University of Alberta

# INVESTIGATING THE COGNITIVE PROCESSES UNDERLYING STUDENT PERFORMANCE ON THE SAT® CRITICAL READING SUBTEST: AN APPLICATION OF THE ATTRIBUTE HIERARCHY METHOD

by

Changjiang Wang ©

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Measurement, Evaluation and Cognition

Department of Educational Psychology

Edmonton, Alberta
Fall 2007

*Your file Votre référence*
*ISBN: 978-0-494-33087-6*
*Our file Notre référence*
*ISBN: 978-0-494-33087-6*

# Abstract

Educational tests are designed to facilitate teaching and learning. However, due to the disjunction between cognitive psychology and educational testing, most large-scale tests typically yield very limited information for teachers, students, and parents about why some students perform poorly or how instructional conditions can be modified to improve teaching and learning (National Research Council, 2001). Although researchers have made progress in modeling the cognitive processes of reading activities (e.g., Butcher & Kintsch, 2003; VanderVeen, et al., 2007), few studies have been conducted to investigate the integration of these models in reading comprehension tests. In the present study, the cognitive processes underlying student performance on the SAT Critical Reading subtest were investigated using a recently developed cognitively-based psychometric approach, the Attribute Hierarchy Method (AHM) (Leighton, Gierl, & Hunka, 2004; Gierl, Cui, & Hunka, in press). The study was conducted in two stages: the cognitive analysis and the psychometric analysis. Three attribute hierarchies for the SAT Critical Reading subtest were first developed. Then, the three attribute hierarchies were validated using student verbal reports and the hierarchy consistency index (*HCI*). Student verbal reports supported the validity of the attributes specified in the hierarchies and the hierarchical relationships among the attributes. Moreover, two additional attributes were discovered. The subsequent *HCI* analysis indicated that Hierarchy 2 had the best model-data fit and, hence, it was used for analyzing student response data in the subsequent psychometric analysis. In the psychometric analysis, attribute probabilities were first calculated, using a neural network, for a sample of 15 students who took the March 2005 administration of the SAT. Reliabilities and standard error of measurement for each of the nine attributes were also estimated. Based on attribute probability and reliability results, exemplar score reports were provided for a sample of students. The results of the study indicate that, with the AHM, cognitive diagnostic information can be extracted from SAT Critical Reading subtest to enhance score reporting and, potentially, guide teaching and learning. Moreover, in the framework of the AHM, important information about the construct underlying student performance on the SAT Critical Reading subtest can be identified and evaluated.

# Acknowledgements

My profound appreciation and sincere thanks go to my supervisor, Dr. Mark Gierl. Mark has been a great mentor both in my academic work, professional development, and in shaping my personality. He constantly pushes me to come up with research proposals, apply for awards, and write for publication. His insight and expertise in educational measurement, his highly-efficient working style, and his philosophy of "thinking big" were, and will continually be, the guidance in my professional development. His "grace under pressure" sets an excellent example for me to learn from. I am also deeply indebted to him for his heart-warming encouragement whenever I need courage to go on. My heartfelt gratitude toward him is far beyond any verbal expression.

My heartfelt appreciation also goes to Dr. W. Todd Rogers, whose wisdom, expertise, quick response, and unselfish support shaped my work and made my doctoral studies and dissertation writing a rewarding experience. He led me, with interesting lectures and challenging questions, into the fields of educational measurement and statistics and pushed me to explore deeper. He also went through numerous drafts for my dissertation to make it a piece with high quality.

I am very grateful to Dr. Jacqueline Leighton for her wonderful lectures in univariate statistics, cognition, and educational assessment. Her teaching style led me to read widely and think critically. Sitting in her lectures is always exciting, challenging, and rewarding. Her expertise in cognition and assessment has made a unique contribution to my dissertation.

I would like to express my sincere thanks to Dr. Rauno Parrila, whose expertise in reading has contributed significantly to my dissertation. His extensive knowledge in reading and exemplary scholarship have impacted my research in many positive ways and will make a unique contribution to my professional career.

My thanks also go to the rest of my dissertation committee, Dr. Christina Gagné and Dr. Joanna Gorin, for their helpful comments and suggestions on various aspects of this dissertation.

I would like to express my gratitude to all the graduate students in CRAME for their comradeship and friendship.

I would like to thank the Social Sciences and Humanities Research Council and the College Examination Board for sponsoring my dissertation research. I am also appreciative of the College Examination Board for providing me with the SAT test materials and data.

I would like to thank my parents and my father- and mother-in-laws. Without their whole-hearted support, my dissertation would not have been possible. My heartfelt thanks go to my beloved wife, Lingyun, for her continued support for me during this physically and emotionally trying period. She is always caring, loving, and understanding. I am also very grateful to Albert, my little sweetie. This guy upgraded me to Version Dad. His innocent love, smiling face, and cute movements brought boundless happiness and encouragement to my life.

## Table of Contents

## List of Tables

## List of Figures

# CHAPTER I: INTRODUCTION

## Context of the Study

Cognitive psychology is fundamental to educational measurement because most tests are based on cognitive problem-solving tasks. But, in many large-scale testing programs, the importance of understanding the psychology underlying student performance has been downplayed relative to the emphasis placed on using statistical models and psychometric techniques for scaling and scoring examinee performance (Glaser, 2000; Leighton, Gierl, & Hunka, 2004; National Research Council, 2001; Nichols, 1994). As a result, most large-scale tests typically yield very limited information for teachers, students, and parents about why some students perform poorly or how instructional conditions can be modified to improve teaching and learning (National Research Council, 2001).

Increasingly, however, researchers and practitioners are calling for the integration of cognitive psychology and educational measurement to enhance learning and instruction (Bejar, 1984; Embretson, 1998; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Gierl, Cui, & Hunka, in press; Leighton et al., 2004; Mislevy & Riconscente, 2006; National Research Council, 2001; Nichols, 1994; Sheehan, 1997; Snow & Lohman, 1989; Tatsuoka, 1995). A variety of cognitively-based psychometric approaches, which attempt to explicitly link principles of cognitive psychology with assessment practice in educational measurement, have been proposed and applied to a number of testing programs (e.g., Embretson, 1998; Leighton et al., 2004; Mislevy & Riconscente, 2006; Tatsuoka, 1995).

Cognitive psychology should exert its impact on different aspects of testing practice, such as the construction of tests, the interpretation of test scores, and the cognitive feedback provided to students. In test construction, a cognitive theory of how people develop competence in a content domain provides clues about the types of item features that would elicit evidence about the degree to which the students have mastered the relevant knowledge and cognitive skills (National Research Council, 2001). Test items developed based on such a cognitive theory would have features allowing test users to make valid inferences from student performance about their cognition in terms of relevant knowledge and skills necessary to answer the test items successfully (National Research Council, 2001). Then, based on the inferences drawn from student performance on the test items, cognitively diagnostic feedback could be provided to the students and teachers to enhance learning and instruction.

Methodologically, many of the research methods commonly used in cognitive psychology, such as verbal reports and expert review, have found their way into testing practice (National Research Council, 2001). These methods help test specialists understand students' knowledge representations in a content domain and their cognitive processes in problem solving. By applying these methods, important information can be obtained which can be used for test construction and analysis.

Although cognitive psychology is exerting its influence on testing practice, its impact to-date has been minimal (Leighton et al., 2004) and unbalanced. The studies which attempt to integrate cognitive theory into testing practice focus, mostly, on the domains of science and mathematics (e.g., Ayala, Shavelson, Yin, & Schultz, 2002; Glaser, 2000; Leighton, Rogers, & Maguire, 1999; Watermann & Klieme, 2002). Although cognitive research in the domain of reading has seen faster development than in many other domains (National Research Council, 2001), and remarkable progress has been made in modeling the cognitive processes of reading activities (e.g., Butcher & Kintsch, 2003; Graesser, Millis, & Zwaan, 1997; Perfetti, 1985), the integration of this progress into tests of reading has been slow. Moreover, even though some studies have integrated cognitive theories in reading into test models (Embretson & Wetzel, 1987; Gorin, Embretson, & Sheehan, 2002; Sheehan & Ginther, 2001), these models have not been widely applied to operational tests. As a result, little is known about what cognitive processes affect student performance on reading tests and hence little information is provided to students about their cognitive skills in reading.

## The Attribute Hierarchy Method

The attribute hierarchy method (AHM) (Leighton et al., 2004; see also Gierl, Cui, & Hunka, in press; Gierl, Leighton, & Hunka, 2000) represents an effort towards the integration of cognitive theory with testing practice. It is a cognitively-based psychometric method which classifies examinees' test item responses into structured attribute patterns according to a cognitive model of task performance. The AHM is based on the assumption that test performance depends on a set of hierarchically-related cognitive competencies called attributes.

In the framework of the AHM, the cognitive model of task performance, as operationalized using the attribute hierarchy, plays an essential role. It specifies the cognitive competencies required to solve test items and the relationships among these competencies. Once the attribute hierarchy is identified for a domain, test developers can create items according to the hierarchical organization of the cognitive attributes. By doing so, the test developer achieves control over the specific attributes each item measures (Leighton et al., 2004). Based on students' performance on the test items designed from the attribute hierarchy, inferences as to the students' strengths and weaknesses can be drawn and cognitive feedback can be provided in order for the students to make necessary remedial effort. However, as a relatively new cognitive psychometric method with desirable features, the AHM has not been extensively applied to testing practice.

## The SAT Critical Reading Subtest

The SAT is one of the best-known, large-scale high-stakes tests in North America. More than two million students take the SAT every year (The College Board, 2005). Their scores on the SAT are used by colleges and universities for making admissions decisions. The SAT measures critical thinking and reasoning skills required for college academic success in the areas of reading, mathematics, and writing. One of the goals of the College Board, the agency responsible for the SAT, is to provide individual students with feedback about their performance on the SAT (Huff, 2004). This goal represents the College Board's ongoing effort towards linking large-scale tests with teaching and learning.

The SAT Critical Reading subtest is designed to measure students' critical reading ability, or the ability to construct a coherent meaning representation of texts (passages, item stems, and response options) and to use these representations to constrain option choices on test items. This ability involves text-based processing, application of background knowledge, reasoning and problem solving, and meta-cognitive execution skills (VanderVeen, 2004).

Consistent with the goal of the College Board to provide feedback to students, the Critical Reading subtest must provide accurate information on students' mastery of cognitive skills in critical reading. To meet this requirement, the cognitive skills measured by the SAT critical reading items must be well understood. Surprisingly, studies which investigate these cognitive skills are rare (e.g., Burton, Welsh, Kostin, & van Essen, 2003; VanderVeen, 2004; VanderVeen et al., 2007).

## Purpose of Study

Thus, the purpose of the present study is to investigate the cognitive attributes underlying student performance on the SAT Critical Reading subtest using the AHM.

An emphasis is placed on the cognitive attributes included in the cognitive model of task performance proposed by VanderVeen et al. (2007) for the SAT critical reading items. Specifically, the study was designed to answer the following two research questions:

1. What cognitive attributes in the VanderVeen et al. (2007) model are involved in answering the SAT critical reading items, what additional attributes are involved, and how are these cognitive attributes related to each other?
2. How can the AHM be used to analyze student response data, report student scores, and provide diagnostic feedback?

To answer research question 1, a review of selected literature in reading is conducted to identify the cognitive attributes generally involved in critical reading and their interrelationships. Then, different attribute hierarchies with specific consideration for the attributes included in the VanderVeen et al. (2007) model for the SAT critical reading items are created. These hierarchies serve as competing cognitive models of task performance. A verbal report study and an analysis using the hierarchy consistency index is also conducted to validate these attribute hierarchies. Hierarchies providing the best fit to the data will be used for the psychometric analyses.

To answer research question 2, the AHM is used to analyze the student response data on an administration of the SAT Critical Reading subtest. Student response data are classified into different attribute patterns according to the hierarchies. Lastly, the features of the AHM that can enhance score reporting and provide cognitively diagnostic feedback are demonstrated.

## Organization of the Dissertation

This dissertation is organized into six chapters. Chapter I described the context for the study, provided a brief introduction to the AHM and the Critical Reading subtest of the SAT, and then presented the purpose of the study. Chapter II builds the theoretical framework for the present study. This chapter includes (a) a detailed introduction of the AHM, the psychometric approach used in the present study, and a review of studies that used the AHM in test analysis; and (b) a review of the cognitive processes involved in reading and the studies on the hierarchical relationships among the cognitive processes involved in reading. Chapter III describes the SAT Critical Reading subtest, the test used in the present study, and the methods and procedures used for evaluating the reading attribute hierarchies and for analyzing the test data. Chapter IV reports the results from the cognitive analysis and identifies the best-fitting hierarchy to be used in the psychometric analysis. Chapter V presents the results from the psychometric analysis. Chapter VI discusses the results and draws the conclusion of the study.

# CHAPTER II: LITERATURE REVIEW

In Chapter I, it was argued that cognitive psychology should be integrated into educational measurement in order to better understand the test performance of examinees and to provide meaningful feedback for their improvement in learning. To demonstrate this integration, the critical reading subtest of the SAT was analyzed using the AHM, a cognitively-based psychometric approach, in the present study. In this chapter, the methodology used in the present study and the cognitive theories in reading that led to the development of different cognitive hierarchies are reviewed to establish the theoretical framework for the study. This chapter is organized in two sections. Section 1 provides a detailed account of the AHM and a review of the studies in which the AHM was applied. Section 2 reviews the theories about the cognitive processes involved in reading and the studies in which the hierarchical relationships among these cognitive processes have been investigated.

## The Attribute Hierarchy Method

The AHM is a cognitively-based psychometric method which classifies examinees' test item responses into structured attribute patterns according to a cognitive model of task performance. In the AHM, cognitive attributes are assumed to be hierarchically related (Leighton & Gierl, 2007; VanderVeen et al., 2007). A cognitive attribute in the AHM is defined as a description of the procedural or declarative knowledge needed to perform a task in a specific domain (Leighton et al., 2004). For the present study, “attribute” is used as an umbrella term to refer to the cognitive processes and skills employed by students to correctly answer the SAT Critical Reading items.

One strength of the AHM lies in its facility to guide test development. Once the attribute hierarchies are identified for a content domain, test developers can create items according to the hierarchical organization of the attributes. By doing so, the test developer achieves control over the specific attributes each item measures. The AHM also offers a more convenient way of providing cognitive feedback to students. This feedback is achieved by mapping observed examinee response patterns onto expected examinee response patterns derived from the attribute hierarchy. A student with a certain observed response pattern is expected to have mastered the attributes implied by the corresponding expected response pattern, but may need more work on other attributes. As a cognitively-based psychometric approach, the AHM consists of two major components, the cognitive component and the psychometric component.

### *Cognitive Component of the AHM*

The cognitive component of the AHM refers to the specification of the attribute hierarchy, which includes the cognitive attributes measured by a test and the interrelationships among these attributes. Ideally, an attribute hierarchy is identified from theories in a content domain. However, in practice, due to the disjunction between testing and cognitive theories (National Research Council, 2001), few *a priori* theories could be found that could be used to specify the attribute hierarchy. In this case, the attribute hierarchy has to be identified retrospectively, using methods such as item reviews and examinee verbal reports on items that have already been developed (Gierl, Cui, & Hunka, 2006). It should also be noted that the specification of an accurate and valid hierarchy requires an iterative process. The initial hierarchy could come from relevant cognitive theory in a content domain or empirical studies. Then, the hierarchy should be validated using empirical test data. The results of validation and the subsequent revision of the hierarchy would result in a more accurate and valid hierarchy for future test development and analysis.

The specification of the attribute hierarchy is of primary importance in the AHM framework because the attribute hierarchy represents the construct for a test and, by extension, the cognitive attributes that underlie test performance. The attribute hierarchy is critical both for test development and for making valid inferences about student performance.

### *Psychometric Component of the AHM*

After attribute hierarchies are specified, psychometric procedures are required to apply the AHM in test analysis. These procedures include the representation of the attribute hierarchy, generation of attribute patterns and expected response patterns, classification of observed response patterns, and evaluation of attribute hierarchies by calculating attribute reliabilities and the hierarchical consistency index.

*Formal Representation of a Hierarchy*

The formal representations of an attribute hierarchy uses the matrices of the rule-space approach, including the adjacency, reachability, incidence, and reduced *Q* matrices (Gierl, Leighton, & Hunka, 2000; Leighton et al., 2004; Tatsuoka, 1995). To illustrate the use of different matrices in the AHM, a hypothetical attribute hierarchy is shown in Figure 1. Figure 1 indicates that attribute A1 is the prerequisite of attributes A2, A3, and A4, and A3 is the prerequisite of A4.

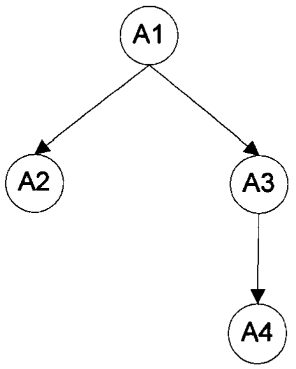


*Figure 1.* A hypothetical attribute hierarchy with four attributes.

In the AHM, the adjacency matrix (*A*), of order (*k, k*), where *k* is the number of attributes, is used to represent the direct relationships among attributes. In the adjacency matrix, the diagonal elements are designated as 0s. For the off-diagonal elements, a 1 in the position (*j, k*) ( $j \neq k$ ) indicates that attribute *j* is directly connected in the form of a prerequisite to attribute *k*, while a 0 in the position (*j, k*) ( $j \neq k$ ) indicates that attribute *j* is not the direct prerequisite to attribute *k*. The adjacency matrix for the hypothetical example shown above is

$$\begin{bmatrix} 0110 \\ 0000 \\ 0001 \\ 0000 \end{bmatrix}. \qquad \text{(Matrix 1)}$$

Row 1 of the matrix indicates that A1 is a direct prerequisite to A2 and A3 (i.e., $a_{12}$=1; $a_{13}$=1) and not the direct prerequisite of A4 (i.e., $a_{14}$=0). A2, on the other hand, is the direct prerequisite of no other attributes because all elements are 0 in row 2 of the matrix.

In the AHM, the reachability matrix (*R*), of order (*k, k*), is used to specify both the direct and indirect relationships among attributes. The *R* matrix can be calculated using $R = (A + I)^n$, where $n = 1, 2, \ldots, k$, is the integer required for *R* to reach invariance, *A* is the adjacency matrix, and *I* is the identity matrix. During calculation, any element that is greater than 1 is replaced by 1. For the hypothetical example, $A + I$ is equal to

$$\begin{bmatrix} 1110 \\ 0100 \\ 0011 \\ 0001 \end{bmatrix}. \qquad \text{(Matrix 2)}$$

The square and cube of $A + I$, after elements of greater than 1 are replaced by 1, are equal to each other. In other words, the resultant matrix has reached invariance with n = 3 and therefore, $(A+I)^3$ is the *R* matrix for the hypothetical example, as shown below:

$$\begin{bmatrix}1111\\0100\\0011\\0001\end{bmatrix}. \qquad \text{(Matrix 3)}$$

The *j*th row of the *R* matrix specifies all the attributes, including the *j*th attribute, that the *j*th attribute can reach through direct or indirect connections. In the sample reachability matrix, row 1 indicates that A1 can reach itself and all other attributes through direct or indirect relations because all elements on row 1 are 1s; rows 2 and 4 indicate that A2 and A4 can only reach themselves (i.e., only $r_{22} = 1$ and $r_{44} = 1$ in the corresponding row); row 3 indicates A3 can reach itself and A4 (i.e., $r_{33} = 1$ and $r_{34} = 1$).

The incidence matrix (*Q*) in the AHM is of order ($k$, $2^k$-$1$). It is composed of all combinations of attributes and represents the set of potential items ($2^k$-$1$) when the attributes are independent. Each column of the *Q* matrix indicates the attributes that are required to solve each item correctly. For the hypothetical example, the *Q* matrix is

$$\begin{bmatrix}101010101010101\\011001100110011\\000111100001111\\000000011111111\end{bmatrix}. \qquad \text{(Matrix 4)}$$

However, in the AHM, the attributes are related hierarchically in that some attributes are prerequisites of other attributes. Consequently, an item that probes an attribute must at the same time probe its prerequisite. For the hypothetical example, item 2 (column 2 of the Q matrix) (0100) indicates that item 2 probes attribute 2. However, the hierarchy indicates that attribute 2 requires attribute 1 as a prerequisite.

Thus, the item must be represented by (1100). But this representation would make item 2 identical to item 3 (column 3 of the Q matrix) with respect to the attributes probed. For this reason, item 2 can be removed. The removal of items in this manner produces a reduced Q matrix ($Q_r$). The $Q_r$ matrix indicates the total number of items that is needed given the restrictions of the attribute hierarchy on the relationships among attributes. For the hypothetical example here, the $Q_r$ is

$$\begin{bmatrix} 111111 \\ 010101 \\ 001111 \\ 000011 \end{bmatrix}. \qquad \text{(Matrix 5)}$$

The $Q_r$ represents the attribute blueprint or cognitive specifications for test construction. Thus, it plays a key role in guiding test development. For the hypothetical example here, the $Q_r$ matrix is of order (4, 6), indicating that, to measure the four attributes in the hypothesized hierarchy, six items must be developed. Column 1 of the $Q_r$ matrix indicates that an item must be created to measure attribute 1. Column 2 indicates that an item must be created to measure attributes 1 and 2. The remaining columns are interpreted in the same manner.

### *Generation of Attribute Patterns and Expected Response Patterns*

Given a hierarchy of attributes, the attribute patterns and expected response patterns can be generated. Attribute pattern refers to the combination of attributes that is consistent with the attribute hierarchy. Attribute patterns are equivalent to the columns in the $Q_r$ matrix except that one additional column should be included. This column contains all 0s for each attribute and indicates that an examinee has mastered none of the attributes specified in the hierarchy. Corresponding to each of the attribute patterns is an expected response pattern. Expected response pattern refers to the response pattern produced by an expected examinee who correctly answers items which require the cognitive attributes that the examinee has mastered, but fails the items which require the cognitive attributes that the examinee has not mastered. Expected response pattern establishes the correspondence between an examinee's test performance and the examinee's attribute pattern, and provides a convenient way for diagnosing the examinee's strengths and weaknesses. An examinee who is classified into an expected response pattern is said to have mastered the cognitive attributes implied by the corresponding attribute pattern, but not others. The following are the matrix of attribute patterns (the matrix on the left) and their corresponding matrix of expected response patterns (the matrix on the right) for the hypothetical example:

$$\begin{bmatrix} 0000 \\ 1000 \\ 1100 \\ 1010 \\ 1110 \\ 1011 \\ 1111 \end{bmatrix}, \begin{bmatrix} 000000 \\ 100000 \\ 110000 \\ 101000 \\ 111100 \\ 101010 \\ 111111 \end{bmatrix}. \qquad \text{(Matrix 6)}$$

As is shown, for an examinee who has an attribute pattern of (1010) (i.e., the examinee has mastered only A1 and A3 as shown in the fourth row of the left matrix above) in the hypothetical example, he/she should correctly answer items 1 and 3 but no other items; consequently the expected response pattern for this examinee is (101000). Conversely, if an examinee has an expected response pattern (101000), then it can be inferred that the examinee has mastered A1 and A3 but still needs work on A2 and A4.

