of this study need to be highlighted. First, the two cognitive models developed for the present study were based on only eight of 21 SAT™ items. Only eight items were used because, as mentioned previously (see First cognitive model of task performance: Expert analysis section), these were the only items found to be sufficiently homogeneous in the knowledge and skills measured to be able to be referenced to a single cognitive model of task performance. That only eight of 21 items were used illustrates the challenge of developing cognitive models for items post hoc. An area of future study is to investigate how multiple cognitive models may be developed to reflect different sets of items or, alternatively, how items can be most effectively developed to reflect a single model of task performance. Second, the student-based cognitive model was generated by a pre-service teacher's interpretation of student verbal reports and not by students themselves. It is therefore possible that in the teacher's interpretation of the reports, he imposed his own knowledge and skills about algebra as he developed the student-based model. While this is certainly possible with any interpretative exercise, it is important to note that the teacher did not have access to the items but only to students' verbal reports. Future studies may consider amalgamating the interpretations of multiple judges when deciding on any final cognitive model—student-based or expert-based. For example, having more than one expert generate a cognitive model would bolster the external validity of the model. Student-based cognitive models developed by having students interpret verbal reports from other students, however, is not recommended because of unnecessary sources of uncontrolled variance—such as experiential and/or professional differences between students and adults. In such cases, poor model-data fit could be attributed to inadequate understanding of verbal reports by students. Third, it is important to note that differences in mean HCI values were not large relative to their standard deviations. Thus, future studies require an investigation of the stability of mean HCI values given that at the present time standard errors of the HCI are unknown.

Notwithstanding these limitations, the present study indicated that a student-based cognitive model strongly predicted the observed responses of moderately-high ability examinees on eight SAT™ algebra items relative to an expert-based model, which predicted the observed responses of moderate and moderately-high ability examinees with only reasonable accuracy. Although the expert-based cognitive model did not achieve the same level of predictive accuracy as the student-based model, expert-based models, developed in the absence of human studies, may still provide a reasonably accurate and efficient route to model development.

Nonetheless, if specific diagnostic inferences are sought about high-stake test performance, using a student-based cognitive model or an expert-based model corroborated with human studies (a hybrid approach) may be necessary for defensible test-based inferences, despite the cost and time associated with generating such models.

Footnotes

¹ Achievement test items are not traditionally developed from psychological theories and/or cognitive models with hierarchical structures. Instead, achievement test items have historically been developed from content specifications alone. Reviewing the psychological literature for a specific cognitive model that might function as an attribute hierarchy for a domain of achievement testing is often not useful. Psychological models often illustrate the knowledge and skills involved in narrow task domains such as categorical and conditional logic (Johnson-Laird & Bara, 1984), balance scale (Siegler & Chen, 1998), or matrix problem solving (Carpenter, Just, & Shell, 1990) and not achievement testing. For this reason, application of the AHM to existing achievement tests developed in the absence of an attribute hierarchy necessitates bypassing stage 2 and developing a hierarchy post hoc.

²In conducting the expert analysis, the expert needs to rely on his or her interpretation of the knowledge and skills examinees might use to meet the content standard required for a correct response to an item. Expert analysis does not involve a blind repetition of the content standards because these are often not hierarchically ordered. Thus, expert analysis is an interpretative process based on the expert's experience with the content domain and the knowledge and skills assumed to operate within that domain.

³ It is unlikely that the HCI results—especially with respect to student-based model—presented in Table 4 are the result of a “type-token” ratio where any countable successes given some number of events will tend to exhibit declining proportions as the number of events increases. Although the student-based model included 10 attributes and the expert-based included 8 attributes, the mean HCI value of the student-based model in fact indicated a better fit than the expert-based model to both the 21 and 100 sample groups. Nevertheless, further study of the factors influencing variations in the HCI is required (see Cui, 2007).

⁴ Another way to assess the adequacy of the two cognitive models for predicting examinee performance would be to test the attributes shown in the hierarchies and included in the reduced Q-matrices as factors in a confirmatory factor analysis (CFA). This was not done in the present paper because the hierarchies included the same number or more attributes than test items, thereby compromising the degrees of freedom for the CFA.

⁵ Readers might assume that the cognitive model created by our expert was advantaged relative to cognitive model created by the pre-service teacher because the expert was the one to decide on the number of items (8

of 21) to be considered for model development. Based upon results from our study we do not think this is the case. The expert-based model did not lead to better average HCI values for two out of the three data samples (as shown in Table 4). If anything, it would appear from these HCI values that the student-based model was advantaged because it produced higher average HCI values for two out of three samples. Moreover, we think it is important to note that confirmatory studies generally require making assumptions about the parameters used to guide data generation, analysis, and reporting of results. In other words, confirmatory studies by their nature begin by advantaging sets of theories or models to be tested because these are chosen from a larger pool of possibilities. In our study, the expert analysis functioned like an a priori model that was verified against a model derived from student reports.