*Classification of Observed Response Patterns*

In a real test, discrepancies will occur between the observed response patterns and expected response patterns because of *slips*. For example, the students may have the required attributes for an item, but due to a mistake in writing the answer sheet, they may get the item wrong. Conversely, some students may not have the required attributes, but by guessing or by applying partial knowledge, they could get the item correct. To achieve the purpose of being diagnostic, an examinee's observed response pattern needs to be classified by matching it against expected response patterns in the presence of these slips.

In the AHM, three classification methods have been developed to date (Cui, Leighton, Gierl, & Hunka, 2006; Gierl, Cui et al., 2006; Leighton et al., 2004). Among the three methods, two are IRT-based procedures (Method A and Method B) while the third is the artificial neural network approach (Method NN). In Method A, an observed response pattern is compared against all expected response patterns and slips (inconsistencies between an observed response pattern and an expected response pattern) of the form 0→1 and 1→0 are identified. The product of the probabilities of each slip is calculated to give the likelihood of the observed response pattern being generated from an expected response pattern for a given ability level (Leighton et al., 2004). Formally, this likelihood is expressed as

$$P_{ijExpected}(\theta_j) = \prod_{k \in S_{io}} P_{jk}(\theta_j) \prod_{m \in S_{i1}} [1 - P_{jm}(\theta_j)], \qquad \text{(Equation 1)}$$

where $S_{i0}$ is the subset of items with slips from 0 to 1 for the observed response vector of examinee, $S_{i1}$ is the subset of items with slips from 1 to 0, and $\theta_j$ is the ability level for a given observed response pattern, which can be estimated using an IRT-model. The observed response pattern would then be classified as being generated from the expected response pattern for which the value of $P_{ijExpected}(\theta_j)$ is the largest. Then, diagnostic information could be inferred from the attribute pattern implied by the corresponding expected response pattern.

For illustration, Table 1 presents the information for classifying the observed response pattern (101100) from a hypothetical 6-item test constructed from the hypothetical hierarchy used previously (the ability levels of the response patterns and the values of likelihood are not estimated using an IRT model, but specified by the author just for illustration purpose). The observed response pattern has a different number of slips when compared with the expected response patterns. As shown, the expected response pattern (111100) corresponds to the largest likelihood. Therefore the observed response pattern (101100) should be classified as this expected response pattern. Since the expected response pattern (111100) corresponds to the attribute pattern (1110), diagnostic information can be provided to the students with this observed response pattern that they have mastered attributes 1, 2, and 3, but need more work on attribute 4.

Table 1

*An Illustration of Classification Method A: Classifying Observed Response Pattern (101100)*

| Examinee Attributes | Expected Response Pattern | Ability Level | No. of Slips | Likelihood |
|---|---|---|---|---|
| 0000 | 000000 | -2.239 | 3 | 0.0001 |
| 1000 | 100000 | -0.823 | 2 | 0.0023 |
| 1100 | 110000 | -0.130 | 3 | 0.0751 |
| 1010 | 101000 | -0.088 | 1 | 0.1123 |
| **1110** | **111100** | **0.529** | **1** | **0.2459** |
| 1011 | 101010 | 0.513 | 2 | 0.0019 |
| 1111 | 111111 | 1.659 | 3 | 0.0001 |

In Method B, all the expected response patterns that are logically contained within the observed response pattern are identified. The attributes implied by these expected response patterns are supposed to have been mastered by the examinee of the observed response pattern. For the expected response patterns that are not logically included in the observed response pattern, the likelihoods of the slips of the form 1→ 0 are computed as

$$P_{ijExpected}(\theta_j) = \prod_{m \in S_{i1}} [1 - P_{jm}(\theta_j)] \,. \qquad \text{(Equation 2)}$$

Based on the likelihood values, judgments can be made about the classification of the observed response pattern according to the criterion set by the researchers.

For illustration, Table 2 presents the information for classifying the observed response pattern (101100) using Method B. The asterisks in the last column of the table indicate that the corresponding expected response patterns are logically included in the observed response pattern. For the expected response patterns that are not logically included, the slips of the form 1→0 are identified as shown in the fourth column of the table. If the researcher sets 0.2 as the criterion, then the observed response pattern (101100) can be classified to the expected response pattern (111100), which corresponds to the attribute pattern (1110). The diagnostic decisions for this response pattern can be made in a similar way as in Method A.

Table 2

*An Illustration of Classification Method B: Classifying Observed Response Pattern (101100)*

| Examinee Attributes | Expected Response Pattern | Ability Level | No. of Slips | Likelihood |
|---|---|---|---|---|
| 0000 | 000000 | -2.239 | 0 | * |
| 1000 | 100000 | -0.823 | 0 | * |
| 1100 | 110000 | -0.130 | 1 | 0.0459 |
| 1010 | 101000 | -0.088 | 0 | * |
| **1110** | **111100** | **0.529** | **1** | **0.2459** |
| 1011 | 101010 | 0.513 | 1 | 0.0415 |
| 1111 | 111111 | 1.659 | 3 | 0.0001 |

Both Methods A and B involve the calculation of joint probabilities for slips to get the values of maximum likelihood. In many cases, such calculations result in very small maximum likelihood values, which make the interpretation of these values difficult. For example, the maximum likelihood value may be 0.01 or even lower, which means, probabilistically, that the expected response pattern associated with the maximum likelihood value is very unlikely. However, according to the classification principle of the two methods, the expected response pattern is most likely because its likelihood is highest among all the expected response patterns. Another weakness common to both methods is that they rely on IRT-models and assumptions about the distribution of examinees, which are restrictive for the application of the two methods. To address these problems, Method NN was proposed (Gierl, Cui et al., 2006).

Method NN does not rely on IRT-models and it does not require stringent assumptions about the distribution of examinees. Moreover, rather than calculating the joint probabilities of slips to find the most likely attribute pattern, Method NN directly calculates the probabilities of individual attributes. In Method NN, the expected response patterns, called exemplars in the terminology of neural network, serve as the input to the neural network, and their associated examinee attribute patterns serve as the desired output from the neural network. The relationship between the expected response patterns and their associated attribute patterns is established by presenting each pattern to the network repeatedly until it "learns" each association (Gierl, Cui et al., 2006). If the network learns the associations successfully, a set of weight matrices will be produced, which can then be used to obtain the probabilities of the individual attributes for any observed response pattern with a tolerably small error term (e.g., RMS < 0.001). Then judgments can be made about whether an examinee has mastered a certain attribute or not according to the criterion set by the researcher.

Table 3

*An Illustration of Classification Method NN: Classifying Observed Response Pattern (101100)*

| Observed Response | Attribute Probability | | | |
|---|---|---|---|---|
| Pattern | 1 | 2 | 3 | 4 |
| 101100 | 0.99 | 0.97 | 0.89 | 0.07 |

Table 3 presents the classification information produced by Method NN for the observed response pattern (101100). Based on the information, we can conclude that examinees with this observed response pattern most likely possess attributes 1, 2, and 3, as the probabilities for these three attributes are high, but do not possess attribute 4, as the probability for this attribute is low.

*Evaluation of Attribute Hierarchies*

Gierl, Cui et al. (2006) developed two approaches for evaluating the cognitive attribute hierarchies: the attribute reliability and the hierarchical consistency index (*HCI*).

Attribute reliability refers to the consistency of the decisions made in a diagnostic test about examinees' mastery of specific attributes. The reliability of an attribute can be estimated by calculating the ratio of true score variance to observed score variance on the items that measure each attribute. In the AHM, items are usually designed to measure a combination of attributes. For these items, each attribute only contributes to part of the total item-level variance. In order to isolate the contribution of each attribute to an examinee's item-level performance, the item score is weighted by the subtraction of two conditional probabilities. The first conditional probability is associated with attribute mastery (i.e., an examinee who possesses the attribute can answer the item correctly) and the second conditional probability is associated with attribute non-mastery (i.e., an examinee who does not possess the attribute can answer the item correctly). The weighted scores for items that measure the attribute are then used in the reliability calculation (for more technical details, see Gierl, Cui et al., 2006). The reliability of each attribute is calculated as a variation of Cronbach's $\alpha$

$$\alpha_i = \frac{k_i}{k_i - 1}\left[1 - \frac{\sum\limits_{j \in S_i} W_{ij}^2 \sigma_{X_j}^2}{\sigma^2_{\sum\limits_{j \in S_i} W_{ij} X_j}}\right], \qquad \text{(Equation 3)}$$

where $\alpha_i$ is the reliability of attribute $i$, $k_i$ is the number of items that are probing attribute $i$ in the $Q_r$ (i.e., the number of elements in $S_i$), $\sigma_{X_j}^2$ is the variance of the observed scores on item $j$, $\sum\limits_{j \in S_i} W_{ij} X_j$ is the weighted observed total score on the items that are measuring attribute $i$, and $\sigma^2_{\sum\limits_{j \in S_i} W_{ij} X_j}$ is the variance of the weighted observed total scores.

Cui, Leighton, Gierl et al. (2006) proposed the *HCI* for the AHM, which examines the degree to which the observed response patterns are consistent with the attribute hierarchy and the reduced Q matrix. The *HCI* for examinee $i$ is given by

$$HCI_i = 1 - \frac{2\sum\limits_{j=1}^{J} \sum\limits_{g \in S_j} X_{i_j}(1 - X_{i_g})}{N_c}, \qquad \text{(Equation 4)}$$

where $J$ is the total number of items, $X_{i_j}$ is examinee $i$'s score (1 or 0) to item $j$, $S_j$

includes items that require the subset of attributes of item $j$, and $N_{c_i}$ is the total number of comparisons for correct-answered items by examinee $i$.

When examinee $i$ correctly answers item $j$, $X_{i_j} = 1$, the examinee is expected to also answer item $g$ that belongs to $S_j$ correctly, $X_{i_g} = 1\ (g \in S_j)$. If $X_{i_g} = 0$, then $X_{i_j}(1 - X_{i_g}) = 1$ and it is a misfit between examinee $i$'s observed response pattern and the expected response patterns specified by the attribute hierarchy. Thus, the fraction on the right side of the formula divided by 2, $\dfrac{\sum_{j=1}^{J} \sum_{g \in S_j} X_{i_j}(1 - X_{i_g})}{N_c}$, represents the proportion of misfits among the total number of comparisons for a given response pattern. This observation is important in the interpretation of *HCI* values: Given an *HCI* value, $\dfrac{1 - HCI_i}{2}$ indicates the proportion of misfits for a given response pattern (Cui, personal communication).

The values of the *HCI* range from $-1$ to $+1$. When an observed response pattern fits an expected response pattern in the hierarchy perfectly, the *HCI* has a value of 1. Conversely, the *HCI* value is -1 when the response pattern maximally misfits the hierarchy. Therefore, *HCI* values close to -1 indicate inconsistency between the observed response patterns and the expected response patterns specified by the attribute hierarchy, suggesting that the attribute hierarchy needs improvement. In addition, the mean and standard deviation of the $HCI_i$, $i = 1, 2, 3, \ldots, n$, can be used as indicators of the overall model-data fit. A high mean and low standard deviation suggest the observed response patterns fit the AHM model well.

### *Previous Applications of the AHM*

The AHM is a recently proposed psychometric method. Its technical details are still being developed. To date, only a few studies have been conducted using the AHM to model different content domains using cognitive models of task performance (Gierl, Wang, & Zhou, 2006; Leighton et al., 2004; Wang, Gierl, & Leighton, 2006).

Leighton et al. (2004) used Johnson-Laird's mental model theory (Johnson-Laird & Bara, 1984) to identify the cognitive attributes involved in syllogistic reasoning. Seven cognitive attributes were identified: (A1) interpretation of quantifiers according to logical criteria, (A2) ability to create first unique representation of logical quantifiers, (A3) ability to infer final conclusion from one-model syllogism, (A4) ability to create second unique representation of logical quantifiers premised on first representation, (A5) ability to infer final conclusion from two-model syllogism, (A6) ability to create third unique representation of logical quantifiers premised on second representation, and (A7) ability to infer final conclusion from three-model syllogism. According to the cognitive complexity of and the logical relationships among the attributes, these seven attributes were organized into the attribute hierarchy shown in Figure 2.

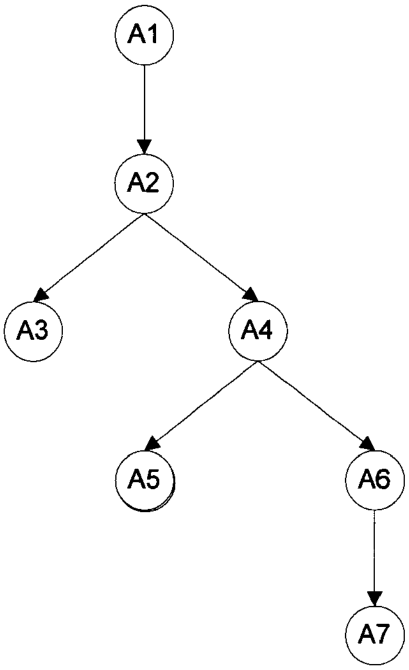


*Figure 2.* The attribute hierarchy of syllogistic reasoning derived from the mental model theory.

Based on this attribute hierarchy, 15 items were developed and 15 expected response patterns generated. Then the item parameters for the 15 items were estimated using the two-parameter logistic IRT model. In the study, the authors demonstrated the use of the two classification methods using two observed response patterns, one with no slips and the other with anomalous slips (e.g., correctly answers a more difficult item but fails an easier item). The authors showed that when an anomalous observed response pattern occurred, it was very unlikely that the response pattern would be classified as being generated from any of the expected response patterns. However, the authors did not evaluate the strengths and weaknesses of the two classification methods. Moreover, more recent technical developments — Method NN, attribute reliability, and the *HCI* — were not yet available.

Wang et al. (2006) conducted the first study that used the AHM with an operational test. Their study investigated the cognitive attributes underlying student performance on an English as a foreign language test. This study involved a three-step analysis: review of selected literature, expert rating, and psychometric analysis. (The substantive details of the study will be presented in the next section when studies investigating the reading hierarchies are reviewed. Only the results related to the AHM are reviewed here.) The mean *HCI* values were 0.3 and 0.5 for Hierarchy 1 and Hierarchy 2, respectively. These values indicated a moderate fit between the data and the two attribute hierarchies and thus supported the hierarchical relationships among cognitive attributes in second language reading.

As the first study that applied the AHM to an operational test, Wang et al. (2006) demonstrated the potential of the AHM in the construction of tests, test analysis, and evaluation of cognitive theories. First, the AHM is able to provide not only overall ability estimates for the examinees, as most current psychometric models do, but also their performance on specific cognitive attributes. Thus, diagnostic feedback can be prepared for the examinees for later performance improvement. Second, by applying the *HCI*, a better fitting cognitive model that more accurately reflects the cognitive attributes involved in problem solving can be found. This better model can be used to guide future test development. Third, the AHM can provide empirical evidence to the theories in a content domain. For example, in Wang et al.'s study, it was found that including the attributes of understanding difficult vocabulary and syntactic structure significantly improved the mean *HCI* values. This finding offers empirical support to Perfetti's (1985) theory that text-based processes are important in explaining reading ability.

Gierl, Wang, et al. (2006) evaluated the newest developments of the AHM, Method NN, attribute reliability, and the *HCI*, using the response data of 5000 randomly selected examinees on 21 SAT algebra items. Due to the lack of existing cognitive models for student performance on algebra tests, the authors used item review to extract the cognitive attributes measured by the items. Then, based on the extracted cognitive attributes, Gierl, Wang, et al. developed four different hierarchies which were used to analyze student response data. The authors found that Method NN was efficient in classifying the attributes mastered by the examinees, and that the *HCI* was efficient in evaluating model-data fit. However, they also found that it was difficult to examine the reliability for some of the attributes in the hierarchy, because, in some cases, only one item measured an attribute. The reason for such a situation is that the hierarchy was fit to existing test items rather than to items developed according to the attribute hierarchy. These results highlight the importance of developing an attribute hierarchy prior to test development and using it to develop test items.

The above studies investigated different aspects of the AHM. Wang et al. (2006) explored the integration of cognitive theory with psychometric analysis within the AHM framework while Gierl, Wang, et al. (2006) and Leighton et al. (2004) evaluated both the technical aspects of the AHM and its application. However, these three studies are far from complete in evaluating the AHM. Therefore, more studies on the technical developments and on the application of the AHM to operational tests are needed.

## The Hierarchical Relationships Among Cognitive Processes in Reading

The AHM is a cognitively-based psychometric approach which attempts to integrate the cognitive theories in a content domain with testing practice. The AHM models not only the cognitive attributes involved in a content domain, but also their interrelationships. As a result, a better understanding of the cognitive processes involved in reading and their interrelationships is necessary to apply the AHM in the analysis of the SAT Critical Reading subtest. In this section, the cognitive processes involved in reading are introduced by reviewing selected literature in both first language and second language reading. Literature in second language reading is reviewed because studies on English language learners can sometimes reveal important cognitive processes which are suggestive to first language reading. The interrelationships among these cognitive processes will be illustrated, and studies in which the hierarchical relationships among cognitive processes on tests of reading were investigated will be reviewed.

## *The Cognitive Processes in Reading*

Reading is an activity characterized by the translation of symbols, or letters, into words and sentences that have meaning for the reader. The ultimate goal of reading is comprehension, or being able to understand written material, to evaluate it, and to use it for one's needs (Chall & Stahl, 2004).

Reading is not a single-factor process (Alderson, 2000; Nassaji, 2003). It is a complex combination and integration of a variety of cognitive processes, ranging from basic linguistic processes to the integration of reader's background knowledge, inferences, and metacognitive processes. To be able to model students' performance on a reading test, it is necessary to understand these major components and how they are related to students' reading ability. These major components are described next.

### *Text-based Processing*

Text-based processing refers to the processing of text based solely on the textual information without the involvement of the reader's background knowledge. The readers' language knowledge and skills are mainly involved in text-based processing. According to the size of the informational units to be processed, text-based processing can be decomposed into different levels, such as word-level processes, sentence-level processes, and text-level processes.

*Word-level processes.* Word-level processes include word perception and word recognition. Word perception refers to the perceptual processes that transform the input stimulus (i.e., the strings of letters and symbols) into a format that is compatible with the format of the word concept in the lexicon (Haberlandt, 1988). Word recognition refers to the processes that activate the semantic properties of the word in the reader's long-term memory (Perfetti, 1985). Word perception and recognition are closely related to the reader's working memory capacity (e.g., Abu-Rabia, 1995; Swanson & Ashbaker, 2000). If these processes do not become automatic, then they may take up much working memory resources, and thus interfere with other processes, thereby causing reading difficulties (Perfetti, 1988). Among normally developing adult readers, word perception and recognition have generally become automatic (Perfetti, 1985; Stanovich, 2000). In other words, these processes only occupy minimal working memory resources in reading unless difficult vocabulary is encountered.

Vocabulary is an important factor for the execution of word-level processes and, in turn, for fluent reading. Texts with difficult vocabulary would make word perception and recognition a strenuous job. In such circumstances, gaps in the meaning of a text would occur and too many gaps would hinder the reader from constructing an overall text representation and a coherent text structure (Daneman, 1988; Perfetti, 1985) and, thus, cause difficulty in comprehension. Consequently, readers with larger vocabularies are generally at an advantage over those with smaller vocabularies (Alderson, 2000) because they can process the text faster and more accurately (Alderson, 2000; Perfetti, 1985). In other words, vocabulary can be regarded as an indicator, among others, of a reader's reading ability.

*Sentence-level processes.* Sentence-level processes refer to the processes of using one's knowledge of syntax to understand the informational structure of a sentence and 'translate' the surface form of a sentence into a coherent meaning representation (Haberlandt, 1988). Syntax is the component of a grammar that determines the way in which words are combined to form phrases and sentences (Radford, 2004). Syntactic knowledge plays an important part in sentence-level processes because the reader needs to apply their syntactic knowledge to segment the sentence into its syntactic constituents, such as phrases and clauses, and understand the referential relationships among the constituents in a sentence. Then, based on syntactic parsing results, they could assemble and integrate the propositions implied in the sentences and build meaningful representations of the sentences.

Among normally developing adult readers, processing sentences with simple syntactic structure has generally become automatic (Alderson, 2000). However, it would be conscious and effortful when complex sentences are encountered in reading (Smith, 2004). Therefore, ability to process syntactically complex sentences could differentiate high-ability readers from low-ability readers (Alderson, 2000; Perfetti, 1985; Vos, Gunter, Schriefers, & Friederici, 2001). Low-ability readers may have difficulty and make more errors in comprehending syntactically complex sentences (Nation & Snowling, 2000). They also tend to rely on limited syntactic strategies to comprehend complex sentences which high-ability readers are able to deal with flexibly and efficiently (Perfetti, 1985).

*Text-level processes.* Text-level processes refer to the processes that produce the unified meaning representation of larger sections of text (Haberlandt, 1988; Perfetti, 1985), such as multiple sentences, paragraphs, and the entire reading passage. To understand larger sections of text, knowledge of textual structure, such as cohesion, is indispensable (Alderson, 2000; Butcher & Kintsch, 2003; Smith, 2004). Cohesion refers to the connections between sentences, which are furnished by pronouns that have antecedents in previous sentences, adverbial connections, known information, and knowledge shared by the reader (Kolln, 1999; Thompson, 2004). Knowledge of cohesion is required when understanding new information in the text depends on understanding the already available information (Hudson, 1996). Without such knowledge, the reader would not be able to connect different parts of the text together and comprehension would be severely impaired.

Research has demonstrated that readers of different reading abilities show differential ability in processing the cohesive devices, and hence in comprehending the text (Ehrlich, Remond, & Tardieu, 1999). For example, low-ability readers had difficulties in the resolution of some anaphoric devices, such as the object pronouns with far antecedents, which impeded them from building a global representation of the text and thus affected their comprehension of the text (Ehrlich et al., 1999).

Apart from knowledge of textual structure, knowledge of genre (e.g., story, newspaper, and journal), text type (e.g., narrative, expository, and argumentative texts), and rhetoric feature (e.g., metaphor, irony, and personification), is also important to the comprehension of text (Alderson, 2000; Smith, 2004). Knowledge of genre and text type helps readers understand how texts are organized, what sort of information to expect in what place, how information is signaled, and how changes of content might be marked so that the reader can obtain a global representation of the text (Alderson, 2000; Smith, 2004). Knowledge of the rhetorical features of the text, together with the reader's global understanding of the text, helps the reader to understand more accurately the information the author is to convey. Although empirical studies evaluating the effects of these types of knowledge on readers' comprehension of text are rare, it has been shown that story grammars, or knowledge about narrative text, facilitate comprehension by allowing readers to quickly construct a model of the text (Alderson, 2000). It has also been shown that knowledge of expository text enables the students to comprehend scientific text better and thus improve their scientific literacy (Baram-Tsabari & Yarden, 2005).

*Application of Background Knowledge*

Background knowledge, or knowledge in the content domain a text is about, is considered essential in reading comprehension by many researchers (Alderson, 2000; Butcher & Kintsch, 2003; Hirsch, 2003; Kintsch, 1988, 1994; Rumelhart, 1994; Smith, 2004). Abundant studies have demonstrated the facilitative effect of background knowledge on comprehension (see Butcher & Kintsch, 2003, for a review). Moreover, it has been shown that background knowledge impacts comprehension at a deeper level than factors external to the reader such as text quality (Butcher & Kintsch, 2003). Conversely, researchers have also demonstrated that lack of background knowledge impedes comprehension. For example, Hirsch (2003) demonstrated that the lack of background knowledge, rather than language deficiency, caused the so-called 'fourth-grade slump,' the emergence of comprehension difficulties at the fourth grade level.

The application of background knowledge can occur at all levels of text-based processing. In word-level processing, proper background knowledge helps the reader activate context-relevant semantic representations of words (Perfetti, 1985). Similarly, in sentence- and text-level processing, proper background knowledge help the reader construct a model that facilitates the comprehension of the text (Perfetti, 1988; Smith, 2004).

Although background knowledge is important in reading comprehension and readers with more text-related background knowledge generally tend to achieve better comprehension of a certain text, such knowledge must be activated and applied properly for the comprehension of the text to be possible (Alderson, 2000; Perfetti, 1985, 1988). The ability to activate and apply the right knowledge can help differentiate high-ability readers from low-ability readers. Background knowledge can be used effectively by high-ability readers to facilitate their comprehension of text but does not affect the reading results of low-ability readers (Clapham, 1998; Perfetti, 1988; Phillips, 1987).

### *Inference*

Inference plays a crucial role in reading comprehension (Butcher & Kintsch, 2003). In communicating with the reader, the writer always assumes a certain amount of shared knowledge between the writer and reader. Thus, the total information necessary for a true understanding of a text is rarely stated explicitly. Much is left unsaid for the reader to fill in (Butcher & Kintsch, 2003), which makes inference necessary and important.

Like the application of background knowledge, inference can occur in word-, sentence-, and text-level processing. To obtain a true understanding of the text, different types of inferences may be needed. Bridging inference and elaborative inference are two common types of inference. Bridging inference is generally involved in text-based processing and is necessary for establishing coherence when cohesive devices, such as *but* or *however*, are lacking or when comprehending anaphoric pronouns (Butcher & Kintsch, 2003; Sanford, 1990; van den Broek, 1990). Rather than establishing coherence across different parts of the text, elaborative inference enriches the text through the addition of information from the reader's background knowledge, experience, or imagination (Butcher & Kintsch, 2003; Keenan, Potts, Golding, &

Jennings, 1990; Sanford, 1990; van den Broek, 1990). The two types of inferences together help the reader establish coherence and achieve an in-depth understanding of the text.

Research has demonstrated that ability to generate an inference is closely related to a reader's reading ability (e.g., Cain, Oakhill, & Bryant, 2004; Perfetti, 1985). Perfetti (1985) demonstrated that while both high-ability readers and low-ability readers may generate similar amounts of inferences during reading, high-ability readers can generate more context appropriate inferences. Researchers have also found that the types of inference may be associated with readers of different abilities. For example, Bowyer-Crane and Snowling (2005) found that while less skilled readers do not show much inferiority in making bridging inferences, they have particular difficulty generating knowledge-related elaborative inferences.

*Metacognitive Processes*

Metacognitive processes in reading refer to the processes of becoming aware of one's knowledge and skills and consciously using these knowledge and skills to plan and monitor one's reading process. Studies have shown that both high and low ability readers have and are able to use metacognitive processes (Perfetti, 1985, 1988). However, differences still exist in the application of these processes among readers of different abilities (Alderson, 2000). In contrast to high-ability readers, low-ability readers are often not aware of how or when to apply the knowledge and skills they have

(Alderson, 2000). They often fail to recognize an increase in the difficulty level of a text, to plan ahead, or to monitor the outcomes of their reading (Perfetti, 1985).

Block (1992) reviewed the relationship between metacognitive processes and reading ability. High-ability readers were found to have more control over the comprehension monitoring process than low ability readers. Moreover, high-ability readers were more aware of how they controlled their reading and more able to verbalize this awareness. Phakiti (2003) also found that high-ability readers tended to use more metacognitive processes and achieve better comprehension performance. However, there is no evidence yet as to whether the lack of ability to use metacognitive processes will impede comprehension and whether metacognitive processes are indispensable in the reading process.

### *Summary*

In the previous sections, the key component processes in reading and their relationship with reading ability were described. These processes include text-based processing, application of background knowledge, drawing of inferences, and use of metacognitive processes. As previously mentioned, these cognitive processes are not isolated, but interrelated. Thus, to model students' test performance and provide accurate cognitive feedback to the students, the relationships among these cognitive processes should also be well understood. Thus, in the next section, the relationships among the cognitive processes in reading are described and the empirical studies investigating these relationships are reviewed.

### *The Hierarchical Relationship Among Cognitive Processes in Reading*

In this subsection, it will be illustrated that the cognitive processes in reading are conceptualized as hierarchically-related by some researchers (e.g., Alderson, 2000; Lunzer, Waite, & Dolan, 1979; Urquhart & Weir, 1998). The hierarchical relationships among cognitive processes in reading make it convenient for modeling student performance on a reading test. For example, from the student performance on a set of test items measuring different cognitive processes in the hierarchy, inferences about the level where a student is reading can be made. As a result, studies on the hierarchical relationships among cognitive processes in a testing situation are also reviewed in this subsection.

In the literature on reading, different, but not mutually exclusive, principles have been proposed to model the hierarchical relationships among the cognitive processes. These principles include the cognitive demand involved, amount of inferences involved, and size of the information unit to be processed (Urquhart & Weir, 1998). Bloom's (1956) taxonomy, applied to reading, would be a good illustration for the hierarchical relationship ordered according to the first principle because the processes appearing later on the scale (e.g., evaluation, synthesis) are cognitively more demanding than those appear earlier (e.g., comprehension) (Alderson & Lukmani, 1989). The principle of the amount of inference involved can be perceived in van Dijk and Kintsch's (1983) model, where the three levels of text representation, the surface code, the textbase, and the situation model, are mainly distinguished by the amount of inference involved. The surface code is the exact wording and syntax of the text without any inference involved, the textbase includes a small number of inferences that are needed to establish local text coherence, and the situation model is constructed inferentially through interactions between the explicit text and background knowledge (Graesser et al., 1997). The demand of working memory resources underlies the principle of the size of the information unit to be processed in ordering cognitive processes into hierarchies. As working memory holds about two sentences (Graesser et al., 1997), the processing of larger sections of text would demand more working memory resources. As a result, sentence-level processes would be less demanding and can be regarded as lower level processes than text-level processes in which more working memory resource is needed.

It should be noted that the belief that the above principles can be used to order cognitive processes into hierarchies is mostly situated at the conceptual level. There have been few studies that explicitly investigated the tenability of these principles (Urquhart & Weir, 1998). Even fewer studies have been conducted to empirically investigate reading hierarchies in a testing situation. Among the few attempts to examine the hierarchical organization of cognitive processes, inconsistent results have been reported due to the different organizing principles of the hierarchies and the different methods used (Alderson, 1990; Alderson & Lukmani, 1989; Andrich &

Godfrey, 1978-1979; Hillocks & Ludlow, 1984; Ludlow & Hillocks, 1985; Lunzer et al., 1979; Wang et al., 2006).

For example, Lunzer et al. (1979) attempted to devise tests aimed at assessing hierarchically-organized cognitive processes in reading comprehension. These processes, which included “word meaning,” “words in context,” “literal comprehension,” “drawing inference from single sentences,” “drawing inference from multiple sentences,” “interpreting metaphors,” “finding main ideas,” and “forming judgments” (p. 44), were intended to be organized from lower- to higher-level processes. The authors used factor analysis with oblique transformation to analyze the test data. However, the authors failed to find evidence for the separability of such cognitive processes and were unable to demonstrate that these skills were hierarchically arranged.

Alderson (1990) and Alderson and Lukmani (1989) also attempted to examine whether the cognitive processes in text comprehension are distinguishable and whether they are hierarchically organized. The participants recruited in their studies were non-native English language learners. The cognitive processes used in their studies, organized from lower- to higher-level processes, include “recognition of words,” “identification,” “discrimination,” “analysis,” “interpretation,” “inference,” “synthesis,” and “evaluation” (Alderson & Lukmani, 1989, p. 260). In both studies, the authors asked a group of instructors experienced in teaching reading to rate what cognitive processes were measured by the items and whether these processes were lower-, middle-, or higher-level processes. However, the raters could not agree on the cognitive processes the items were measuring, nor on the levels of cognitive complexity for most items. For the items on which the raters had agreement, Alderson and Lukmani linked item parameters, such as item difficulty and discrimination, with the cognitive processes the items measure. They also found a slight but not marked tendency for higher-level cognitive processes to be more difficult and that items measuring lower-level cognitive processes were more discriminating than items measuring higher-level cognitive processes. Such findings support Perfetti's (1985) argument that text-based processes are functional in accounting for ability differences in reading. However, these findings were not verified by Alderson (1990), which led Alderson to conclude that no systematic relationship exists between item difficulty and level of text processing or between item discrimination and level of text processing. In other words, the hierarchical organization of the cognitive processes, though conceptually appealing, was not supported empirically in these two studies.

Contrary to the findings in the above studies, research completed by Andrich and Godfrey (1978-1979), Hillocks and Ludlow (Hillocks & Ludlow, 1984; Ludlow & Hillocks, 1985) and Wang et al. (2006) produced evidence for the hierarchical organization of the cognitive processes in reading. These studies assume an inherent relationship between item difficulty and the level of cognitive processes in reading. Specifically, they assume that items measuring higher level cognitive processes tend to be more difficult than items measuring lower level cognitive processes. Then, based on the difficulty levels of the items measuring specific cognitive processes, they could make claims about the hierarchical relationships among cognitive processes in reading.

Andrich and Godfrey (1978-1979) used the Rasch latent trait model to examine the responses of 188 native English speakers on a reading comprehension test. The participants were Grade 9 to first-year university students. By examining the item difficulties, the authors found that the items tended to cluster into different difficulty levels on a single dimension. For example, items measuring processes such as "remembering word meanings" and "understanding content stated explicitly" were found to be easier than processes such as "making inferences about the content" and "recognizing the author's tone, mood and purpose." As the reading passages were of similar difficulty, these items which display different difficulty levels led the authors to conclude that the cognitive processes measured by the items might be at different levels of a hierarchical structure.

Hillocks and Ludlow (1984) and Ludlow and Hillocks (1985), rather than simply examining the item difficulties, also examined the response patterns of individual students. The participants in their studies were native English speakers from Grade 9 to Grade 12 and some graduate students. The authors designed test items that measured two major levels of cognitive processes: literal questions (those whose answers appear directly in the text) and inferential questions (those whose answers are cued in the text but are not stated therein). The cognitive processes included the comprehension of "basic stated information," "key detail," "stated relationship," "simple implied relationship," "complex implied relationship," "author's generalization," and "structural generalization" (p. 16). Students who mastered the latter cognitive processes were supposed to have mastered the previous cognitive processes, but not vice versa. That is to say, students who correctly answered an item measuring a higher-level cognitive process should be able to correctly answer items measuring lower-level cognitive processes. Hillocks and Ludlow (1984) used the Rasch model to analyze the test data and Ludlow and Hillocks (1985) also used the Guttman model to compare the performance of the two models. The results from both studies indicated that, as predicted, students who correctly answered items measuring higher-level processes tended to correctly answer items measuring lower-level processes. At the same time, students who incorrectly answered items measuring lower-level processes tended to answer incorrectly items measuring higher-level processes. Such results led the authors to conclude that the cognitive processes measured by the test items were hierarchical and cumulative.

Wang et al. (2006) used the AHM to examine the hierarchical relationships among cognitive processes underlying student performance on an English as a foreign language reading test. The study was conducted on the item response data of 1,500 examinees. A three-step analysis was involved in the study: review of selected literature, expert rating, and psychometric analysis. The authors first reviewed the literature (e.g., Alderson, 2000; Urquhart & Weir, 1998; VanderVeen, 2004) in second language reading and reading assessment. Using the findings, the two attribute hierarchies displayed in Figure 3 were constructed.

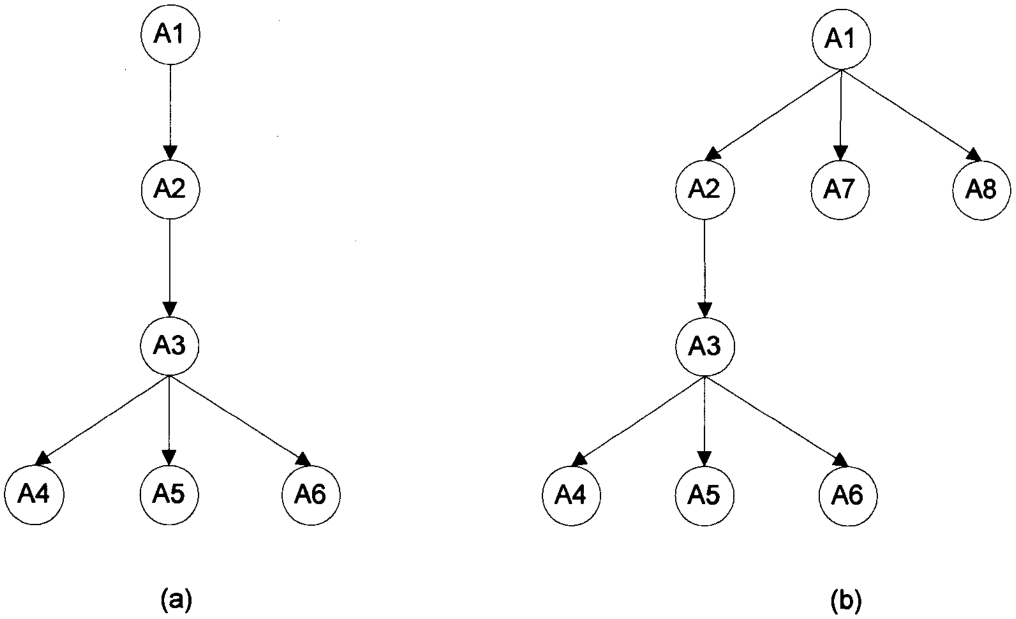


*Figure 3.* Two attribute hierarchies in second language reading in Wang et al. (2006).

Eight cognitive attributes were involved in the two hierarchies: (A1) basic language knowledge, such as word recognition and basic syntactic knowledge; (A2) understanding the content, form, and function of sentences; (A3) understanding the content, form, and function of larger sections of text; (A4) analyzing authors' purposes, goals, and strategies; (A5) determining word meaning in context; (A6) making inferences based on background knowledge; (A7) understanding text with difficult vocabulary; and (A8) understanding text with complex syntactic structure. Then, expert raters who were familiar with the target examinee population, were recruited to rate the above cognitive attributes measured by the test items. The items whose cognitive attributes were consistently rated by the experts were then submitted to the AHM analysis for each of the two hierarchies.

The results of the study indicated that a moderate hierarchical relationship existed among attributes. More specifically, attributes A4, A5, and A6 were more difficult than attribute A3, which, in turn, was more difficult than attributes A2 and A1. Moreover, inclusion of attributes A7 and A8 significantly improved the model-data fit, as indicated by the mean *HCI* values (0.3 for Hierarchy 1 and 0.5 for Hierarchy 2). Such a finding reinforced the role of text-based processes in explaining reading ability (Perfetti, 1985).

The divergent results from the studies reviewed have several implications. First, the difficulty level of the reading passages based on which test items are developed could be a factor in capturing the hierarchical relationships among cognitive processes in reading. For example, Wang et al. (2006) accounted for the effect of text difficulty by including two cognitive attributes, understanding text with difficult vocabulary and understanding text with complex syntactic structure in Hierarchy 2 and a stronger hierarchical relationship was detected. Hillocks and Ludlow (1984) and Ludlow and Hillocks (1985), who discovered the hierarchical relationships among cognitive processes in reading, also took the effect of text difficulty into consideration by including questions measuring the whole spectrum of skills in the hierarchy for each passage. However, the effect of text difficulty was not accounted for in the other studies which investigated the hierarchical relationship among cognitive processes. The effect of text difficulty is important to consider because reading involves the interaction between the reader and the text. If the text *per se* is difficult, then even text-based processes such as word recognition and the assembly and integration of propositions across sentences will become cognitively demanding. These processes will use up much of the available working memory resources and little will be left for the execution of other processes (Cain, Oakhill, & Lemmon, 2004; Long, Oppy, & Seely, 1997; Oakhill, Cain, & Bryant, 2003; Perfetti, 1985, 1988). Thus, in future studies, the difficulty level of the reading passages should be systematically manipulated in order to examine the hierarchical relationships among cognitive processes.

Second, if hierarchical relationships among cognitive processes are to be found, then appropriate statistical methods should be used to test the hierarchy. For example, Alderson and Lukmani (1989) and Alderson (1990) did not apply a rigorous statistical method due to the low reliability of the rating data. While Lunzer et al. (1979) used factor analysis with oblique transformation to analyze the test data, they did not discover any systematic relationship among cognitive processes. On the other hand, in the studies that did discover hierarchical relationships among cognitive processes (Andrich & Godfrey, 1978-1979; Hillocks & Ludlow, 1984; Ludlow & Hillocks, 1985;

Wang et al., 2006), the psychometric models, including the Rasch model, the Guttman model, and the AHM, were used. Two reasons could account for the failure of factor analysis in comparison to the Rasch and Guttman models and the AHM. One reason may be the moderate to high correlations among the cognitive processes in reading comprehension (e.g., Alderson, 2000). The correlations among the cognitive processes may cause the factor analysis method to fail to discover distinguishable factors, which may further affect its capability to discover hierarchical relationships among factors. If items measuring lower level cognitive processes have, on average, lower item difficulties than those measuring higher level cognitive processes, then the hierarchical relationship among cognitive processes is implied. However, factor analysis is mainly based on inter-item correlations, and is not sensitive to item difficulty. Thus, it is not an efficient method in comparison to the Guttman and Rasch models and the AHM to discover hierarchical relationships among hierarchically-ordered cognitive processes.

Third, the ways in which the test items are developed and the cognitive processes are determined may, to a certain degree, affect the results obtained from the studies. In Alderson and Lukmani (1989), Alderson (1990), and Andrich and Godfrey (1978-1979), existing test items were used and the cognitive processes measured by these items were determined post hoc by experts. At least two problems may arise in this situation. In Alderson and Lukmani (1989) and Alderson (1990), the authors could not get the raters to agree on the cognitive processes measured by most of the items. If the researchers are involved in determining the cognitive processes measured by the items, as in Andrich and Godfrey (1978-1979), then their judgment may be distorted by the statistical results they had, which would eventually erode the results of the study. On the other hand, the test items used by Lunzer et al. (1979), Hillocks and Ludlow (1984), and Ludlow and Hillocks (1985) were developed by the authors or by expert item writers, who were guided by the cognitive processes to be measured. This design feature gave the researchers control on what kind of items to develop and how to develop these items, thereby reducing the magnitude of unintended variance. Thus, if future studies are to be conducted to study the hierarchical relationships among cognitive processes, then items specifically designed for this purpose, rather than existing items, should be used. If this is not practical, reliable results about the cognitive processes measured by the items should be obtained before further steps are taken.

To summarize, studies on the hierarchical relationships among cognitive processes in reading were reviewed in this section. However, it must be noted that empirical studies are rare. This situation is unfortunate because studying the hierarchical relationships among cognitive processes in reading will help us better understand reading ability, an important ability people acquire and use throughout our life (Salinger, 2003). Once the hierarchical relationships among cognitive processes in reading are confirmed, test items measuring specific levels of cognitive processes could be designed. Then based on students' performance on these items, their reading problems could be diagnosed and the levels of their reading ability could be determined. This information could then be used for designing remediation techniques to help improve students' reading ability. Some statistical models, which are designed specifically for studying the hierarchical relationships among cognitive processes, such as the AHM, have rarely been applied to the reading domain. If more applications of these models are conducted on this topic, more light will be shed on the relationships among cognitive processes in reading, thereby leading to improved understanding of reading comprehension. The present study was an attempt to fill this gap.

# CHAPTER III: METHODOLOGY

The purpose of the current study was to investigate the cognitive attributes underlying student performance on the SAT Critical Reading subtest using the AHM. An emphasis is placed on the cognitive attributes included in the cognitive model of task performance proposed by VanderVeen et al. (2007) for the SAT critical reading items. Specifically, the study was designed to answer the following research questions:

1. What cognitive attributes in the VanderVeen et al. (2007) model are involved in answering the SAT critical reading items, what additional attributes are involved, and how are these cognitive attributes related to each other?
2. How can the AHM be used to analyze student response data, report student scores, and provide diagnostic feedback?

In the present chapter, the instrument used in the study and the detailed procedures for answering the research questions are described.

## Instrument

Student response data from the March 2005 administration of the SAT Critical Reading subtest were used. The SAT Critical Reading subtest is designed to measure students' critical reading ability or the ability to construct a coherent meaning representation of texts (passages, item stems, and response options). Critical reading ability is required for success in college and it involves text-based processing, knowledge-based processing, cognitive reasoning and problem solving, and metacognitive executive skills (Burton et al., 2003; VanderVeen, 2004).

The March 2005 administration of the SAT Critical Reading subtest contained two 25-minute sections and one 20-minute section. There are 67 items in two types: 19 sentence completion items and 48 passage-based reading items. All items are in the multiple-choice format with five options for each item. The sentence completion items have one or two blanks in the sentences for students to fill in by selecting the appropriate answer from the five options provided for each sentence. The passage-based reading items are based on eight reading passages of varying length, ranging from less than 100 words to approximately 700 words. The passages have different number of items, ranging from 2 to 13 items.

For the present study, data from passage-based reading items rather than sentence completion items were used. This decision was made because cognitive attributes that are not measured by sentence completion items, such as understanding larger sections of text, can be measured by the passage-based reading items.

## Procedures

The study was conducted in two stages. The first stage was a *cognitive* analysis to identify a final model for the SAT critical reading items. This stage involved the specification of attribute hierarchies, the identification of cognitive attributes used by students, and the validation of the attribute hierarchies using student verbal reports and the hierarchy consistency index (*HCI*). The second stage involved the *psychometric* analysis of the students' responses to the test items. This stage included the generation of the $Q_r$, attribute pattern, and expected response pattern matrices, calculation of attribute probabilities and reliabilities, and cognitive feedback.

## *Cognitive Analysis*

### *Specifying the Attribute Hierarchy*

The AHM starts with the specification of the attribute hierarchy using the cognitive attributes measured by the test items. In the present study, three attribute hierarchies for the SAT Critical Reading items were initially specified based on theories in the assessment of reading and research results related to the SAT Critical Reading subtest.

*Hierarchy 1.* Several studies have been conducted to identify the cognitive attributes measured by the SAT Critical Reading items (Burton et al., 2003; VanderVeen, 2004; VanderVeen et al., 2007). Based on a broad literature review and a review of the SAT critical reading items, VanderVeen (2004) proposed a preliminary cognitive model for the SAT Critical Reading that included seven cognitive attributes needed by examinees to successfully answer the items. The seven cognitive attributes included:

1. determining the meaning of words;
2. understanding the content, form and function of sentences;
3. understanding the situation implied by a text;
4. understanding the content, form, and function of larger sections of text;

5. analyzing authors' purposes, goals, and strategies;
6. use knowledge-based reasoning and problem solving; and
7. exercise metacognitive control over other processes by monitoring comprehension and processing capacity and selecting strategies such as task switching.