References

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Author. Washington, DC.
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061-1071.
- Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*, 33-63.
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review*, *97*, 404-431.
- Chi, M.T.H. (2006). Laboratory methods for assessing experts' and novices' knowledge. In K.A. Ericsson, N. Charness, P.J. Feltovich, & R.R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 167-184). Cambridge University Press.
- Chi, M.T.H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, *6*, 271-315.
- Chi, M. T. H., Glaser, R., & Farr, M. J. (Eds.). (1988). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cui, Y. (2007). *The hierarchy consistency index: A person-fit statistic for the attribute hierarchy method*. Unpublished doctoral dissertation. University of Alberta, Edmonton, Alberta, Canada.
- 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.
- Embretson, S. E. (1999). Cognitive psychology applied to testing. In F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay, & M. T. H. Chi (Eds.), *Handbook of applied cognition*, (pp. 629-66). New York: Wiley.

- Embretson, S. (1998). A cognitive design system approach to generating valid tests. Application to abstract reasoning. *Psychological Methods, 3*, 380-396.
- Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*, 343-368.
- Ericsson, K.A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K.A. Ericsson, N. Charness, P.J. Feltovich, & R.R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 223-241). Cambridge University Press.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.
- Gierl, M. J., Leighton, J. P., & Hunka, S. (2000). Exploring the logic of Tatsuoaka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice, Fall*, 34-44.
- Gierl, M.J., Leighton, J.P., Wang, C., Zhou, J., Gokiert, R., & Tan, A. (2006). *Validating the cognitive model underlying examinees' algebra performance on the SAT®*. Technical Report #4. College Entrance Examination Board.
- Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. Ronning, J. Glover, J.C. Conoley & J. Witt (Eds.), *The influence of cognitive psychology on testing and measurement: The Buros-Nebraska symposium on measurement and testing* (Vol. 3, pp. 41-85). Hillsdale, NJ: Erlbaum.
- Hamilton, L.S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*, 181-200.
- Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition, 16*, 1-61.
- Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Washington, DC: American Council on Education.
- Leighton, J. P. (2004). Avoiding misconceptions, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*, 6-15.

- Leighton, J. P., & Dawson, M. R. W. (2001). A parallel processing model of Wason's card selection task. *Cognitive Systems Research, 2-3*, 207-231.
- Leighton, J.P., & Gierl, M.J. (2007). Verbal reports as data for cognitive diagnostic assessment. In J.P. Leighton & M.J. Gierl (Eds.), *Cognitive diagnostic assessment for education. Theory and applications* (pp. 146-172). Cambridge, MA: 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.
- Leighton, J. P., Gokiert, R.J., & Cui, Y. (2007). Using exploratory and confirmatory methods to identify the cognitive dimensions in a large-scale science assessment. *International Journal of Testing, 7*, 1-49.
- Leighton, J. P., & Sternberg, R. J. (2003). Reasoning and problem solving. In A. F. Healy & R. W. Proctor (Volume Eds.), *Experimental psychology* (pp. 623-648). Volume 4 in I. B. Weiner (Editor-in-Chief) *Handbook of psychology*. New York: Wiley.
- Lohman, D.F. (2000). Complex information processing and intelligence. In R.J. Sternberg (Ed.), *Handbook of intelligence* (pp. 285-340). NY: Cambridge University Press.
- Lohman, D. F & Nichols, P. (2006). Meeting the NRC panel's recommendations: Commentary on the papers by Mislevy and Haertel, Gorin, and Abedi and Gandara. *Educational Measurement: Issues and Practice, 25*, 58-64.
- Luecht, R.M. (2007). Using Information from Multiple-Choice Distractors to Enhance Cognitive-Diagnostic Score Reporting. In J.P. Leighton & M.J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 319-340). Cambridge, UK: Cambridge University Press.
- Mislevy, R.J. (2006). Cognitive psychology and educational assessment. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 257-305). 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.
- Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 438-483.
- Nichols, P. (1994). A framework of developing cognitively diagnostic assessments. *Review of Educational Research, 64*, 575-603.