After multiple and iterative expert coding rounds, the first five attributes were retained in VanderVeen et al. (2007). A more complete description of the five attributes is provided in Table 4.

Table 4

*The Five Cognitive Attributes in VanderVeen et al.'s Cognitive Model of the SAT Critical Reading Subtest*

| **Skill Category** | **Description** | **Comment** |
|---|---|---|
| Determining the Meaning of Words | Student determines the meaning of words in context by recognizing known words and connecting them to prior vocabulary knowledge. Student uses a variety of skills to determine the meaning of unfamiliar words, including pronouncing words to trigger recognition, searching for related words with similar meanings, and analyzing prefixes, roots, and suffixes. | This skill category includes more than just lexical access, as word identification and lexical recall are combined with morphological analyses. |

| Skill Category | Description | Comment |
|---|---|---|
| Understanding the Content, Form, and Function of Sentences | Student builds upon an understanding of words and phrases to determine the meaning of a sentence. Student analyzes sentence structures and draws on an understanding of grammar rules to determine how the parts of speech in a sentence operate together to support the overall meaning. Student confirms that his or her understanding of a sentence makes sense in relationship to previous sentences, personal experience, and general knowledge of the world. | This skill category focuses on the syntactical, grammatical, and semantic case analyses that support elementary proposition encoding *and* integration of propositions across contiguous sentences. |
| Understanding the Situation Implied by a Text | Student develops a mental model (i.e., image, conception) of the people, things, setting, actions, ideas, and events in a text. Student draws on personal experience and world knowledge to infer cause-and-effect relationships between actions and events to fill in additional information needed to understand the situation implied by the text. | This skill category is a hybrid of the explicit text model and the elaborated situation model described by Kintsch (1998). As such, category three combines both lower level explicit text interpretation and higher level inferential processes that connect the explicit text to existing knowledge structures and schemata. |

| Skill Category | Description | Comment |
|---|---|---|
| Understanding the Content, Form, and Function of Larger Sections of Text | Student synthesizes the meaning of multiple sentences into an understanding of paragraphs or larger sections of texts. Student recognizes a text's organizational structure and uses that organization to guide his or her reading. Student can identify the main point of, summarize, characterize, or evaluate the meaning of larger sections of text. Student can identify underlying assumptions in a text, recognize implied consequences, and draw conclusions from a text. | This skill category focuses on the integration of local propositions into macro-level text structures (van Dijk & Kintsch, 1983) and more global themes. It also includes elaborative inferencing that supports interpretation and critical comprehension, such as identifying assumptions, causes, and consequence and drawing conclusions at the level of the situation model. |
| Analyzing Authors' Purposes, Goals, and Strategies | Student identifies an author's intended audience and purposes for writing. Student analyzes an author's choices regarding content, organization, style, and genre, evaluating how those choices support the author's purpose and are appropriate for the intended audience and situation. | This skill category includes contextual and pragmatic discourse analyses that support interpretation of texts in light of inferred authorial intentions and strategies. |

According to the three principles of ordering hierarchies reviewed in the last chapter, hierarchical relationships can be established among the five cognitive attributes in the VanderVeen et al. model. The first two cognitive attributes can be considered as lower-level in comparison to the latter three attributes. This interpretation is valid because the first two attributes deal with smaller information units and fewer inferences are involved in comparison to the latter three cognitive attributes. Moreover, neither of these two attributes involves attributes such as evaluation and synthesis, which are more cognitively demanding. To fit the hierarchical relationship among the five attributes in the AHM, two amendments were made. First, an attribute that measures the examinees' basic language knowledge, such as basic vocabulary and grammatical knowledge, was added. This attribute is assumed to be mastered by all examinees who take the SAT and it is the prerequisite for all other attributes. Second, the attributes of "understanding the situation implied by a text" and "understanding the content, form and function of larger sections of text" are essentially indistinguishable and thus were combined into one attribute. This consolidation occurred because, to build a situation model (i.e., a mental model of the people, things, setting, actions, ideas, and events in a text), a student has to have a global understanding of larger sections of text. Moreover, both attributes involve the integration of students' background knowledge with the textual information. Thus, even though these two attributes may be distinguishable conceptually, the distinction is difficult to make in practice, especially in the situation of a reading test. For example, VanderVeen et al. (2007) could not obtain a sufficient number of test items to measure the attribute "understanding the situation implied by a text" and thus had to remove the attribute from their analysis. As a result, these two attributes were combined into one attribute.

Based on VanderVeen et al. (2007) model and the above discussion, the initial version of Hierarchy 1 was specified, as shown in Figure 4a. The five cognitive attributes in Hierarchy 1 include: (A1) basic language knowledge, (A2) determining the meaning of words, (A3) understanding the content, form and function of sentences, (A4) understanding the content, form and functions of larger sections of text, and (A5) analyzing author's purposes, goals and strategies. The hierarchy reflects the interrelationships among cognitive attributes as discussed above: (1) A1 is the prerequisite for all the other attributes and is assumed to have been mastered by the target examinees of the SAT; (2) the execution of attributes A4 and A5 depends on the successful execution of attribute A3, which deals with smaller information units and involves less inferences.

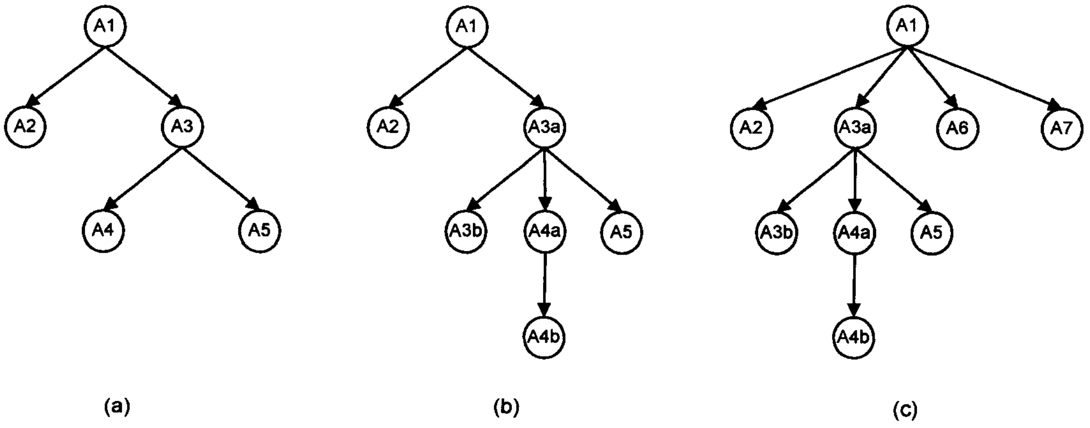


*Figure 4.* The three initial hierarchies specified for the present study.

*Hierarchy 2.* The cognitive attributes used to construct the initial version of

Hierarchy 1 were taken from the VanderVeen model with minimal revision. Although these cognitive attributes can be ordered hierarchically according to the three principles reviewed in Chapter 2, it must be noted that these attributes collapse across multiple cognitive processes and may affect the precision of cognitive feedback if used for diagnostic purpose. As a result, changes must be made to these attributes. First, the attribute "determine the meaning of words" could be regarded as a composite attribute consisting of two attributes with different cognitive demands. The first attribute involves the recognition and meaning retrieval of simple words, and the second attribute involves the determination of the meanings for more difficult words. The first attribute is generally automatic and could be regarded as part of basic language knowledge (i.e., attribute A1 in the current study) (Perfetti, 1985; Stanovich, 2000). Thus, it could be included in the prerequisite attribute discussed in Hierarchy 1. For the second attribute, contextual clues or language skills, (e.g., pronouncing words to trigger recognition, searching for related words with similar meanings, and analyzing prefixes, roots, and suffixes) must be used before the word meaning is determined, which would be cognitively more demanding than the first attribute. Thus it could be considered as a higher-level attribute than the first attribute. With the partitioning of the attribute of "determining the meaning of words," the denotation of A2 should be changed into "determining the meaning of words by referring to the context or by applying language skills." One ramification for this partitioning might be changing the structure of the hierarchy, as attribute A3 (understanding the content, form and function of sentences) might become a prerequisite of attribute A2 because of the involvement of context. However, at the current stage, the relationship between A2 and A3 will be retained until verbal report data are collected and analyzed.

Second, both A3 and A4 in Hierarchy 1 can be broken down into two attributes which involve different amount of inferences. For A3, the first attribute (A3a) involves literal understanding of sentences with minimal amount of inferences while the second attribute (A3b) requires the reader to use their experience and background knowledge to make inferences in order to build coherence at the sentence level. Similarly, the first attribute that can be extracted from A4 (A4a) does not involve much inference while the second attribute (A4b) requires the reader to use their experience and world knowledge to make inferences in order to build coherence across, summarize, and evaluate larger sections of text. Attributes A3a and A4a can be regarded as lower-level attributes than A3b and A4b, respectively. Moreover, A4b can be regarded as higher level than A3b because A4b requires the processing of larger information units than A3b. Based on such considerations, the initial version of Hierarchy 2 is specified in Figure 4b.

*Hierarchy 3.* As reviewed in Chapter 2, text difficulty could be a factor in capturing the hierarchical relationships among cognitive attributes in reading. However, in the VanderVeen et al. (2007) model, text factor was not accounted for. To address this omission, two additional attributes were included in Hierarchy 3 to account for the effect of the text difficulty. These two attributes are (A6) "understanding text with difficult vocabulary" and (A7) "understanding text with complex syntactic structure." These two attributes were included because vocabulary difficulty and syntactic complexity are two important indicators of text difficulty (e.g., Alderson, 2000; Smith, 2004; Veenman & Beishuizen, 2004; Vos et al., 2001). The inclusion of these two attributes also reflected Perfetti's (1985) argument that low-level linguistic processes play an essential role in reading. For texts with simple vocabulary and sentence structure, low-level linguistic processes could become automatic and occupy very little working memory space during reading. However, when difficult vocabulary or sentences are encountered, the reader would have to rely on morphological clues or contextual cues, which would sometimes become a more demanding task that requires more controlled processing (e.g., Perfetti, 1985). In other words, understanding texts with difficult vocabulary and sentences represents different cognitive attributes from understanding easy vocabulary and sentences in the reading hierarchy.

One note about the difference between attribute A2 and A6, which both involve the understanding of difficult vocabulary, is in place here. Attribute A6 is used as an indicator of text difficulty and it applies at the text level. On the other hand, attribute A2 refers to the understanding of specific difficult words. For example, in the following passage:

> In physics, a **laser** is a device that emits light through a specific mechanism for which the term **laser** is an acronym: **l**ight **a**mplification by **s**timulated **e**mission of

> radiation. This is a combined quantum-mechanical and thermodynamical process discussed in more detail below. As a light source, a laser can have various properties, depending on the purpose for which it is designed. A typical laser emits light in a narrow, low-divergence beam and with a well-defined wavelength. This is in contrast to a light source such as the incandescent light bulb, which emits into a large solid angle and over a wide spectrum of wavelength. These properties can be summarized in the term coherence.

If an item asks about the main idea of the passage, then attribute A6 is measured, because the vocabulary of the passage is, in general, difficult. However, if an item asks about the characteristics of the light emitted from a typical laser, then attribute A2 might be involved because the word "low-divergence" might need to be understood if the correct answer option uses a synonym of the word.

The initial version of Hierarchy 3 is as shown in Figure 4c. A summary of the cognitive attributes involved in the three hierarchies and their definitions are presented in Table 5. The definition of A2 in Hierarchies 2 and 3, rather than in Hierarchy 1, is presented.

Table 5

*Summary of Cognitive Attributes in the Initial Versions of Hierarchies 2 and 3*

| | |
|---|---|
| A1 | Basic language knowledge, such as word recognition and basic grammar. |
| A2 | Determining word meaning by referring to context or by applying language skills |
| A3 | Understanding the content, form and function of sentences |
| A3a | Literal understanding of sentences with minimal amount of inferences |
| A3b | Understanding sentences by making inferences based on the reader's experience and background knowledge. |
| A4 | Understanding the content, form and function of larger sections of text |
| A4a | Literal understanding of larger sections of text with minimal amount of inferences |
| A4b | Understanding larger sections of text by making inferences based on the reader's experience and world knowledge; building coherence across, summarizing, and evaluating larger sections of text. |
| A5 | Analyzing author's purposes, goals and strategies. |
| A6 | Understanding text with difficult vocabulary |
| A7 | Understanding text with complex syntactic structure |

*Revision of the hierarchies.* While the three attribute hierarchies above were developed by referring to the relevant reading theories and the previous research results related to the SAT Critical Reading subtest, it was not yet known how accurately the hierarchies reflected the way examinees used their cognitive attributes on the present administration of the test. Therefore, allowances were made to make revisions to the three initial hierarchies to account for any inconsistencies between these hierarchies and the examinee verbal report analysis.

### *Verbal Report Study*

For the AHM to work properly, the attribute hierarchy and the cognitive attributes within it must be a valid reflection of the cognitive processes and skills used by students. After attribute hierarchies are specified, the cognitive attributes measured by the test items and their interrelationships should be identified and validated. For this purpose, verbal report study was completed as part of another study (Wang & Gierl, 2007). A group of eight first year university students who were directly out of high school were recruited for this study. The interviews were conducted at the beginning of the university year. It was felt that beginning university students interviewed at that time would be most like high school students who take the SAT toward the end of Grade 12.

To avoid the negative effects caused by fatigue, a shorter test was created from the SAT Critical Reading subtest for the verbal report study. There were two considerations in creating the shorter test. First, the passages should represent the three types of passages used on the SAT Critical Reading subtest, namely, short passages, long passages, and parallel passages. Second, the difficulty levels of the items in the shorter test should represent those of the passage-based reading items on the SAT Critical Reading subtest, which range from 1, indicating easy items, to 5, indicating hard items. Based on these two considerations, the shorter test created for the verbal report study contained three passages and 20 items. The three passages included one short passage, one long passage, and one parallel passage. The 20 items were at different difficulty levels. A comparison of the difficulty levels of these 20 items with those of the items on the full subtest can be found in Table 6.

Table 6

*A Comparison of the Difficulty Levels of the 20 items with Those of the Items on the Whole Subtest*

| Difficulty levels | Shorter Test (20 items) | All Passage-based Reading Items (48 items) |
|---|---|---|
| 1 | 1 | 3 |
| 2 | 5 | 12 |
| 3 | 9 | 25 |
| 4 | 4 | 4 |
| 5 | 1 | 4 |

The verbal report study was conducted in September 2006. Before conducting the verbal report study, information sheets about the study were sent to a group of 200 first-year undergraduate students at the University of Alberta. Students who responded the information sheet were contacted. Using this process, eight first-year undergraduate students (five females and three males) who were directly out of high school were recruited for the verbal report study. The students were from a variety of academic backgrounds: Two of the students were majoring in Education, two in Arts, and one each in Psychology, Sociology, Anthropology, and Physics. As only academically competitive students were admitted to the university, the reading proficiencies of these students might be higher than the average of the SAT test-takers, which includes all college-bound students.

To ensure that complete information about the cognitive attributes used to answer the test items was obtained, both concurrent and retrospective verbal reports were collected from the individual participants as they worked through the reading items (Ericsson & Simon, 1993; Leighton, 2004). During the data collection, the participants were asked to read the passages aloud, verbally express their thought processes while responding to the items, and upon completing each passage, retrospectively describe aloud what they remembered about the thought processes they used to answer the items. The participants' responses were recorded using high-quality audio recorders. Considering that participants might not be familiar with the concurrent and retrospective verbal reports, they were provided with an opportunity to practice their verbal reporting skills, using a sample passage with two items. Once the participants were accustomed to the verbal report procedures and had no more questions, they were administered the three passages together with the passage items. If the participants remained silent for longer than 10 seconds, they would be prompted by the researcher to keep talking. After the verbal reports were collected, they were transcribed verbatim and typed into the computer.

Two graduate students (including the researcher) in educational psychology, who had experience in conducting verbal report study and were familiar with reading comprehension and the assessment of reading, coded the cognitive attributes in the three hierarchies specified in Table 5. The coders were also instructed to identify any additional cognitive attributes that occurred in the verbal report data but were not included in the three hierarchies. Additional cognitive attributes that occurred frequently and consistently were used to revise the three attribute hierarchies specified in Figure 4. Disagreements during coding were resolved through discussion.

During the coding process, the two coders first met to discuss the ratings for the cognitive attributes involved in the three hierarchies. After achieving agreement on the meanings of the cognitive attributes, the coders rated two items from the verbal reports of two students collaboratively in order to get a sense of how the coding should be conducted. Then the coders independently evaluated the remaining verbal report data. Because different cognitive attributes were involved in the attribute hierarchies, two rounds of coding were conducted. The first round of coding was conducted using the cognitive attributes in Hierarchy 1, and the second round using the attributes in Hierarchies 2 and 3. After the coding was completed, the two coders met, once again, to discuss and resolve any disagreements that occurred during coding.

### *The HCI Analysis*

After the cognitive attributes and the final attribute hierarchies were determined, the *HCI* (Cui et al., 2006) was used to evaluate whether the attribute hierarchies accurately reflected the cognitive attributes employed by the examinees on the SAT Critical Reading subtest. The response data from two random samples of 2000 examinees who took the 2005 administration of the SAT were used for the calculation of the *HCI* results. The selection of two random samples allowed evaluation of the performance of the *HCI* across samples for the same attribute hierarchy.

The *HCI* for examinee *i* was provided in Equation 4 (see Chapter 2, p.23). The mean and standard deviation of the *HCI*$_i$ across all the students was used as the indicator of the overall model-data fit. A high mean and low standard deviation suggest that student response data fit the attribute hierarchies well. For the present study, the Mathematica syntax developed by Cui, Leighton, Gierl et al. (2006) was used to calculate the mean and standard deviations of the *HCI* values.

After the *HCI* values for the three hierarchies were calculated, they were compared and the best hierarchy was selected for the next stage—analysis of student response data. Two criteria were used in selecting the hierarchy: model-data fit and parsimony. If two hierarchies indicated different degrees of model-data fit, then the one with better model-data fit was selected. However, if two hierarchies fitted the data equally well, then the more parsimonious one was used because a more parsimonious model is easier for interpretation and for providing diagnostic feedback to students. Given that there is no established guidelines or statistical tests for evaluating model-data fit using the *HCI* at this stage, the evaluation for the model-data fit of the attribute hierarchies has to be judgmental.

### *Psychometric Analysis*

#### *Generation of the $Q_r$, Attribute Pattern and Expected Response Pattern Matrices*

After the best-fitting hierarchy was identified, the attribute patterns and expected examinee response patterns were generated. Attribute patterns refer to the combinations of attributes that are consistent with the attribute hierarchy. They can be obtained by transposing the $Q_r$ matrix of the attribute hierarchies.

After the attribute patterns were generated, they were matched with the cognitive attributes required by the items to generate the expected examinee response patterns. An expected examinee is a hypothetical examinee who correctly answers items that require only specific cognitive attributes that the examinee has mastered. If the attribute pattern of an expected examinee includes the cognitive attributes required by the item, then this examinee is expected to answer this item correctly. Conversely, if at least one of the cognitive attributes required by the item is missing in an expected examinee's attribute pattern, then the examinee may not answer this item correctly.

*Estimating Attribute Probabilities*

To classify student response data into structured attribute patterns, the neural network classification method (Method NN) was used to calculate the probabilities that examinees possess specific attributes. Method NN was used because it does not rely on IRT-models, nor does it require stringent assumptions about the distribution of examinees. Moreover, the method produces easily interpretable results in comparison to the other two methods reviewed in Chapter 2.

To use the neural network, the relationship between the expected response patterns and their associated attribute patterns was established by presenting each pattern to the network repeatedly until it "learned" each association (Gierl, Cui et al., 2006). After the network learned the associations successfully, a set of weight matrices was produced. These weight matrices were then used to obtain the probabilities of the individual attributes for any observed response pattern. A probability close to 1 indicates that the corresponding attribute is likely to be mastered by the examinee. Conversely, a probability close to 0 indicates that the corresponding attribute is likely not to be mastered by the examinee.

*Estimation of Attribute Reliabilities and Standard Error of Measurement (SEM)*

Attribute reliability refers to the consistency of the decisions made in a cognitive diagnostic test about examinees' mastery of specific attributes. A high value of reliability is always preferred. In the present study, the reliabilities of the attributes were calculated using Equation 3 (see Chapter 2, p.23).

After the attribute reliabilities were estimated, the SEM was calculated for the attributes using the following formula

$$SEM = s_x\sqrt{1-r_{xx}}\,, \qquad \text{(Equation 5)}$$

where $s_x$ is the standard deviation of attribute probabilities in the context of this study and $r_{xx}$ is the reliability of a certain attribute. With the SEM, confidence intervals can be built for attribute probabilities in score reporting.

*Score Reporting and Providing Cognitive Feedback*

According to the attribute probabilities for the examinees and the reliabilities of each specific attribute, profiles were created for a sample of students to demonstrate the features of the AHM in score reporting and illustrate how cognitive diagnostic feedback can be presented to future examinees.

# CHAPTER IV: IDENTIFICATION OF THE FINAL MODEL

This chapter presents the results from the cognitive analysis and identifies the final hierarchy to be used for psychometric analysis. The chapter is divided into two sections. The first section contains the results from the verbal report study. Based on the verbal report results, the three attribute hierarchies specified in Chapter III were revised. In the second section, the *HCI* was calculated for each of the two random samples of students who took the March 2005 administration of the SAT. The hierarchy that provided the best model-data fit was selected and used for analyzing student response data.

## Results from the Verbal Report Study

In this section, the results from the verbal report study are presented. The characteristics of the items used for collecting student verbal reports are first presented. Then, the frequencies of the cognitive attributes that appeared in the student verbal reports are reported. Based on these frequencies, the cognitive attributes measured by the test items were determined. The relationships among the cognitive attributes in the three hierarchies are also described using the excerpts from student verbal reports. Finally, the three hierarchies were revised based on the results obtained from the verbal report study.

### *Item Characteristics*

An analysis was done to evaluate the item characteristics of the 20-item test created for the verbal report study, and to compare these results to a larger sample of students (*n* = 5000) who wrote the same items in March 2005. Using data from the eight students, the mean performance on the 20-item reading test was 15.25 (SD = 4.33); the mean item difficulty value was 0.76 (SD = 0.18); and the mean item discrimination value was 0.64 (SD = 0.50). The mean performance of the 5000 students who wrote the March 2005 administration of the SAT was 12.41 (SD = 4.19), the mean item difficulty values was 0.62 (SD = 0.17); and the mean item discrimination was 0.56 (SD = 0.06). In other words, the 20 items were slightly easier and more discriminating for the eight verbal report students than for the 5000 sample. A summary of the analysis can be found in Table 7. As expected, the sample of eight students tended to perform at a higher level than the sample of students who wrote the SAT in March 2005.

Table 7

*Psychometric Characteristics for the Verbal Report and March 2005 Samples on the 20 Critical Reading Items*

| Sample | Verbal Report | March 2005 |
|---|---|---|
| No. of Examinees | 8 | 5000 |
| No. of Items | 20 | 20 |
| Mean | 15.25 | 12.41 |
| SD | 4.33 | 4.19 |
| Mean Item Difficulty | 0.76 | 0.62 |
| SD Item Difficulty | 0.18 | 0.17 |
| Mean Item Discrimination[a] | 0.64 | 0.56 |
| SD Item Discrimination | 0.50 | 0.06 |

[a]Biserial correlation

*Frequencies of the Cognitive Attributes Used by the Students*

The frequencies for the cognitive attributes A2 through A5 in the initial version of Hierarchy 1 are summarized in Table 8[1]. All four attributes were used, to differing degrees, by the students. Moreover, attributes A3 and A4 were more frequently used than attributes A2 and A5. Attribute A3 was used 88 times, A4 80 times, A5 48 times, and A2 21 times.

[1] As attribute A1 is specified as a prerequisite to all other attributes in the three hierarchies, it is assumed to be measured by all items and mastered by all students. Therefore, it was not coded or reported.

Table 8

*Frequencies of the Cognitive Attributes in the Initial Hierarchy 1*

| Item | A2 | A3 | A4 | A5 |
|---|---|---|---|---|
| 1 | 2 | 7 | 2 | 2 |
| 2 | 0 | 2 | 0 | 1 |
| 3 | 1 | 2 | 7 | 2 |
| 4 | 0 | 4 | 8 | 3 |
| 5 | 0 | 3 | 7 | 4 |
| 6 | 4 | 5 | 4 | 2 |
| 7 | 1 | 0 | 3 | 6 |
| 8 | 0 | 0 | 7 | 3 |
| 9 | 4 | 6 | 3 | 0 |
| 10 | 0 | 7 | 1 | 1 |
| 11 | 4 | 2 | 1 | 6 |
| 12 | 0 | 4 | 2 | 1 |
| 13 | 0 | 6 | 6 | 4 |
| 14 | 0 | 7 | 4 | 0 |
| 15 | 3 | 2 | 6 | 4 |
| 16 | 0 | 3 | 5 | 4 |
| 17 | 2 | 7 | 4 | 0 |
| 18 | 0 | 6 | 8 | 2 |
| 19 | 0 | 8 | 1 | 2 |
| 20 | 0 | 7 | 1 | 1 |
| Total | 21 | 88 | 80 | 48 |

The results for attributes A2 through A7 in the initial versions of Hierarchies 2 and 3 are summarized in Table 9. Similar to the outcomes in the initial version of Hierarchy 1, all eight attributes were used, to differing degrees, by the students. Attributes A3a and A4a were used more frequently than the other attributes. Attribute A3a was used 81 times and A4a 74 times. The frequencies for the attributes A2, A3b, A4b, A5, A6, and A7 are 21, 52, 26, 48, 29, 14 times, respectively.

Table 9

*Frequencies of the Cognitive Attributes in Hierarchies 2 and 3*

| Item | A2 | A3a | A3b | A4a | A4b | A5 | A6 | A7 |
|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 5 | 2 | 2 | 1 | 2 | 1 | 0 |
| 2 | 0 | 1 | 2 | 0 | 0 | 1 | 0 | 0 |
| 3 | 1 | 2 | 0 | 4 | 3 | 2 | 1 | 0 |
| 4 | 0 | 4 | 0 | 6 | 3 | 3 | 0 | 0 |
| 5 | 0 | 3 | 3 | 7 | 2 | 4 | 0 | 0 |
| 6 | 4 | 5 | 2 | 4 | 1 | 2 | 5 | 0 |
| 7 | 1 | 0 | 0 | 3 | 0 | 6 | 2 | 0 |
| 8 | 0 | 0 | 0 | 6 | 2 | 3 | 0 | 0 |
| 9 | 4 | 5 | 3 | 3 | 0 | 0 | 1 | 0 |
| 10 | 0 | 6 | 5 | 1 | 1 | 1 | 0 | 0 |
| 11 | 4 | 2 | 6 | 1 | 0 | 6 | 0 | 0 |
| 12 | 0 | 4 | 4 | 2 | 0 | 1 | 4 | 3 |
| 13 | 0 | 6 | 0 | 6 | 4 | 4 | 5 | 4 |
| 14 | 0 | 7 | 6 | 4 | 0 | 0 | 0 | 0 |
| 15 | 3 | 2 | 5 | 6 | 2 | 4 | 4 | 7 |
| 16 | 0 | 3 | 0 | 5 | 5 | 4 | 0 | 0 |
| 17 | 2 | 6 | 4 | 4 | 2 | 0 | 6 | 0 |
| 18 | 0 | 6 | 2 | 8 | 4 | 2 | 0 | 0 |
| 19 | 0 | 7 | 2 | 1 | 5 | 2 | 0 | 0 |
| 20 | 0 | 7 | 6 | 1 | 0 | 1 | 0 | 0 |
| Total | 21 | 81 | 52 | 74 | 26 | 48 | 29 | 14 |

Apart from the cognitive attributes specified in the three initial hierarchies, two additional attributes were found. These two additional attributes are (A8) "using rhetorical knowledge," and (A9) "evaluating response options." Attribute A8 was defined as "use rhetorical knowledge to identify the rhetoric devices used by the author, such as imagery, metaphor, and parallelism, and/or to improve the reader's comprehension of the text." The verbal report of student G.A. when answering item 2 provides an example of attribute A8:

> G.A.: And for that, if you look back line8 to 10, it's talking about, like, actually going out, and doing things to get the final product, like getting fish by going into the boat and catching it and having a chicken by raising it. So, in the end all you have is the fish or the chicken, there is no indication of how you got it. So to me it is a metaphor of how he got his music, like final product of his music.

In the above verbal report, the student used her rhetorical knowledge to identify that the author was using metaphor in the text. Another example of attribute A8 was reflected in the verbal report of student D.C. when answering item 15, where identifying the rhetorical device improved the student's comprehension of the text:

> D.C.: Yeah, I am thinking of trying to grasp the concept of what he is trying to say. Disparage the present-day treatment of the arts. Uh...this is a concept the author is talking about throughout this. Could be this one because he uses "conventional" several times. Maybe he is repeating to emphasize this concept.

In this verbal report, the student identified that repetition was used by the author. Using her rhetorical knowledge that repetition can be used to emphasize, she could further recognize that the repetition was used to emphasize the concept of "disparaging the present-day treatment of the arts," a central idea of the paragraph.

Attribute A9 was defined as "select the one response option that best fits the requirements of the question and the idea structure of the text through evaluating the alternative response options and eliminating the option(s) that appear unreasonable based on paragraph or overall passage meaning." It must be noted that attribute A9 differs from the other attributes in that it is an attribute specific to multiple-choice test items only. Thus, it is more of a test-taking attribute than a pure reading attribute. An example of attribute A9 can be found in the verbal report by student G. A. when answering item 7:

> G.A.: Ok, so it's totally not E. And it's not A either. Um…it could be C but that depends on what the rest is …[talking about?]. It's only asking about the opening paragraph, so let's say it's not C. And D, is only saying that …they are remaking these old theater houses, he is not saying why are providing an explanation as to why any one want to do that. So it's not D, I am gonna go with B.

In the above report, the student evaluated the reasonableness of the five options and eliminated two least reasonable options. Then, she ruled out two other options according to her understanding of the text and arrived at the correct answer.

The frequencies of attribute A8 and A9 are shown in Table 10. Altogether, attribute A8 was used 34 times and attribute A9 71 times.

Table 10

*Frequencies of Attributes A8 and A9*

| Item | A8 | A9 |
|---|---|---|
| 1 | 0 | 2 |
| 2 | 8 | 1 |
| 3 | 0 | 2 |
| 4 | 0 | 0 |
| 5 | 0 | 3 |
| 6 | 0 | 6 |
| 7 | 4 | 5 |
| 8 | 0 | 2 |
| 9 | 1 | 2 |
| 10 | 0 | 1 |
| 11 | 3 | 7 |
| 12 | 0 | 5 |
| 13 | 1 | 7 |
| 14 | 5 | 1 |
| 15 | 7 | 8 |
| 16 | 0 | 7 |
| 17 | 1 | 5 |
| 18 | 0 | 2 |
| 19 | 0 | 1 |
| 20 | 4 | 4 |
| Total | 34 | 71 |

The total frequencies and percentages of the cognitive attributes used by the participants are summarized in Table 11 in descending order. The table indicates that attributes A3, A3a, A4, A4a, and A9 are more used attributes, exceeding 10% of the total attribute usage. Attributes A6, A4b, A2, and A7, on the other hand, are used less than 5% of the total attribute usage.

Table 11

*The Total Frequencies and Percentages of All Cognitive Attributes*

| Attribute | Total<br>Frequencies | % |
|---|---|---|
| A3[2] | 88 | 14.2 |
| A3a | 81 | 13.1 |
| A4[2] | 80 | 12.9 |
| A4a | 74 | 12.0 |
| A9 | 71 | 11.5 |
| A3b | 52 | 8.4 |
| A5 | 48 | 7.8 |
| A8 | 34 | 5.5 |
| A6 | 29 | 4.7 |
| A4b | 26 | 4.2 |
| A2 | 21 | 3.4 |
| A7 | 14 | 2.3 |

### *Determining the Cognitive Attributes Measured by Test Items*

As the AHM uses only dichotomous values to conduct the psychometric analysis—meaning the attribute can only be deemed absent (i.e., a value of 0) or present (i.e., a value of 1)—the frequencies of the cognitive attributes had to be transformed into 0s and 1s. In determining whether a cognitive attribute was measured by a test item, two considerations were made. First, if a cognitive attribute is coded as 1, then its prerequisite would also be coded as 1. Second, as eight students were included in the verbal report study, if at least half of the participants used a cognitive attribute, then this

[2] As A3 is the composite attribute of A3a and A3b, and A4 of A4a and A4b, there are overlaps between A3, and A3a and A3b, and between A4, and A4a and A4b. They are all included in the table to give a sense of how frequently each attribute is used by the students.

cognitive attribute could be deemed prevalent enough to be regarded as present. Using these two assumptions, the recoded cognitive attributes measured by the test items are summarized in Table 12.

Table 12

*Cognitive Attributes Measured by the 20 Test Items*

| Item | A2 | A3 | A4 | A3a | A3b | A4a | A4b | A5 | A6 | A7 | A8 | A9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 3 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 6 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 7 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| 8 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 12 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 13 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 14 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 15 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 16 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
| 17 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 18 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 19 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 20 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

### *Relationships Among Cognitive Attributes*

In addition to evaluating the veracity of the cognitive attributes, their hierarchical relationships were also examined during the protocol analysis. For Hierarchies 2 and 3, attribute A3a was found to be the prerequisite of attributes A2, A3b, A4a, A5, and A8 because the participants must possess A3a before they could use other attributes to process the text. For similar reasons, attribute A4a was the prerequisite of A4b. As A3a and A4a were coded, respectively, as A3 and A4 in Hierarchy 1, A3 is considered the prerequisite of A2, A4, and A5. As no prerequisite relationships were found between attributes A6, A7, and A9 and any other attributes, they were put under attribute A1 to indicate that, with basic language knowledge, examinees are able to execute these attributes. The relationships between attribute A3a and attributes A2, A3b, A4a, A5, and A8, and the relationship between attribute A4a and A4b are illustrated below using excerpts from student verbal reports.

*A3a vs. A2.* Item 9 measures attributes A3a and A2. Five students correctly answered this item. Students are required to determine the meaning of the word "*death*" by referring to a wider context (A2). However, when solving this item, the students must understand the literal meaning of the sentences that are related to this word (A3a). These sentences include:

*(1) After 50 years of life and 20 years of death, the great Adler and Sullivan Auditorium in Chicago is back in business again.*
*(2) Closed after that, it settled into decay for the next 20 years.*

Understanding these two sentences will lead students to rule out the two plausible distractors "demolition" and "flagging attendance," and select the correct one "neglect." This is illustrated in the verbal reports provided by students D.C. and P.B.:

> D.C.: **They weren't demolishing, like it hasn't been demolished yet (A3a).** Uh, neglect, I chose neglect because it talked about **until 1967, and said, oh, here it is, closed after that, it settled into decay for the next 20 years (A3a). It was neglected (A2).** So it had to be that.

> P.B.: So death is really, what it's saying or the essence of **what it's talking about in the paragraph is a time when it was unoccupied or it was not being used, or ignored (A3a)**. So flagging attendance, no, because **it's closed (A3a)**. So I would say neglect because it's being rejected by the general public during the…or after the war. **So I would say neglect (A2).**

Both students understood the above two sentences and identified the correct meaning for the word "death."

On the contrary, failure to understand these two sentences led students to select a wrong option. Consider the verbal report provided by L.G.:

> L.G.: Based on the 1st sentence of the passage, death here contrasted "life" and "is back in business". I think it refers to the demolition, which is C.

Obviously, the student failed to understand or did not notice the second sentence listed above. Therefore, she chose "demolition" rather than "neglect."

*A3a vs. A3b.* Item 14 measures attributes A3a and A3b. All eight students correctly answered this item. To answer the item correctly, the students must first understand literal meaning of the sentence *"The practice has been to treat the arts in chamber-of-commerce, rather than in creative, terms. "* (A3a). Then, they need to draw inferences about the sentence (A3b). The verbal reports by G.A. and K.R. indicate the prerequisite relationship between A3a and A3b:

> G.A.: …cause, **chamber-of-commerce, you go there when you want to start a business (A3a).** So, **the business which would be a commercial thing, if it is cultural, it is definitely the commercialization of culture (A3b).** I don't know, it's just a logical thought process to get to A.

> K.R.: The practice, what's the practice. The practice has been to treat the arts in chamber-of-commerce, rather than in creative, terms" That talks about the

> government, I think. **Chamber-of-commerce is government and it's saying the practice means what they have been doing, I mean, like, rather than in creative terms, treat the arts, treat the arts in government terms, rather than in creative terms (A3a).** ... Arts in chamber-of-commerce terms, rather than in creative terms, We are not being creative of the arts. **We are making money, maybe, that's what they mean. Commercialization...that's what commercialization means (A3b).** And money, make money. ... The commercialization rather than creative terms, I just go with A.

In both reports, the students started by trying to understand the key term "chamber-of-commerce" in terms of its literal meaning. G.A. understood it as a place where you go to start a business and K.R. understood it as a government institution. Based on that understanding, they could then infer that the "practice" refers to "commercialization of culture."

*A3a vs. A4a.* Item 4 measures both attribute A3a and attribute A4a. All eight students correctly answered the item. To answer the item correctly, the students need to understand the literal meaning of a larger section of text (A4a). However, before the larger section of text is understood, they must first understand each individual sentence in the text (A3a), as indicated by the verbal reports of L.G. and P.B.:

> L.G.: ...**it mainly focuses on the economic impact of recycling (A4a)**, which can be seen from these sentences: **They offer mainly short-term benefits to a few groups; ...Diverting money from genuine social and environmental problem; Recycling programs actually consume resources (A3a),** and the last sentence. I think that last sentence is the main idea of passage 2. ...**It is mainly about the cost of recycling, cost of money and human and natural resources (A4a).**
>
> P.B.: Passage 2 ...passage 2 is talking about how it's really only beneficial to a few groups. Um...and its taking away from things that can be considered beneficial ...Does talk about ...**how it's three times more expensive than**

> **collecting a ton of garbage. ... and recycling is a waste of time and money (A3a).** So **I am gonna say economic impact because most of them are focusing on the monetary things in passage 2 (A4a).**

As indicated above, both students referred to several key sentences in the paragraph and reported how they understood these sentences to get the meaning of the whole paragraph.

*A3a vs. A5.* Item 13 measures attributes A3a and A5. Seven students correctly answered the item. Students are required to understand the author's purpose in writing the segment *"temples to bourgeois muses with all the panache of suburban shopping centers"* (A5). To answer the item correctly, the literal meaning of the following sentence must first be understood (A3a):

> *(1) It has seen a few good new theaters and a lot of bad ones, temples to bourgeois muses with all the panache of suburban shopping centers.*

Consider the verbal reports of G.A. and D.C.:

> G.A.: The last decade has seen city after city...Ok, **it is the last decade,...so the last decade has a few good new theaters, and a lot of bad ones, temples to bourgeois muses, with all the panache of suburban shopping centers (A3a),** So that is just describing, ...the bad theaters, in this line it is a descriptive of the bad ones, ok, ...**temples to bourgeois muses line could be a dig at the character. And the shopping center thing could be a dig at the appearance (A5)**. Mmm, I'll put a star beside C, cause it might be good.
>
> D.C.: Let's see...(read the sentence) they are talking about cities, so description best serves to ...well, before that, in lines before that, **it said, it has seen a few good new theaters and a lot of bad ones (A3a), which really influenced me choosing C because it says deprecate the appearance and character of many new theaters (A5).** Not all theaters but many of them were bad.

Both students first understood the literal meaning of the sentence and then they could infer the author's purpose in writing "*temples to bourgeois muses with all the panache* *of suburban shopping centers.*"

On the other hand, failure to understand the sentence might lead students to select a wrong option, as indicated in the report by P.S.:

> P.S.: I read the sentence and I didn't know the exact meaning of this word bourgeois, I didn't know what exactly it meant, but I just thought that the answer from the 11th question could carry forward to the 12th question, So I said that it's A.

Due to word difficulty, P.S. could not understand the sentence. Therefore, she answered the item incorrectly.

*A3a vs. A8.* Item 2 measures attributes A3a and A8. All eight students answered the item correctly. Students are required to identify what rhetorical device was used in the paragraph (A8). To answer the item correctly, the literal meaning of one or more of the following sentences must be understood (A3a):

(1) *For him the sounds of the world were the ingredients he mixed into appetizers, main courses, and desserts to satisfy the appetite of his worldwide audience.*
(2) *He wasn't averse to going out in a boat to catch the fish himself.*
(3) *He would raise the fowl himself.*
(4) *But when that musical meal appeared before you none of the drudgery showed.*

Consider the verbal reports provided by K.R. and G.A.:

> K.R.: So what does he mean, when he is saying all of this. Okay...um...**the sounds of the world were the ingredients, it's talking about how he takes the sound that he hears and it's relating it to food (A3a)**, he wasn't averse to going out in a boat to catch the fish himself. ...Oh, so he ...he didn't like things that he didn't sort of have insistence, creating it, raising the fowl, the chicken, but the music meal appear before you...**Metaphor, through relating food and making food to his making music (A8)**...I think I probably say metaphor...because from rereading, from line 5 to the end, **he is comparing things that aren't shouldn't be compared with music that**

> **most people would think. He was comparing food, and like food mostly, and making food, and how you get to what you get to the restaurant from like the farm (A8).**
>
> G.A.: ...if you look back line8 to 10, **it's talking about, like, actually going out, and doing things to get the final product, like getting fish by going into the boat and catching it and having a chicken by raising it (A3a)**. So, in the end all you have is the fish or the chicken, there is no indication of how you got it. **So to me it is a metaphor of how he got his music, like final product of his music (A8).**

As indicated in these reports, before the students arrived at their answers, they must understand the literal meaning of the sentences. Then, by comparing the literal meaning of the sentences with the theme of the passage—creating music, they recognize the rhetorical device used by the author—metaphor.

*A4a vs. A4b.* Item 16 measures both attribute A4a and attribute A4b. Six students correctly answered this item. Students are required to identify the author's purpose in writing a paragraph (A5). However, to answer this item, the students must first understand the literal meaning of the paragraph (A4a) and then, make inferences about the paragraph (A4b). Consider the verbal reports provided by G.B. and P.B.:

> G.B.: To answer this question, you have to find what the author meant in the first place. **That wisdom, as it comes, blah-blah, ...is expressing what the author believes to be the economic reasons for people building the new centers of the arts (A4a).** Property values are not being very concerned about. Well, tradition, or perhaps, the older buildings are nicer. Just because you can make the new one shiny doesn't mean it is necessarily better. ...So, primarily the paragraph serves to ...criticize the way in which **cultural buildings are viewed as commodities. That's what it is (A4b).**
>
> P.B.: So...**what is this paragraph saying...uh...So, this is all about how to get, you know, more, how to stretch your dollar kind of thing, ...by building cultural centers (A4a).** So, yeah, they are commodity and that's

> just what the whole paragraph is talking about. Really I think **what the paragraph was conveying was how culture's being destroyed by modern values (A4b),** so criticize the way in which cultural buildings are viewed as commodities, so I picked it which was A.

In both verbal reports, the students first attempted to grasp the literal meaning of the paragraph. Then, based on their understanding of the paragraph, they made further inferences about the author's purpose in writing the paragraph.

*Revision of the Attribute Hierarchies*

The previous section helps illustrate the cognitive attributes used by the students in solving the SAT Critical Reading items. Results from the verbal analysis also helped specify the relationships among cognitive attributes. These results have implications for revising the initial attribute hierarchies specified based on the VanderVeen et al. (2007) model and a selected review of reading theories. The first revision was to incorporate the two additional cognitive attributes into the three attribute hierarchies. As attribute A8 requires A3a as its prerequisite, A8 should form another branch below attribute A3 in Hierarchy 1 and below attribute A3a in Hierarchies 2 and 3. Attribute A9, on the other hand, should form another independent branch directly connected to A1 as no prerequisite relationship was found between A9 and other attributes in the hierarchies. The second revision of the hierarchies concerned the placement of attribute A2. Originally, attribute A2 was an independent branch directly related to attribute A1. However, the verbal report data indicated that A2 required A3a as its prerequisite, hence, it should be placed below A3 in Hierarchy 1 and below A3a in Hierarchies 2 and 3. The revised version of Hierarchies 1, 2, and 3 are shown, respectively, in Figure 5a, 5b, and 5c.

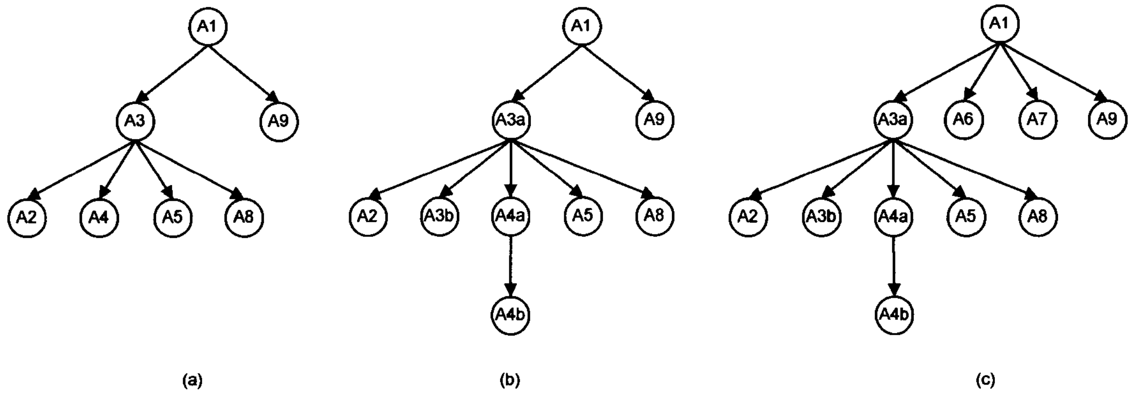


*Figure 5.* The three revised hierarchies in the present study.