- Poggio, A., Clayton, D. B., Glasnapp, D., Poggio, J., Haack, P., & Thomas, J. (April, 2005). *Revisiting the item format question: Can the multiple choice format meet the demand for monitoring higher-order skills?* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.
- Rindermann, H., & Neubauer, A. C. (2004). Processing speed, intelligence, creativity, and school performance: Testing of causal hypotheses using structural equation models. *Intelligence, 32*, 573-589.
- Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. NY: Oxford University Press.
- Siegler, R.S. & Chen, Z. (1998). Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology, 36*, 273-310.
- Sternberg, R.J. (1999). Intelligence as developing expertise. *Contemporary Educational Psychology, 24*, 359-375.
- Sternberg, R.J. & Pretz, J.E. (2005). (Eds.). *Cognition & Intelligence*. Cambridge, UK: Cambridge University Press.
- Taylor, K. L., & Dionne, J-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92*, 413-425.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

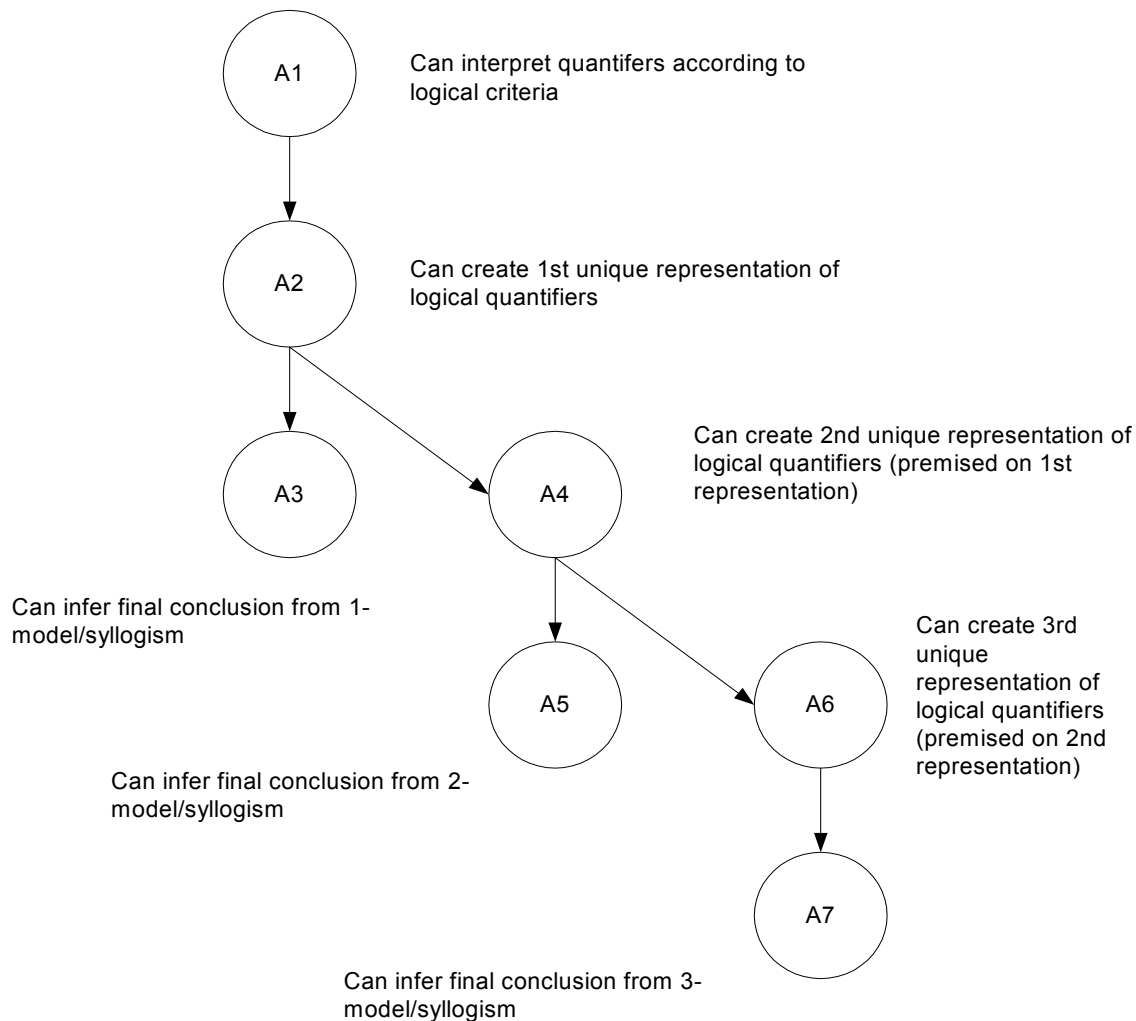


Figure 1. A seven-attribute hierarchy of categorical syllogism performance reproduced with permission from Leighton, J.P., Gierl, M.J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement*, 41, 205-236. Note attribute 1: The interpretation of quantifiers according to formal criteria; attribute 2: The ability to create an initial model or representation of the premises; that is, combining the representation of the first and second premises into a whole representation; attribute 3: The ability to draw a conclusion from the initial model created from the premises; attribute 4: The ability to generate a second unique model of the premises; that is, to generate another representation of the premises; attribute 5: The ability to generate a conclusion that is consistent with the initial model and this second model of the premises; attribute 6: The ability to generate a third unique model of the premises; and attribute 7; The ability to generate a conclusion that takes into account all three models of the premises.