Although the three initial hierarchies were revised according to the verbal report data, it must be noted that, at the present stage, it was not yet known which of the three revised hierarchies was the most effective in analyzing the response data for large samples of students who took the March 2005 SAT Critical Reading subtest. Thus, the mean *HCI* was used to evaluate which of the revised hierarchies provides the best model-data fit. Once discovered, the revised hierarchy that provided the best fit was used to analyze student response data on the SAT Critical Reading subtest.

## Results from the *HCI* Analysis

In this section, the results from the *HCI* analysis are reported. The psychometric features of the two randomly-selected samples are first described. Then, the *HCI* results for the three revised attribute hierarchies are reported and compared. Finally, using the criteria of model-data fit and parsimony, the best attribute hierarchy was selected.

### *Psychometric Features of the Two Samples*

Two samples of 2000 examinees were randomly selected from the March 2005 administration of the SAT. The psychometric features of the total scores on the 20 items for the two samples are presented in Table 13. The values of the means and standard deviations for the two samples are very similar. Moreover, the two samples have the same minimum and maximum total scores. To further examine whether the two samples were equivalent, an $F$-test and a Levene's test of homogeneity of variance were conducted. The results are reported in Table 14. The two tests indicate that there is no significant difference between the two sample means ($p > 0.10$) and that the variances of the two samples are equal ($p > 0.10$). As both the descriptive features and statistical tests indicate that the two samples were equivalent to one another, these two samples were used separately to calculate the *HCI* values for the three attribute hierarchies, thereby providing a cross-validation of the model fits.

Table 13

*Psychometric Features of the Two Samples*

| Sample | N | Mean | Std. Deviation | Minimum | Maximum |
|---|---|---|---|---|---|
| 1 | 2000 | 12.34 | 4.26 | 0 | 20 |
| 2 | 2000 | 12.40 | 4.17 | 0 | 20 |

Table 14

*Comparison of the Mean Total Scores for the Two Samples*

| | | Mean | | | Levene's | |
|---|---|---|---|---|---|---|
| | df | Square | F | Sig. | Statistic | Sig. |
| Between Groups | 1 | 2.916 | 0.164 | 0.686 | 0.685 | 0.408 |
| Within Groups | 3998 | 17.795 | | | | |
| Total | 3999 | | | | | |

### *HCI Results for the Three Hierarchies*

The *HCI* values were calculated using the Mathematica syntax developed by Cui, Leighton, Gierl et al. (2006). The means and standard deviations of the *HCI* values for the three hierarchies are summarized in Table 15.

Table 15

*A Summary of the HCI Values for the Three Attribute Hierarchies*

| | Hierarchy 1 | | Hierarchy 2 | | Hierarchy 3 | |
|---|---|---|---|---|---|---|
| | $\bar{X}$ | *SD* | $\bar{X}$ | *SD* | $\bar{X}$ | *SD* |
| Sample 1 | 0.44 | 0.48 | 0.55 | 0.43 | 0.57 | 0.43 |
| Sample 2 | 0.45 | 0.47 | 0.56 | 0.42 | 0.58 | 0.42 |

The values in Table 15 suggest moderate model-data fit for all three hierarchies. From Hierarchy 1 to Hierarchy 2, the mean *HCI* values for both samples increased by 0.11. However, from Hierarchy 2 to Hierarchy 3, the mean *HCI* values for both samples only increased by 0.02. Thus, comparatively, Hierarchy 2 improved model-data fit considerably but Hierarchy 3 did so only negligibly[3]. Given the fact that Hierarchy 2 involves fewer attributes and inter-attribute connections and, thus, is more parsimonious than Hierarchy 3, Hierarchy 2 was regarded as the best among the three hierarchies and, thus, was used for analyzing student response data.

[3] As there is no established guidelines or statistical tests for evaluating model-data fit using the *HCI* at this stage, the evaluation for the model-data fit of the attribute hierarchies has to be purely judgmental.

# CHAPTER V: RESULTS FROM THE PSYCHOMETRIC ANALYSIS

In the previous chapter, Hierarchy 2 was selected for analyzing student response data due to its parsimony and its degree of model-data fit. In this chapter, Hierarchy 2 was used to demonstrate how the AHM can be used in test analysis. The $Q_r$ matrix for the 20 items was specified according to the cognitive attributes identified from student verbal reports. Then, attribute pattern and expected response pattern matrices based on Hierarchy 2 were generated and used for calculating attribute probabilities and determining examinees' attribute mastery. Attribute reliabilities and the SEM were also calculated. Based on the information of attribute probability, attribute reliability, and the SEM, descriptive score reports were compiled for a sample of students to demonstrate how the AHM could be used to provide examinees with cognitive diagnostic feedback.

## Generating the $Q_r$, Attribute Pattern, and Expected Response Pattern Matrices

The $Q_r$, attribute pattern, and expected response pattern matrices provide necessary information for a neural network to calculate attribute probabilities. In theory, the $Q_r$ matrix is obtained according to the hierarchical relationships among cognitive attributes. In practice, however, when the test is not developed according to the specification of a $Q_r$ matrix, the $Q_r$ matrix has to be specified by other means. For the present study, the $Q_r$ matrix was specified according to the cognitive attributes identified from student verbal reports.

The cognitive attributes measured by each of the 20 items were identified in Table

12. Based on this information, the $Q_r$ matrix for Hierarchy 2 was specified, as follows

$$\begin{bmatrix} 11111111111111111111 \\ 00000100101000000000 \\ 11111111111111111111 \\ 00000000011101101001 \\ 00111101000011111110 \\ 00000000000010010110 \\ 00001010001010110000 \\ 01000010000001100001 \\ 00000110001110111001 \end{bmatrix}. \qquad \text{(Matrix 7)}$$

The $Q_r$ matrix for Hierarchy 2 has 20 columns and nine rows, indicating the 20 test items included in the present study and the nine cognitive attributes in Hierarchy 2, respectively. A "1" in the matrix means the corresponding attribute is measured and a "0" means the corresponding attribute is not measured by the item. For example, column 1 of the matrix indicates that item 1 measured attribute A1 and attribute A3a, as there is a "1" on row 1 and row 3, respectively.

After the $Q_r$ matrix was specified, the attribute pattern matrix based on Hierarchy 2 was generated. Attribute pattern refers to the combination of attributes that is consistent with the attribute hierarchy. The transposed attribute pattern matrix of Hierarchy 2 is shown in Matrix 8. Matrix 8 has 77 columns and 9 rows, indicating that Hierarchy 2, with its hierarchical structure among the nine cognitive attributes, can generate 77 unique attribute patterns for the population of hypothetical examinees, when there are no slips.

$$\begin{bmatrix}
11111111111111111111111111111111111111111111111111111111111111111111111111111\\
00001111000000111111110000000111111110001110011000000011111110000111100001111\\
01111111111111111111111111111111111111111111111111111111111111111111111111111\\
00110111001111001100110001111001100110110110101001111001100110011001100110011\\
00000000111111111111111111111111111110000000000111111111111111111111111111111\\
00000000010101010101010101101010101010101000000000010101010101010101010101010101\\
00000001000000000000000111111111111111100000011100000000000000111111111111111111\\
00000000000000000000000000000000000001111111111111111111111111111111111111111\\
00010011000011000011110010011000011110010011111000011000011110000000011111111
\end{bmatrix}. \tag{Matrix 8}$$

Next, the expected response patterns for Hierarchy 2 were generated. The transposed expected response matrix is shown in Matrix 9. The 77 columns in Matrix 9 indicate the 77 expected response patterns, corresponding to the 77 attribute patterns shown in Matrix 8. The 20 rows of Matrix 9 indicate examinees' expected responses to the 20 test items included in the present study. For example, consider the third column of Matrix 8 and Matrix 9. The third column of Matrix 8 indicates the expected examinee has mastered attributes A1, A3a, and A3b. The corresponding column in Matrix 9 means that this expected examinee is able to correctly answer items 1 and 10 but not the other 18 items. If an examinee has correctly answered only items 1 and 10, then diagnostic feedback can be provided, indicating that this examinee has mastered attributes A1, A3a, and A3b but still needs improvement on the other attributes.

## Estimating Attribute Probabilities

After the $Q_r$, attribute pattern, and expected response pattern matrices were generated, attribute probabilities were estimated according to students' response data. The estimation was done using Method NN. Three value ranges were specified to aid in interpreting the attribute probabilities produced from Method NN estimation:

a. $0 \le p < 0.50$,

b. $0.50 \le p < 0.80$, and

c. $0.80 \le p \le 1.00$.

Range *a* is regarded as non-mastery, range *b* as partial mastery, and range *c* as mastery

$$\begin{bmatrix}
01111111111111111111111111111111111111111111111111111111111111111111111111111 \\
00000000000000000000000000000000000001111111111111111111111111111111111111111 \\
00000000111111111111111111111111111110000000000111111111111111111111111111111 \\
00000000111111111111111111111111111110000000000111111111111111111111111111111 \\
00000000000000000000001111111111111110000000000000000000000001111111111111111 \\
00000000000000000011110000000000011110000000000000000000111100000000000001111 \\
00000000000000000000000000000000000000000001111000000000000000000000011111111 \\
00000000111111111111111111111111111110000000000111111111111111111111111111111 \\
00001111000000111111110000000111111110001110011000000011111110000111100001111 \\
00110111001111001100110001111001100110110110101001111001100110011001100110011 \\
00000001000000000000000000000000000110000000001000000000000000000000000000011 \\
00010011000011000000110000011000000110010010101000011000000110000000000110011 \\
00000000000000000000000010001000001010000000000000000000000000000000001010101 \\
00000000000000000000000000000000000000000000000001111001100110011001100110011 \\
00000000000000000000000000000000000000000000000000000000000000000000000110011 \\
00000000000000000000000010001000001010000000000000000000000000000000001010101 \\
00000000000011000000110000011000000110000000000000011000000110000000000110011 \\
00000000010101010101011010101010101010000000000010101010101010101010101010101 \\
00000000010101010101011010101010101010000000000010101010101010101010101010101 \\
00000000000000000000000000000000000000010010101000011000000110000000000110011
\end{bmatrix}. \qquad \text{(Matrix 9)}$$

in the present study. Other classification ranges could be proposed and used. The present three level system was adopted to illustrate the reporting approach.

The attribute probabilities for a random sample of 15 students were estimated and the results are displayed in Table 16 (in an actual testing situation, the attribute probabilities for all 2000 students would be estimated). The first column of Table 16 is a listing of the item response patterns of the 15 students. The second column contains the total scores of the 15 students. The next nine columns display the attribute probabilities of the 15 students on the nine cognitive attributes in Hierarchy 2. In other words, in the framework of the AHM, the examinees are provided with not only a total score but also detailed information about their attribute performance levels. For example, examinee 1 has a response pattern of "11110001110100111011." Out of the 20 items, the examinee answered 13 items correctly for a total score of 13. From the attribute probabilities, it can be said that the examinee has mastered attributes A1, A3a, A3b, A4a, A8, and A9, as the probabilities for these attributes are all over 0.80. This examinee has also partially mastered attributes A2 and A4b because the attribute probabilities were 0.733 for attribute A2 and 0.582 for attribute A4b. On the other hand, the examinee has not mastered attribute A5, as the attribute probability is only 0.003.

Table 16 summarizes the numbers of masters, partial-masters, and non-masters for each of the nine cognitive attributes in Hierarchy 2 (the last three rows of the table) and

Table 16

*Attribute Probabilities for a Random Sample of 15 Students Who Wrote on the 20 SAT Critical Reading Test Items in March 2005*

| | | | Attribute Probabilities | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Student | Total | | | | | | | | | | | | |
| Examinee | Response Data | Score | A1 | A2 | A3a | A3b | A4a | A4b | A5 | A8 | A9 | M | PM | NM |
| 1 | 11110001110100111011 | 13 | 0.999 | 0.733 | 0.990 | 0.998 | 0.982 | 0.582 | 0.003 | 0.987 | 0.998 | 6 | 2 | 1 |
| 2 | 11101101111000011101 | 13 | 0.997 | 1.000 | 0.999 | 0.994 | 0.993 | 0.218 | 0.999 | 0.994 | 1.000 | 8 | 0 | 1 |
| 3 | 11111101111100000000 | 11 | 0.998 | 1.000 | 0.999 | 0.994 | 0.987 | 0.019 | 0.999 | 0.989 | 1.000 | 8 | 0 | 1 |
| 4 | 11111101110011010110 | 14 | 0.999 | 0.998 | 0.999 | 0.998 | 0.999 | 0.998 | 0.988 | 0.986 | 1.000 | 9 | 0 | 0 |
| 5 | 11111111111111111110 | 19 | 0.999 | 0.996 | 0.998 | 0.995 | 0.999 | 0.985 | 0.998 | 0.994 | 1.000 | 9 | 0 | 0 |
| 6 | 11111110101100101011 | 14 | 0.998 | 0.980 | 0.998 | 0.800 | 0.909 | 0.381 | 1.000 | 0.998 | 1.000 | 8 | 0 | 1 |
| 7 | 10110011110001000101 | 10 | 0.999 | 0.995 | 0.999 | 0.998 | 0.999 | 0.741 | 0.930 | 0.347 | 0.913 | 7 | 1 | 1 |
| 8 | 01110000100000100000 | 5 | 0.994 | 1.000 | 0.813 | 0.026 | 0.905 | 0.042 | 0.013 | 0.997 | 0.003 | 5 | 0 | 4 |
| 9 | 11111001111111111110 | 17 | 0.999 | 0.992 | 0.998 | 1.000 | 1.000 | 0.997 | 0.985 | 0.981 | 0.999 | 9 | 0 | 0 |
| 10 | 11110010000101011100 | 10 | 0.999 | 0.005 | 0.999 | 0.492 | 1.000 | 0.959 | 0.999 | 0.931 | 1.000 | 7 | 0 | 2 |
| 11 | 11111001011101111010 | 14 | 0.999 | 0.011 | 0.998 | 0.994 | 1.000 | 0.460 | 0.999 | 0.997 | 0.999 | 7 | 0 | 2 |
| 12 | 11010000000000000000 | 3 | 0.999 | 0.015 | 0.995 | 0.013 | 0.690 | 0.013 | 0.015 | 0.998 | 0.033 | 3 | 1 | 5 |
| 13 | 11011010110111011111 | 15 | 0.998 | 0.998 | 0.999 | 0.987 | 0.991 | 0.986 | 1.000 | 0.994 | 1.000 | 9 | 0 | 0 |
| 14 | 10101100100010110000 | 8 | 0.998 | 0.998 | 0.967 | 0.005 | 0.761 | 0.085 | 0.879 | 0.000 | 1.000 | 5 | 1 | 3 |
| 15 | 01100000101010000100 | 6 | 0.998 | 1.000 | 0.987 | 0.098 | 0.977 | 0.930 | 0.657 | 0.989 | 0.838 | 7 | 1 | 1 |
| M | | | 15 | 11 | 15 | 10 | 13 | 6 | 11 | 13 | 13 | | | |
| PM | | | 0 | 1 | 0 | 0 | 2 | 2 | 1 | 0 | 0 | | | |
| NM | | | 0 | 3 | 0 | 5 | 0 | 7 | 3 | 2 | 2 | | | |

*Note.* In Table 16, M indicates Mastery, PM indicates Partial Mastery, and NM indicates Non-Mastery.

for each of the 15 examinees (the last three columns of the table). As indicated in the last three rows, attributes A1 and A3a were mastered by all 15 examinees while other attributes were mastered by different numbers of examinees. As attribute A1 is assumed to have been mastered by all target examinees of the SAT, the fact that this attribute was estimated to be mastered by all 15 examinees indicates that Method NN was successful in estimating the probabilities of this attribute. Mastery of attribute A3a was also expected because, in Hierarchy 2, A3a is a prerequisite to six other cognitive attributes and it is measured by all 20 items. In other words, to be able to answer most of the test items, the examinees must have this attribute. Apart from attributes A1 and A3a, other attributes were mastered by different numbers of examinees. In descending order, attributes A4a, A8, A9, A5, A2, A3b, and A4b were mastered by 13, 13, 13, 11, 11, 10, and 6 examinees, respectively.

From Table 16, a general finding is apparent: the higher the total score, the more attributes an examinee masters. This finding gives some legitimacy for reporting total scores to students, as is done in most operational testing programs. For example, examinees 4 and 5 received scores of 14 and 19, respectively, and the estimation results indicate that they have mastered all nine attributes in Hierarchy 2. On the other hand, examinees 8 and 12 received scores of 5 and 3, respectively, and the estimation results indicate that they have mastered five and three attributes, respectively. A correlation analysis was conducted to confirm this finding. The results of the correlation analysis, as shown in Table 17, indicate positive and high correlation between the total scores and the number of mastered attributes, indicating a strong relationship between the total score an examinee obtains and the number of attributes the examinee has mastered.

Table 17

*Correlation between Total Scores and the Attribute Mastery*

| | Mastery | Medium-Mastery | Non-Mastery |
|---|---|---|---|
| Total Score | 0.844* | -0.396* | -0.825* |

* $p < 0.05$.

However, it must be noted that the correlation between total scores and attribute mastery is not perfect. One possible reason for such results lies in the fact that examinees who received a lower total score may have mastered more cognitive attributes than those who received a higher total score. For example, examinee 14 received a total score of 8, and attribute probabilities indicate that this examinee has mastered five attributes, partially mastered one attribute, and lacks three attributes. However, attribute probabilities indicate that examinee 15, who received a total score of only 6, has mastered seven attributes, partially mastered one attribute, and lacks only one attribute. These results indicate that providing an examinee with only a single total score may misrepresent their cognitive skill profiles. The reason for such an anomaly may be traced back to the response patterns of the examinees, the attributes measured by the test items, and the contribution of each attribute to the items. Examinee 14 has a response pattern of "10101100100010110000" and examinee 15 has a response pattern of "01100000101010000100". Both examinees correctly answered one item that measures attribute A8: Examinee 14 correctly answered Item 15 and examinee 15 correctly answered Item 2. However, because Item 2 measures only three attributes (A1, A3a, and A8) but Item15 measures as many as seven attributes (A1, A3a, A3b, A4a, A5, A8, and A9), attribute A8 contributes more to Item 2 than to Item 15 (see Table 18 for the detailed weights of attribute A8 in Item 2 and Item 15 in the next section). Hence, when examinee 15 correctly answered Item 2, it is very likely that the examinee has mastered attribute A8. However, when examinee 14 correctly answered Item 15, no such conclusion can be drawn because many other attributes can lead to the correct answer. The same interpretation can be used to explain why it was estimated that examinee 14 has not mastered attribute A4b but examinee 15 has.

Table 16 also indicates that examinees with the same total scores may possess different cognitive attributes if their response patterns are different. For example, examinees 1 and 2 both received a total score of 13 out of 20. However, because their response patterns are different ("11110001110100111011" for examinee 1 and "11101101111000011101" for examinee 2), the estimated attribute probabilities indicate the two examinees have mastered different cognitive attributes. Examinee 1 has mastered attributes A1, A3a, A3b, A4a, A8, and A9, partially mastered attributes A2 and A4b, but lacks attribute A5, while examinee 2 has mastered all attributes but A4b. These results suggest that students who obtained the same total score on a test do not necessarily have the same cognitive skill profiles, especially when a test measures a variety of cognitive attributes. These results also point to the disadvantages of classical test theory and unidimensional item response theory where only a total score or a latent ability score is reported, when the goal is also to make diagnostic inferences about an examinee's skill profile.

## Estimating the Reliabilities and SEM for the Attributes in Hierarchy 2

Attribute probabilities were estimated for the random sample of 15 examinees and the results of their attribute mastery have been summarized. However, it is not clear how consistent the information of attribute mastery would be if a different test was administered. To evaluate the consistency of the decisions about the examinees' attribute mastery, reliabilities for the attributes in Hierarchy 2 were calculated. Attribute reliabilities would also be used to calculate the standard error of measurement for each attribute. With the SEM, confidence intervals can be built for attribute probabilities in score reporting.

### *Estimation of Attribute Reliabilities*

As the 20 test items used for the present study measure a combination of different attributes (see Table 12), each attribute only contributes to part of the total item-level variance. In order to isolate the contribution of each attribute to an examinee's item-level performance, the item score has to be weighted by the subtraction of two conditional probabilities. The first conditional probability is associated with attribute mastery (i.e., an examinee who has mastered the attribute can answer the item correctly) and the second conditional probability is associated with attribute non-mastery (i.e., an examinee who has not mastered the attribute can answer the item correctly). The weightings of the item scores were obtained by first simulating a normal distribution of examinees. For the present study, a normal distribution of 5000 examinees was simulated in calculating the weightings of the items scores. The weighting results of the nine attributes in Hierarchy 2 for the 20 items are displayed in Table 18.

Table 18

*The Weighting Results of the Nine Attributes for the 20 Items*

| Item | A1 | A2 | A3a | A3b | A4a | A4b | A5 | A8 | A9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.141 | 0.000 | 0.141 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 2 | 0.074 | 0.000 | 0.074 | 0.000 | 0.000 | 0.000 | 0.000 | 0.519 | 0.000 |
| 3 | 0.102 | 0.000 | 0.102 | 0.000 | 0.182 | 0.000 | 0.000 | 0.000 | 0.000 |
| 4 | 0.102 | 0.000 | 0.102 | 0.000 | 0.182 | 0.000 | 0.000 | 0.000 | 0.000 |
| 5 | 0.048 | 0.000 | 0.048 | 0.000 | 0.085 | 0.000 | 0.499 | 0.000 | 0.000 |
| 6 | 0.024 | 0.235 | 0.024 | 0.000 | 0.043 | 0.000 | 0.000 | 0.000 | 0.153 |
| 7 | 0.016 | 0.000 | 0.016 | 0.000 | 0.000 | 0.000 | 0.164 | 0.111 | 0.101 |
| 8 | 0.102 | 0.000 | 0.102 | 0.000 | 0.182 | 0.000 | 0.000 | 0.000 | 0.000 |
| 9 | 0.070 | 0.693 | 0.070 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 10 | 0.075 | 0.000 | 0.075 | 0.377 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 11 | 0.007 | 0.072 | 0.007 | 0.037 | 0.000 | 0.000 | 0.076 | 0.000 | 0.047 |
| 12 | 0.038 | 0.000 | 0.038 | 0.191 | 0.000 | 0.000 | 0.000 | 0.000 | 0.244 |
| 13 | 0.011 | 0.000 | 0.011 | 0.000 | 0.019 | 0.088 | 0.113 | 0.000 | 0.069 |
| 14 | 0.029 | 0.000 | 0.029 | 0.147 | 0.052 | 0.000 | 0.000 | 0.206 | 0.000 |
| 15 | 0.003 | 0.000 | 0.003 | 0.017 | 0.006 | 0.000 | 0.036 | 0.024 | 0.022 |
| 16 | 0.011 | 0.000 | 0.011 | 0.000 | 0.019 | 0.088 | 0.113 | 0.000 | 0.069 |
| 17 | 0.026 | 0.000 | 0.026 | 0.131 | 0.046 | 0.000 | 0.000 | 0.000 | 0.167 |
| 18 | 0.051 | 0.000 | 0.051 | 0.000 | 0.091 | 0.413 | 0.000 | 0.000 | 0.000 |
| 19 | 0.051 | 0.000 | 0.051 | 0.000 | 0.091 | 0.413 | 0.000 | 0.000 | 0.000 |
| 20 | 0.020 | 0.000 | 0.020 | 0.100 | 0.000 | 0.000 | 0.000 | 0.139 | 0.127 |

Table 18 shows that the same cognitive attribute may have different weighted scores in items which measure different combinations of attributes, indicating that different cognitive attributes may have differing contribution to an item. For example, Item 1 measures attributes A1 and A3a, and the weighted scores for these two attributes are both 0.141. However, the weighted scores of the same two attributes reduce to 0.074 in Item 2 because Item 2 measures an additional attribute, A8, whose weighted score is a much larger number of 0.519. Hence, for Item 2, the contribution of attribute A1 and A3a becomes relatively less due to the contribution of attribute A8. Such results can also help explain the anomaly in the attribute mastery results for examinees 14 and 15 presented in the previous section. Both examinees correctly answered one item that measures attribute A8: Examinee 14 correctly answered Item 15 and examinee 15 correctly answered Item 2. However, the weighted score of A8 in Item 15 is only 0.024 but is 0.519 in Item 2, indicating more contribution of A8 in Item 2 than in Item 15. Hence, it was estimated that examinee 14 had not mastered attribute A8 while examinee 15 had mastered attribute A8.

The weighting results in Table 18 and the response data of the two samples of 2000 examinees were used to calculate the attribute reliabilities. The results are displayed in Table 19. The values of the attribute reliabilities are comparable across the two samples, indicating the relative consistency of the attribute reliability index. Moreover, the relationship between the number of items that measure an attribute and the reliability estimate of the attribute is apparent from Tables 18 and 19. Generally, the more items that measure a certain attribute, the higher the attribute reliability estimate. When the number of items that measure a certain attribute decreases, the reliability estimates also decrease. For example, attributes A1 and A3a are measured by all 20 items and thus had the highest reliability estimates of 0.74. Attribute A2, on the other hand, had the lowest reliability estimates of 0.21 and 0.20 because only three items measure this attribute.

Table 19

*Attribute Reliabilities for the 20-item Test and a Hypothetical 60-item Test*

| Sample | No. of Items | A1 | A2 | A3a | A3b | A4a | A4b | A5 | A8 | A9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 20 | 0.74 | 0.21 | 0.74 | 0.45 | 0.64 | 0.47 | 0.37 | 0.33 | 0.58 |
|  | 60 | 0.85 | 0.35 | 0.85 | 0.62 | 0.78 | 0.64 | 0.54 | 0.50 | 0.73 |
| 2 | 20 | 0.74 | 0.20 | 0.74 | 0.44 | 0.65 | 0.46 | 0.36 | 0.34 | 0.55 |
|  | 60 | 0.85 | 0.33 | 0.85 | 0.61 | 0.79 | 0.63 | 0.53 | 0.51 | 0.71 |

Another finding that can be noted in Table 19 is that all the attribute reliability estimates, which are between 0.20 and 0.74, are relatively low (see Row 1). If a reliability of 0.80 is regarded as a threshold value for decision consistency, then the decisions made on none of the nine attributes would be considered consistent. Such results would make it difficult for the test to provide diagnostic feedback to examinees because such diagnostic feedback is unreliable. Fortunately, the previous finding that a relationship exists between the number of items that measure an attribute and the reliability estimate of the attribute points to a possible solution. If we want to improve the reliability about examinee's attribute mastery, then a simple way would be to increase the number of items that measure each attribute. The Spearman-Brown formula can be used to estimate attribute reliabilities if the test were increased to a certain length by adding parallel item sets (Gierl et al., in press). The Spearman-Brown formula adapted to the AHM is specified as

$$\alpha_{AHM-SB_k} = \frac{n_k \alpha_{AHM}}{1 + (n_k - 1)\alpha_{AHM}} \qquad \text{(Equation 6)}$$

where $\alpha_{AHM-SB_k}$ is the Spearman-Brown reliability of attribute *k* if $n_k$ additional item sets that are parallel to items measuring attribute *k* are added to the test. The reliability estimates for the nine attributes in Hierarchy 2 with a hypothetical 60-item test (i.e., three parallel sets of 20 items) are displayed in the second row of Table 19. With three sets of parallel items, the reliability estimates for all attributes except attribute A2 increased considerably. However, the reliability estimates for attribute A2 remained low at 0.35 and 0.33 for the two samples.

One problem arises from the above results: Although the reliability estimates could be improved by increasing the number of items measuring each attribute, the fact that the cognitive attributes in the AHM are hierarchically related makes it difficult for these attributes to achieve the same level of reliability. This result occurs because, in the AHM, if a test item measures a certain cognitive attribute, it will also measure the prerequisite of this attribute. In other words, an attribute will generally be measured by fewer items than its prerequisite attribute and, thus, will get a lower reliability estimate.

This situation requires further study.

*Calculation of SEM for the Attributes in Hierarchy 2*

After the attribute reliabilities are estimated, the SEM can be calculated for the attributes using Equation 5. The SEM was calculated using the response data of the two samples of 2000 examinees who took the March 2005 administration of the SAT. The results are displayed in Table 20. These values can be used to build confidence intervals for the examinees' attribute probabilities. For example, if an examinee has a probability of 0.80 on attribute A3a, then the 95% confidence interval of the probability that this examinee has mastered attribute A3a is between 0.742 and 0.858. In other words, if the examinee takes a parallel form of the test, his/her probability of mastering attribute A3a falls within this interval 95% of the time.

Table 20

*The SEM for the Nine Attributes in Hierarchy 2*

| Sample | A1 | A2 | A3a | A3b | A4a | A4b | A5 | A8 | A9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.001 | 0.337 | 0.029 | 0.284 | 0.106 | 0.306 | 0.264 | 0.321 | 0.165 |
| 2 | 0.001 | 0.339 | 0.029 | 0.287 | 0.105 | 0.309 | 0.266 | 0.318 | 0.171 |

Although the SEM is useful in building confidence intervals for attribute probabilities, it must be noted that, in this study, large values of SEM were produced due to low reliabilities for most of the attributes in Hierarchy 2. These large values will necessarily result in wide confidence intervals, indicating that the decisions about students' attribute mastery would be relatively inconsistent across tests. To correct this situation, efforts must be made to first improve the reliabilities for the attributes (e.g., by adding a larger number of parallel items to the test).

## Score Reporting and Providing Diagnostic Feedback

A key advantage of the AHM is that it provides a specific method for diagnostic score reporting, as examinees receive individualized score reports based on their attribute mastery levels. The score reports produced by the AHM not only have a total score but also have detailed information about what cognitive attributes were measured by the test and the degree to which the examinees have mastered these cognitive attributes. Also, because attributes are hierarchically related (e.g., A3a and A3b), the score reports can reflect such relationships among the attributes and this information can be conveyed to the examinee and their instructors. For example, one of the principles in specifying the attribute hierarchies concerns whether inferences are involved in reading comprehension. The score reports can then be composed in such a way that they reflect at which level (literal level or inferential level) the examinees' text comprehension is situated. The reliabilities and the SEM of individual attributes, as illustrated in the previous section, can also be incorporated into examinees' attribute probabilities to build confidence intervals for examinees' attribute mastery.

To demonstrate how the AHM can be used to report test scores and provide diagnostic feedback, four examinees were chosen from the 15 examinees whose response data were used in estimating attribute probabilities in a previous section. The four examinees are examinees 5, 1, 8, and 12.

Examinees 5 and 1 are chosen to demonstrate how detailed information regarding the examinees' attribute mastery can be provided to the examinees. Examinee 5 correctly answered 19 items out of 20. According to the estimated attribute probabilities, this examinee mastered all nine attributes in Hierarchy 2. Examinee 1 correctly answered 13 items. According to the estimated attribute probabilities, this examinee mastered attributes A1, A3a, A3b, A4a, A8, and A9, partially mastered attributes A2 and A4b, but lacks attribute A5. Thus, this examinee is considered to lack the ability to analyze the author's purposes, goals and strategies. Moreover, Examinee 1 may not efficiently determine the meaning of difficult words from context (A2) or understand larger sections of text by making inferences or integrating background knowledge (A4b). These results demonstrate that the score reports from the AHM provide detailed information to the examinees about their cognitive attribute mastery levels. Such information can be used by examinees to improve specific cognitive skills, thereby increasing their test performances.

Examinees 8 and 12 are chosen to demonstrate how the hierarchical relationships among cognitive attributes can be incorporated into score reports to increase the specificity of diagnostic feedback. Examinee 8 correctly answered five items. The estimated attribute probabilities indicate that this examinee mastered five attributes (A1, A2, A3a, A4a, and A8), but lacks three attributes (A3b, A4b, A5). As both attributes A3b and A4b represent the inferential text comprehension, the absence of these two attributes suggests that Examinee 8's comprehension of text rests at a literal level while he has difficulty in understanding text inferentially. Examinee 12 only answered three items correctly. The estimated attribute probabilities indicate that this examinee only mastered three attributes (A1, A3a, and A8), partially mastered one attribute (A4a), and lacks the other five attributes. As the examinee lacks both attributes A3b and A4b, the same diagnostic feedback provided to examinee 8 can be provided to this examinee. Moreover, as the examinee only partially mastered attribute A4a and did not master attribute A4b, which both deal with larger sections of text, the conclusion may be drawn that this examinee can only process text at the sentence level, but may have difficulty in processing larger sections of text (A4a, A4b). Descriptions of the cognitive strengths and weaknesses for the four examinees are summarized in Table 21. The descriptions would make it possible for these examinees to identify their cognitive strengths and weaknesses in reading. It must also be noted that, because the descriptions were based mainly on the VanderVeen et al. (2007) model, which had coarse grain size and was validated from the verbal reports of only eight students, some of the attributes might not be precise. For the feedback to be operationally more applicable, more precise attributes must be obtained through additional studies.

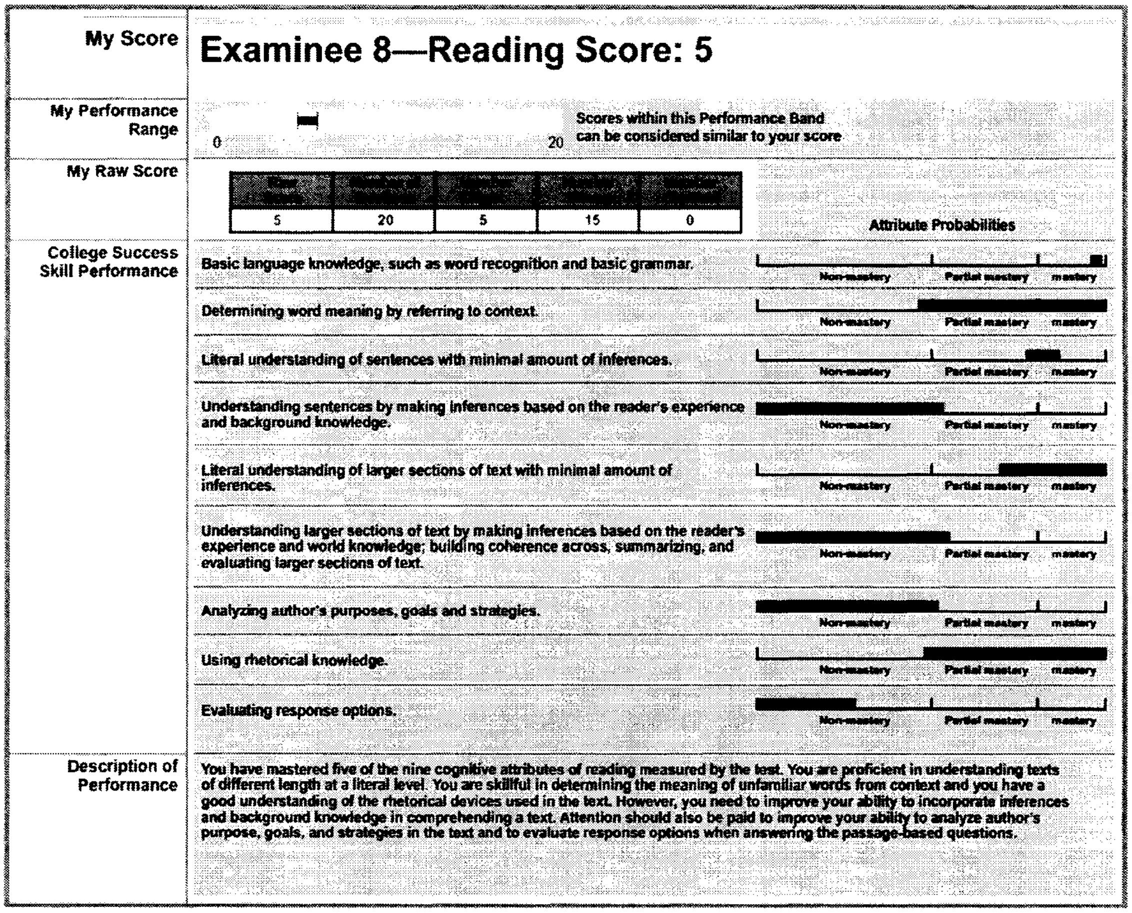
Table 21

*Sample Descriptive Score Reports for Examinees 5, 1, 8, and 12*

| Examinee | Total Score | Description of Performance |
|---|---|---|
| 5 | 19 | You have mastered the nine cognitive attributes of reading measured by the test. You are proficient in understanding texts of different length at both literal level and inferential level. You are also proficient in analyzing author's purpose, goals, and strategies in the text. You are skillful in determining the meaning of unfamiliar words from context and you have a good understanding of the rhetorical devices used in the text. In addition, you have good ability in evaluating response options to get the correct answer. |
| 1 | 13 | You have mastered six, and partially mastered another two of the nine cognitive attributes of reading measured by the test. You are proficient in understanding texts of different length at literal level. Although you can incorporate inferences and background knowledge in comprehending text, you need strengthen this ability when reading larger chunk of text. You have a good understanding of the rhetorical devices used in the text and a good ability in evaluating response options to get the correct answer. There is still room for improving your skill in determining the meaning of unfamiliar words from context. Attention should also be paid to improve your ability to analyze author's purpose, goals, and strategies in the text. |
| 8 | 5 | You have mastered five of the nine cognitive attributes of reading measured by the test. You are proficient in understanding texts of different length at literal level. You are skillful in determining the meaning of unfamiliar words from context and you have a good understanding of the rhetorical devices used in the text. However, you need to improve your ability to incorporate inferences and background knowledge in comprehending a text. Attention should also be paid to improve your ability to analyze author's purpose, goals, and strategies in the text and to evaluate response options when answering the passage-based questions. |
| 12 | 3 | You have mastered three of the nine cognitive attributes of reading measured by the test. You are proficient in understanding |

| | |
|---|---|
| | sentences at literal level. You have a good understanding of the rhetorical devices used in the text. However, you should learn how to incorporate inferences and background knowledge into comprehending texts of different lengths. Moreover, you need to improve ability in reading larger chunks of text. You also need to improve your skill in determining the meaning of unfamiliar words from context. Attention should also be paid to improve your ability to analyze author's purpose, goals, and strategies in the text and to evaluate response options when answering the passage-based questions. |

An exemplary score report for Examinee 8, as displayed in Figure 6, is also used to illustrate how the results from the AHM could be used in score reporting. The left panel of Figure 6 indicates that the score report has five parts: the examinee's score, performance range, raw score, the cognitive attributes measured by the test, and description of the examinee's performance. The examinee's score may be any scaled score, depending on which scale the testing program is using. In this example, the raw score was used for an illustration. The examinee's performance range is a confidence interval for the examinee's scaled score. The raw score part of the report displays the examinee's raw score and the numbers of correct, incorrect, and omitted items. Such information is generally included in traditional score reports. What distinguishes the score report produced using the results of the AHM from traditional score reports is the remaining two parts of the report. The fourth part, displayed in the central part of Figure 6, includes information about the cognitive attributes measured by the test and the examinee's mastery levels of each individual attribute. The mastery levels of the attributes are also expressed in the form of confidence intervals, which were obtained by incorporating the SEM of the attribute probabilities. However, it must be noted that, due to the low reliability estimates, the confidence intervals for most of the attributes are wide. The fifth part of the report, displayed in the lower part of Figure 6, is a description of the examinee's cognitive strengths and weaknesses, which can be used to guide teaching and learning.



My Score

**Examinee 8—Reading Score: 5**

My Performance Range

0 20 Scores within this Performance Band can be considered similar to your score

My Raw Score

| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
|---|---|---|---|---|
| 5 | 20 | 5 | 15 | 0 |

Attribute Probabilities

College Success Skill Performance

Basic language knowledge, such as word recognition and basic grammar.

Determining word meaning by referring to context.

Literal understanding of sentences with minimal amount of inferences.

Understanding sentences by making inferences based on the reader's experience and background knowledge.

Literal understanding of larger sections of text with minimal amount of inferences.

Understanding larger sections of text by making inferences based on the reader's experience and world knowledge; building coherence across, summarizing, and evaluating larger sections of text.

Analyzing author's purposes, goals and strategies.

Using rhetorical knowledge.

Evaluating response options.

Description of Performance

You have mastered five of the nine cognitive attributes of reading measured by the test. You are proficient in understanding texts of different length at a literal level. You are skillful in determining the meaning of unfamiliar words from context and you have a good understanding of the rhetorical devices used in the text. However, you need to improve your ability to incorporate inferences and background knowledge in comprehending a text. Attention should also be paid to improve your ability to analyze author's purpose, goals, and strategies in the text and to evaluate response options when answering the passage-based questions.

*Figure 6.* Sample score report for examinee 8.

## Summary of Results

Chapter IV reported the results from the cognitive analysis and identified the final model for the psychometric analysis. The cognitive analysis consisted of a verbal report study and the *HCI* analysis. Eight participants from a variety of educational background were recruited and their verbal reports were coded by two graduate students in educational psychology, who have experience in conducting verbal report studies and are familiar with reading comprehension and the assessment of reading. The verbal report data confirmed that all cognitive attributes in the three hierarchies (with the exception of attribute A1, which is assumed to have mastered by all potential examinees of the SAT and thus was not coded) were used by the examinees in their problem solving. Two additional cognitive attributes, using rhetorical knowledge (A8) and evaluating response options (A9), were discovered from the verbal report data. The attributes measured by each of the 20 test items were determined according to their frequencies appearing in the verbal reports and the final results are displayed in Table 12. Regarding the relationships among the cognitive attributes, most of the relationships specified in the three hierarchies were confirmed. However, it was found that attribute A2, rather than being independent of attribute A3a, should have it as its prerequisite. Moreover, it was found that attribute A8, which was discovered from the verbal reports, also has attribute A3a as its prerequisite. According to verbal report results, the three initial hierarchies presented in Chapter III were revised and are displayed in Figure 5.

In the second part of the cognitive analysis, the *HCI* for the three revised hierarchies were calculated using two random samples of 2000 examinees who wrote the March 2005 administration of the SAT. From the *HCI* results, it was found that Hierarchy 2 improved model-data fit considerably but Hierarchy 3 did so only negligibly. Given the fact that Hierarchy 2 involves less attributes and inter-attribute connections and, thus, is more parsimonious than Hierarchy 3, Hierarchy 2 was regarded as the best among the three hierarchies and, thus, was used for analyzing student response data.

The present chapter provided the results from the psychometric analysis, in which student response data were analyzed. The $Q_r$ matrix, attribute pattern and expected response pattern matrices were generated according to Hierarchy 2. Then, attribute probabilities were calculated for a random sample of 15 students. It was found that, in general, examinees with higher total scores mastered more attributes. However, examinees with the same total score may have mastered different combinations of cognitive attributes. Such results point to the necessity for not only reporting an overall score to examinees, as is done in most operational testing programs, but also detailed information about what attributes were mastered and what not by the examinees. Next, attribute reliabilities and the SEM were estimated for the nine attributes in Hierarchy 2. Low reliabilities and large values of SEM were obtained. However, when the test length was increased, the attribute reliabilities were improved. Finally, sample score reports were created for four examinees to demonstrate how the AHM can be used to report test scores and provide diagnostic feedback.

# CHAPTER VI: DISCUSSION AND CONCLUSIONS

The chapter is organized in six sections. In the first section, the research questions and a brief description of the methods used in the present study are presented. A summary and discussion of the key findings are presented in the second section. The limitations of the study are presented in the third section. The conclusions of the study are drawn in the fourth section. The fifth and sixth sections present implications for educational practice and recommendations for future research, respectively.

## Summary of Research Questions and Methods

The present study addressed one important problem in educational testing, the disjunction between cognitive psychology and testing practice. Specifically, in many large-scale testing programs, the importance of understanding the psychology underlying students' test performance is downplayed relative to the emphasis placed on using statistical models and psychometric techniques for scaling and scoring examinees' performance (Glaser, 2000; Leighton et al., 2004; National Research Council, 2001; Nichols, 1994). One of the consequences of this situation is that most large-scale tests typically yield very limited information about why some students perform poorly or how instructional conditions can be modified to improve teaching and learning (National Research Council, 2001). To correct this situation, efforts must be made to understand the cognitive attributes underlying student performance on tests. The present study served this purpose and investigated the cognitive attributes underlying student performance on the SAT Critical Reading subtest based on the cognitive model proposed by VanderVeen et al. (2007).

The present study is also among the first applications of the AHM to data from an operational testing program. The AHM was used because it represents an effort towards the integration of cognitive psychology into educational testing practice. It starts with the specification of attribute hierarchies, which include the cognitive attributes measured by test items and their relationships. Then, according to students' performance on the test items, diagnostic feedback about their cognitive strengths and weaknesses can be provided to them.

Two research questions were addressed in the present study:

*Research Question 1: What cognitive attributes in the VanderVeen et al. (2007) model are involved in answering the SAT critical reading items, what additional attributes are involved, and how are these cognitive attributes related to each other?*

*Research Question 2: How can the AHM be used to analyze student response data, report student scores, and provide diagnostic feedback?*

The study was conducted in two stages. In the first stage, the *cognitive* features of the SAT critical reading items were analyzed. Three attribute hierarchies for the SAT Critical Reading subtest were initially identified. Then, a verbal report study was conducted to validate and, when required, revise the three initial attribute hierarchies. Next, an *HCI* analysis was conducted to further validate the three revised hierarchies on large samples and to select the best-fitting hierarchy for analyzing examinee response data. Based on the *HCI* results, Hierarchy 2, due to its model-data fit and parsimony, was selected for analyzing examinee response data.

The second stage was a *psychometric* analysis of the response data from 15 examinees on the March 2005 administration of the SAT was conducted. Probabilities for the nine attributes in Hierarchy 2 were calculated for the 15 examinees. Reliabilities for the nine attributes were also estimated based on the response data of two samples of 2000 examinees. According to the attribute probabilities and reliabilities, descriptive score reports were composed for four examinees to demonstrate the facility of the AHM in providing diagnostic feedback.