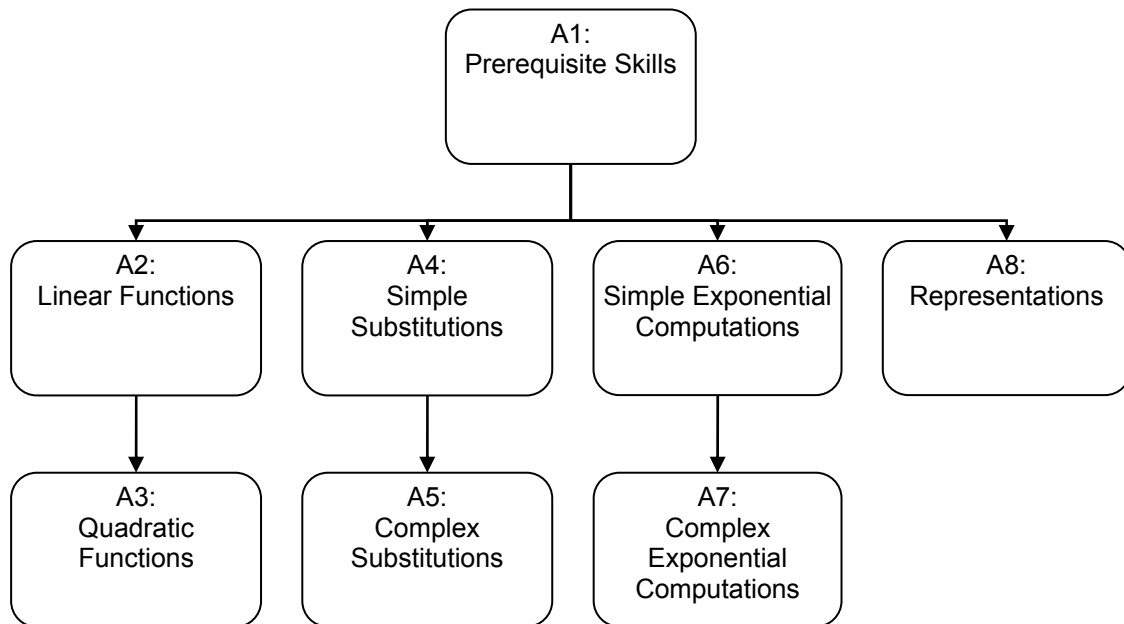


Figure 2. The expert-based cognitive model developed from expert analysis. Note: attribute 1: The understanding of the arithmetic operations implied by $+$, $-$, \times , $/$, $=$, absolute value, square, square root, exponent, $>$, $<$, \leq , \geq , and signed numbers; The ability to carry out basic computations, such as addition, subtraction, multiplication and division of whole numbers; attribute 2: The ability to solve linear functions; attribute 3: The ability to factor quadratic expressions and solve quadratic functions; attribute 4: The ability to substitute the value of a variable for the letter; attribute 5: The ability to substitute abstract expressions and rules; attribute 6: The ability to carry out basic exponential computations, such as multiplication and division with two terms; attribute 7: The ability to carry out more complicated exponential computations, such as multiplication and division with more than two terms; and attribute 8: The ability to translate words into mathematical expressions.

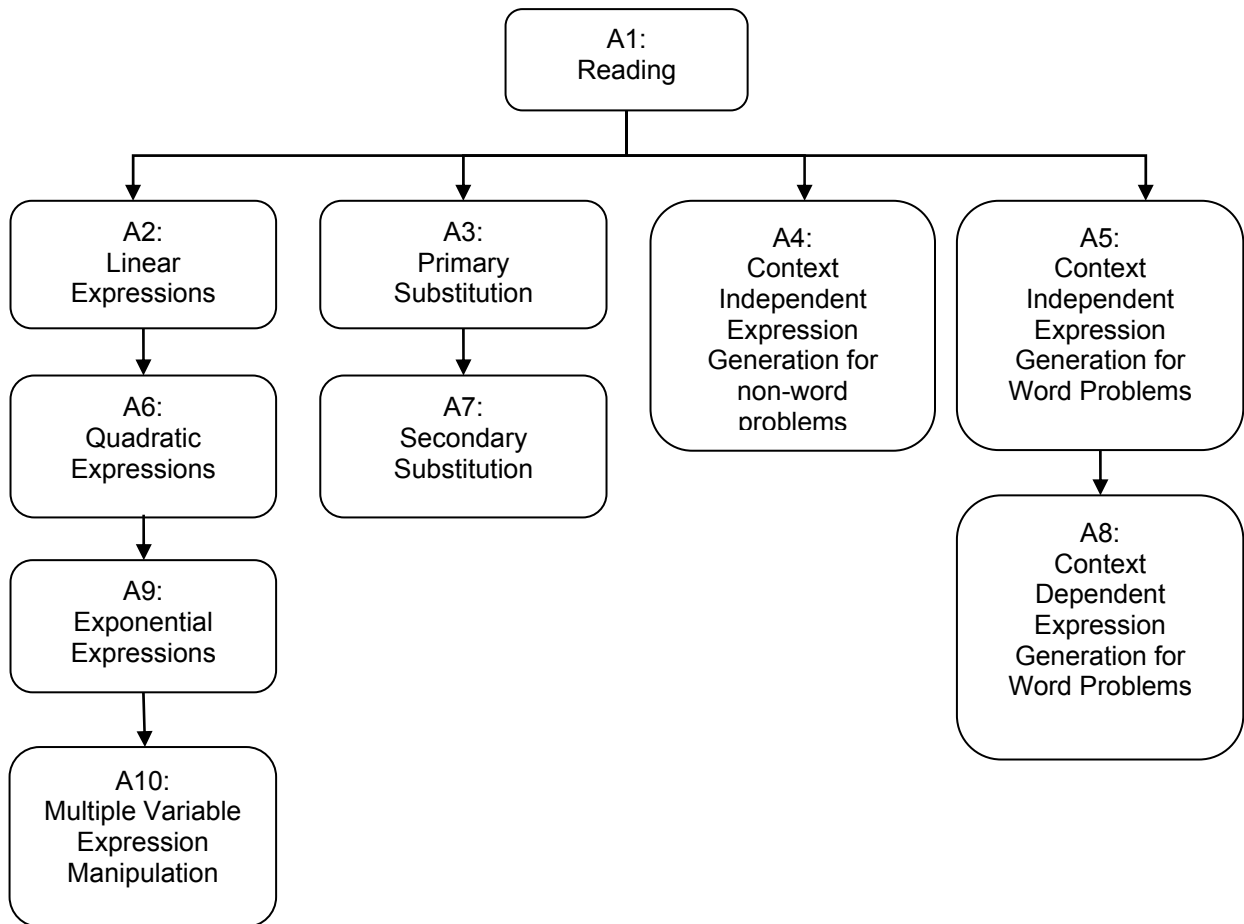


Figure 3. The student-based cognitive model developed from verbal analysis. Note attribute 1: Ability to read the question; attribute 2: Ability to manipulate linear algebraic expressions; attribute 3: Ability to perform primary substitution; attribute 4: Ability to generate algebraic expressions using previous knowledge; attribute 5: Ability to generate an algebraic expression from a context independent word problem; attribute 6: Ability to manipulate quadratic algebraic expressions; attribute 7: Ability to perform secondary substitution; attribute 8: Ability to generate an algebraic expression from a context dependent word problem; attribute 9: Ability to manipulate exponential algebraic expressions; and attribute 10: Ability to manipulate linear algebraic expressions with more than two variables.

Table 2

Reduce Q-Matrix for Expert-Based Model (Attribute Hierarchy Shown in Figure 2)

		ITEMS							
ATTRIBUTES		1	1	1	1	1	1	1	1
		1	0	0	1	1	0	0	1
		0	0	0	1	0	0	0	0
		0	0	1	0	0	1	0	0
		0	0	1	0	0	0	0	0
		1	1	0	0	0	0	0	0
		0	1	0	0	0	0	0	0
		0	1	0	1	0	0	1	1

Table 4

The mean and standard deviation of the HCI values as a Function of Cognitive Model and Sample Group

Cognitive Model	Data Set Sample Size		
	21	5000	100
Verbal-Analysis	0.95 (0.15)	0.48 (0.60)	0.93 (0.18)
Expert-Analysis	0.82 (0.19)	0.71 (0.46)	0.74 (0.19)

Note. Observed response vectors in each data set reflect answers to eight SAT™ items.