## Findings

### *Research Question 1*

Three hierarchies were initially specified by reviewing selected literature in reading and the research results related to the SAT Critical Reading subtest (Figure 4). These hierarchies reflected different grain sizes of cognitive models of task performance. The grain size of Hierarchy 1 was coarse because the attributes in this hierarchy collapsed across multiple cognitive processes. As hierarchies with such coarse attributes may affect the precision of cognitive feedback if used for diagnostic purpose, Hierarchy 2, which had a smaller grain size than Hierarchy 1, was proposed with attributes A2, A3, and A4 in Hierarchy 1 redefined or decomposed. A problem with both Hierarchy 1 and

Hierarchy 2 was that they did not take into account the effect of text factor. Therefore, attributes A6 and A7, which were used to reflect text difficulty, were added to form Hierarchy 3.

Eleven cognitive attributes were included in the three initial hierarchies. These attributes are summarized in Table 5. Among the attributes, A3 was considered a composite of A3a and A3b, and A4 a composite of A4a and A4b, respectively. The attributes in the three initial hierarchies were organized according to three non-mutually exclusive principles: the cognitive demand involved, the size of information unit to be processed, and the amount of inferences involved. For example, A3a and A4a were regarded as lower-order attributes than A3b and A4b because the latter two attributes involve more inference-making. On the other hand, A3a was considered as a lower order attribute than A4a because A4a involves the processing of larger sections of text while A3a only involves the processing of individual sentences. However, because these principles of organizing cognitive attributes have not been empirically validated, they were investigated using data from student verbal reports.

The results from the verbal report study indicated that, except for attribute A1, which was assumed to have been mastered by all target examinees of the SAT, all remaining attributes in the three hierarchies were used, to differing degrees, by the participants. Moreover, two more attributes were discovered from the verbal reports. These two attributes are (A8) "using rhetorical knowledge" and (A9) "evaluating response options." Attribute A8 was originally proposed in VanderVeen's (2004) seven-attribute model, but was deleted from the model after multiple rounds of expert rating in his study. However, the results from students' verbal reports indicate that this attribute should be included. This decision is justified because, although the attribute was used for only a small number of items, it was used by almost all students for these items (e.g., Item 2 and Item 15), indicating its importance to the correct solution of these items. In comparison to attribute A8, attribute A9 was a more frequently used attribute. The discovery of this attribute, which concerns the response options in the items, echoes the findings of many other researchers (Embretson & Wetzel, 1987; Kirsch & Mosenthal, 1990; Mosenthal & Kirsch, 1991; Hannon & Daneman, 2001). For example, Embretson and Wetzel (1987) applied a cognitive processing model of reading comprehension to account for variance in item difficulty on the Armed Services Vocational Aptitude Battery (ASVAB). The authors analyzed the test items and identified two general processes underlying performance on multiple-choice reading comprehension questions based on reading a passage: text representation and test item response decisions. In other words, attribute A9 has been regarded as a key process in reading tests when the items are in multiple-choice form. As a result, although this attribute was given only brief mention in the specification of the SAT Critical Reading subtest, it should not be ignored, especially when the test items are in multiple-choice form.

The results from the verbal reports also revealed the hierarchical relationships among the attributes. In Hierarchy 1, attribute A3 was found to be the prerequisite of attributes A2, A4, and A5. In Hierarchies 2 and 3, attribute A3a was found to be the prerequisite of attributes A2, A3b, A4a, A5, and A8, and attribute A4a was the prerequisite of A4b. From these findings, it can be said that comprehension of text starts from the sentence level and then goes to the text level. Moreover, literal comprehension of text is essential for inferential comprehension. As processing text and drawing inferences are cognitively more demanding than processing sentences and literal comprehension, the three principles used to specify the attribute hierarchies, cognitive demands involved, the amount of inferences involved (van Dijk & Kintsch, 1983; Urquhart & Weir, 1998), and size of the information unit to be processed (Urquhart & Weir, 1998), were confirmed. However, it is still not clear how these principles interact. For example, it is not known how attributes A3b and A4a, which both require attribute A3a as their prerequisite, relate to each other and which, if either, attribute is a higher-order attribute.

As only eight students participated in the verbal report study and their performance may not reflect the performance of large samples of examinees, an *HCI* analysis was conducted on larger samples to further validate the cognitive attributes and their relationships. The *HCI* analysis evaluated the model-data fit of the revised hierarchies and provided another line of validation evidence in addition to the findings from the verbal report study. The *HCI* analysis was conducted on two random samples of 2000 examinees. The mean *HCI* values for the three revised hierarchies are summarized in Table 15. Because the *HCI* ranges from a maximum misfit of -1 to a maximum fit of 1, these values indicate moderate model-data fit for all three hierarchies. Moreover, as the hierarchies get more fine-grained from Hierarchy 1 to Hierarchy 3, better model-data fit results. However, relative to the improvement of model-data fit from Hierarchy 1 to Hierarchy 2, the improvement was negligible from Hierarchy 2 to Hierarchy 3. In other words, Hierarchy 2 improved model-data fit considerably but Hierarchy 3 did so only negligibly. Given the fact that Hierarchy 2 involved less attributes and inter-attribute connections and, thus, was more parsimonious than Hierarchy 3, Hierarchy 2 was regarded as the best hierarchy and, thus, was used for analyzing student response data.

With the selection of Hierarchy 2, attributes A6 and A7 were not used in describing student performance on the SAT critical reading subtest. However, this does not mean that attributes A6 and A7 were not important in describing students' reading ability. Rather, the exclusion of these two attributes was done on the basis of parsimony as their inclusion improved model-data fit only marginally. It is also possible that, in a high-stakes test such as the SAT, the effect of text difficulty has been controlled in some way and, therefore, attributes A6 and A7 did not contribute appreciably to students' test performances. To further investigate why the two attributes did not improve the model-data fit more, future studies should examine the interaction between text difficulty and characteristics of test items.

*Summary.* In the cognitive analysis, both verbal report study and the *HCI* were used to validate the cognitive attributes and their relationships. Although the results of the verbal report study supported all attributes in the three revised hierarchies, the *HCI* results suggested that only the attributes involved in revised Hierarchy 2 (A1, A2, A3a, A3b, A4a, A4b, A5, A8, A9) were most representative of those used by the SAT test-takers. The results of the verbal report study also supported the hierarchical relationships among cognitive attributes. Specifically, attributes involving text-level comprehension are higher-order than attributes involving only sentence-level comprehension; attributes involving inferential comprehension are higher-order than attributes involving only literal comprehension.

*Research Question 2*

In the current study, attribute probabilities were calculated for a sample of 15 examinees who took the March 2005 administration of the SAT. A strong relationship was detected between examinees' total scores and their attribute mastery: the higher the total score, the more attributes an examinee possesses. This finding gives some legitimacy for reporting a total score to students, as is done in most operational testing programs. However, some of the examinees who received a lower total score possessed more cognitive attributes than some of the examinees who received a higher total score.

Moreover, examinees with the same total scores possessed different cognitive attributes. In other words, providing examinees with a single total score may, sometimes, misrepresent students' cognitive skill profiles.

With attribute probabilities, information about examinees' attribute mastery levels was obtained. Attribute reliability, on the other hand, provides a means to judge the consistency of decisions on students' attribute mastery. In the current study, the reliability estimates for the nine attributes in Hierarchy 2, were, in general, low, with the highest reliability estimates at 0.74 and lowest at 0.20. With such low values, the reliability of the diagnostic feedback would be called into question. It was found that the values of attribute reliabilities were closely related to the number of items measuring each attribute: the more items that measured an attribute, the higher the attribute reliability. Therefore, to increase the reliability estimates of attributes, one simple way would be to increase the length of the test by adding parallel items. For example, when two more parallel item sets were added to the 20-item test, eight out of the nine attributes had reliability estimates of over 0.5. Another possible solution lies in the precision of the cognitive attributes and the method of item development. As suggested in the previous chapter, because the cognitive attributes used in the current study were based mainly on the VanderVeen et al. (2007) model, which had coarse grain size, and were validated from the verbal reports of only eight students, some of the attributes might not be defined precisely. Thus, cognitive attributes having sounder theoretical and empirical basis could be used and test items could be written according to an *a priori* attribute hierarchy, rather than by retrofitting the hierarchy to existing test items, so that the items measure accurately the attributes in the hierarchy.

Although the reliability estimates could be improved by increasing the length of the test, the fact that the cognitive attributes in the AHM are hierarchically related made it difficult for these attributes to achieve the same level of reliability. This result occurs because, in the AHM, if a test item measures a certain cognitive attribute, then it will also measure the prerequisite of this attribute. In other words, an attribute will generally be measured by fewer items than its prerequisite attribute will and thus will get a lower reliability estimate. Hence, further study in the future is needed to handle this situation.

With information on examinees' attribute mastery levels and the reliability estimates, the AHM provides individualized score reports to the examinees. The score reports produced from the AHM not only have a total score, but also provide detailed information about examinees' attribute mastery and their cognitive strengths and weaknesses. These score reports also reflect examinees' proficiency level in a content domain. In the context of reading tests, examinees' level of text comprehension (e.g., literal vs. inferential levels, sentence vs. text levels) can be determined according to the score reports, as is done for Examinees 8 and 12 in the previous chapter. With the information about examinees' cognitive strengths, the score reports also suggest to examinees where to direct their efforts for improvement in learning.

## Limitations of the Study

One key stage of the current study is the cognitive analysis, which was conducted to identify the attributes measured by the SAT critical reading items. For practicality issues, first-year undergraduates at a Canadian university were recruited for a verbal report study to obtain the attributes measured by the SAT critical reading items. Although efforts were made to ensure the representativeness of the participants, they may not be a perfect sample for the SAT test-taker population for two reasons. First, their reading ability may be more proficient (see Table 7, p.72) than the general student who takes the SAT because only students who are academically competitive could be admitted to the university while all college-bound students, regardless of their academic ability, are entitled to take the SAT. As a result, less cognitive attributes might have been recovered than if a group with more heterogeneous reading abilities were used. Second, although the SAT is internationally administered, a majority of the test-takers are Americans. For example, in 2005, about 95% test-takers were U.S. citizens or permanent resident (The College Board, 2005). However, in the current study, Canadian students were used for the verbal report study. Due to culture differences and topic familiarity with the reading passages, their performance might have been different from the performance of their American counterparts. In both cases, the validity of attribute hierarchies might be affected.

The second limitation of the study concerns the test items used in the study. In the

AHM, it is ideal to use test items developed according to the specification of the attribute hierarchies. However, hierarchies were developed from scratch by reviewing related literature, validated using student verbal reports, and then retrofitted to existing test items. Despite these procedures, the *HCI* results indicated only moderate model-data fit. In other words, there were still considerable misfits between the hierarchies and the attributes used by examinees when they solved the test items. Due to these misfits, the cognitive attributes identified in the study and the attribute probability results must be interpreted with caution.

The third limitation of the study concerns the attribute reliabilities. In the current study, the reliabilities for most attributes in Hierarchy 2 were low. Even after lengthening the test by three times, the reliability for some attributes remained low. One possible reason for the low reliability estimates may be that the items were not developed according to predetermined attribute hierarchies and that the attributes measured by these items were determined post-hoc using student verbal reports. The low attribute reliabilities made it difficult for the AHM to yield precise score reports. For example, in the sample score report for Examinee 8, it was hard to decide whether the examinee has mastered attributes A2 and A3a because the confidence intervals for these attributes cross different mastery levels.

## Conclusions

To conclude, the AHM draws on information from students, assessment specialists, and cognitive psychologists and, in turn, yielded information to benefit these three groups. In other words, the AHM serves to unify the different components of educational assessment, including instruction, cognitive theory, and assessment. However, it must be noted that the links among these different components are weak, at the current stage. For example, few large-scale tests are developed with the guidance of an explicit cognitive model. Therefore, these links must be strengthened in the future. The present study provides an illustration of how these links can begin to be formed.

## Implications for Educational Practice

The findings in this study have implications for learning and instruction, construct validation, and development of cognitive theories. First, the AHM provides detailed information about examinees' cognitive strengths and weaknesses. As this information is derived from validated cognitive models, it allows the examinees to focus their efforts on remedying attributes they have not mastered. This information also allows the teachers to tailor their instruction to the specific needs of the students. In other words, the AHM may be instrumental in improving teaching and learning.

Second, the AHM provides useful information for construct validation. The AHM requires a cognitive analysis, which can include a verbal report study, to probe the specific cognitive skills measured by each item. For example, in the present study, two attributes not included in the initial hierarchies were found: using rhetorical knowledge (A8) and evaluating response options (A9). If *HCI* results indicate good fit between the attribute hierarchies and the observed response data, then evidence about the construct validity of the test can be obtained.

Third, the AHM provides information for researchers to further investigate the nature of reading ability and, thereby, develop cognitive theories in reading. For example, once examinees' strengths and weaknesses in reading are reliably and validly identified, remediation techniques specific to each cognitive attribute may be developed. Then, a pretest-posttest experimental study could be conducted to produce contrasting subscores on the component attributes and the total score. By regressing the posttest subscores and total score on the pretest scores, the treatment effect can be determined. The relative contributions of treatments targeting the component attributes to the total score can be contrasted, and inferences drawn about which component attributes are most efficient in their contribution to reading ability (VanderVeen et al., 2007).

## Directions for Future Research

The current study suggests at least two directions for future research. One line of future research is how to apply the cognitive information produced from the AHM to teaching and learning practice. Although the AHM produces detailed information about examinees' cognitive strengths and weaknesses, it does not suggest specific procedures for examinees to make improvement on non-mastered attributes. For example, if an examinee has been diagnosed as not possessing the attribute of analyzing author's purpose, goals, and strategies, no suggestions were made to the examinees about how to improve this attribute. Therefore, future research can be directed towards the development of specific procedures which link the cognitive information produced from the AHM with teaching and learning practice. With the AHM results and these procedures, teachers and students will not only know what attributes should be improved, but also how to improve these attributes. In so doing, cognition, testing, and teaching and learning will be integrated.

The second line of future research can be directed towards exploring the potential of the AHM and its application in testing practice. In the current study, attribute hierarchies were retrofitted to existing test items and validated using student verbal reports. A different sample of students, more representative of the SAT test-taker population, can be recruited to further validate these hierarchies. More precise attribute hierarchies can be developed to guide test construction. Then, new test items can be constructed based on these attribute hierarchies and administered to a larger sample of students. The response data from these students could then be used to examine whether new attributes should be included and whether the model-data fit could be improved. Studies should also be conducted to find out the optimal length of a test which produces satisfactory reliability estimates for all attributes in a hierarchy. This outcome may contribute to the application of the AHM in more practical testing situations.

## References

Abu-Rabia, S. (1995). Learning to read in Arabic: Reading, syntactic, orthographic and working memory skills in normally achieving and poor Arabic readers. *Reading Psychology, 16*, 351-394.

Alderson, J. C. (1990). Testing reading comprehension skills (part one). *Reading in a Foreign Language, 6*, 425-438.

Alderson, J. C. (2000). *Assessing reading.* Cambridge, UK: Cambridge University Press.

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language, 5*, 253-270.

Andrich, D., & Godfrey, J. R. (1978-1979). Hierarchies in the skills of Davis' Reading Comprehension Test, Form D: An empirical investigation using a latent trait model. *Reading Research Quarterly, 14*, 182-200.

Ayala, C. C., Shavelson, R. J., Yin, Y., & Schultz, S. E. (2002). Reasoning dimensions underlying science achievement: The case of performance assessment. *Educational Assessment, 8*, 101-121.

Baram-Tsabari, A., & Yarden, A. (2005). Text genre as a factor in the formation of scientific literacy. *Journal of Research in Science Teaching, 42*, 403-428.

Bejar, I. I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement, 21*, 175-189.

Block, E. L. (1992). See how they read: Comprehension monitoring of L1 and L2 readers. *TESOL Quarterly, 26*, 319-343.

Bloom, B. S. (1956). *Taxonomy of educational objectives. Book I cognitive domain.* London: Longman.

Bowyer-Crane, C., & Snowling, M. J. (2005). Assessing children's inference generation: What do tests of reading comprehension measure? *British Journal of Educational Psychology, 75*, 189-201.

Burton, N., Welsh, C., Kostin, I., & van Essen, T. (2003). *Toward a definition of verbal reasoning in higher education.* New York: College Examination Board.

Butcher, K. R., & Kintsch, W. (2003). Text comprehension and discourse processing. In A. F. Healy & R. W. Proctor (Eds.), *Handbook of psychology: Experimental psychology* (pp. 575-595). New York: John Wiley & Sons.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 95*, 31-42.

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology, 96*, 671-681.

Chall, J. S., & Stahl, S. (2004). Reading. In *Microsoft Encarta encyclopedia standard 2004.* Redmond, WA: Microsoft Corporation.

Clapham, C. (1998). The effect of language proficiency and background knowledge on EAP students' reading comprehension. In A. Kunnan (Ed.), *Validation in language assessment: Selected papers from 17th Language Testing Research Colloquium, Long Beach* (pp. 141-168). Mahwah, NJ: Erlbaum.

Cui, Y., Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2006, April). *The Hierarchical Consistency Index: A person-fit statistic for the Attribute Hierarchy Method.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Daneman, M. (1988). Word knowledge and reading skill. In M. Daneman, G. E. MacKinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 6, pp. 145-177). New York: Academic Press.

Ehrlich, M., Remond, M., & Tardieu, H. (1999). Processing of anaphoric devices in young skilled and less skilled comprehenders: Differences in metacognitive monitoring. *Reading and Writing: An Interdisciplinary Journal, 11*, 29-63.

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

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

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analyses: Verbal reports as data* (revised ed.). Cambridge, MA: MIT Press.

Gierl, M. J., Bisanz, J., Bisanz, G. L., Boughton, K. A., & Khaliq, S. N. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*(2), 26-36.

Gierl, M. J., Cui, Y., & Hunka, S. M. (in press). The attribute hierarchy method for cognitive assessment: Technical developments. *Applied Measurement in Education.*

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement:*

*Issues and Practice, 19*(3), 34-44.

Gierl, M. J., Wang, C., & Zhou, J. (2006, July). *Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT*©. New York: College Examination Board.

Glaser, R. (Ed.). (2000). *Advances in instructional psychology: Educational design and cognitive science* (Vol. 5). Mahwah, NJ: Erlbaum.

Gorin, J. S., Embretson, S. E., & Sheehan, K. M. (2002, April). *Cognitive and psychometric modeling of text-based reading comprehension GRE items: Building tests so we know what students know.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48,* 163-189.

Haberlandt, K. (1988). Component processes in reading comprehension. In M. Daneman, G. E. MacKinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 6, pp. 67-108). New York: Academic Press.

Hillocks, G. J., & Ludlow, L. H. (1984). A taxonomy of skills in reading and interpreting fiction. *American Educational Research Journal, 21,* 7-24.

Hirsch, E. D. J. (2003). Reading comprehension requires knowledge--of words and the world. *American Educator, 27,* 10-48.

Hudson, T. (1996). *Assessing second language academic reading from a communicative competence perspective: Relevance for TOEFL 2000.* Princeton, NJ: Educational Testing Service.

Huff, K. (2004). *A practical application of evidence centered designed principles:*

*Coding items for skills.* Paper presented at the 2004 annual meeting of the National Council on Measurement in Education, San Diego, CA.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition, 16,* 1-61.

Keenan, J. M., Potts, G. R., Golding, J. M., & Jennings, T. M. (1990). Which elaborative inferences are drawn during reading? A question of methodologies. In D. A. Balota, G. B. Flores d'Arcais & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 377-402). Hillsdale, NJ: Erlbaum.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95,* 163-182.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49,* 294-303.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, MA: Cambridge University Press.

Kolln, M. (1999). *Rhetorical grammar: Grammatical choices, rhetorical effects.* Needham Heights, MA: Allyn & Bacon.

Leighton, J. P. (2004). Avoid misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*(4), 6-15.

Leighton, J. P., & Gierl, M. J. (2007). Identifying and evaluating cognitive models in educational measurement. *Educational Measurement: Issues and Practice, 26*(2), 3-16.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of*

*Educational Measurement, 41*, 205-236.

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

Long, D. L., Oppy, B. J., & Seely, M. R. (1997). Individual differences in readers' sentence- and text-level representations. *Journal of Memory and Language, 36*, 129-145.

Ludlow, L. H., & Hillocks, G. J. (1985). Psychometric considerations in the analysis of reading skill hierarchies. *Journal of Experimental Education, 54*, 15-21.

Lunzer, E., Waite, M., & Dolan, T. (1979). Comprehension and comprehension tests. In E. Lunzer & K. Gardner (Eds.), *The effective use of reading* (pp. 37-71). London: Heinemann Educational.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Erlbaum.

Nassaji, H. (2003). Higher-level and lower-level text processing skills in advanced ESL reading comprehension. *The Modern Language Journal, 87*, 261-276.

Nation, K., & Snowling, M. J. (2000). Factors influencing syntactic awareness skills in normal readers and poor comprehenders. *Applied Psycholinguistics, 21*, 229-241.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

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

Oakhill, J., Cain, K., & Bryant, P. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes, 18*, 443-468.

Perfetti, C. A. (1985). *Reading ability.* London: Oxford University Press.

Perfetti, C. A. (1988). Verbal efficiency in reading ability. In M. Daneman, G. E. MacKinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 6, pp. 109-144). New York: Academic Press.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing, 20*, 26-56.

Phillips, L. M. (1987). *Inference Strategies in Reading Comprehension* (No. 410). Urbana Champaign: Illinois University.

Radford, A. (2004). *English syntax: An introduction.* Cambridge, UK: Cambridge University Press.

Rumelhart, D. E. (1994). Toward an interactive model of reading. In R. B. Ruddell, M. R. Ruddell & H. Singer (Eds.), *Theoretical models and processes of reading* (4th ed., pp. 864-895). Newark, DE: International Reading Association.

Salinger, T. (2003). Helping older, struggling readers. *Preventing School Failure, 47*(2), 79-85.

Sanford, A. J. (1990). On the nature of text-driven inference. In D. A. Balota, G. B. Flores d'Arcais & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 515-535). Hillsdale, NJ: Erlbaum.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*, 333-352.

Sheehan, K. M., & Ginther, A. (2001, April). *What do multiple-choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on a standardized test of reading comprehension skill.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Smith, F. (2004). *Understanding reading: A psycholinguistic analysis of reading and learning to read.* Mahwah, NJ: Erlbaum.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-332). New York: Macmillan.

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York: The Guilford Press.

Swanson, H. L., & Ashbaker, M. H. (2000). Working memory, short-term memory, speech rate, word recognition and reading comprehension in learning disabled readers: Does the executive system have a role? *Intelligence, 28*, 1-30.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis. In P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.

The College Board. (2005). *2005 College-bound seniors: Total group profile report.* New York: The College Board.

Thompson, G. (2004). *Introducing functional grammar.* New York: Arnold.

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. London: Longman.

van den Broek, P. (1990). The causal inference maker: Towards a process model of inference generation in text comprehension. In D. A. Balota, G. B. Flores d'Arcais & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 423-445). Hillsdale, NJ: Erlbaum.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

VanderVeen, A. (2004, April). *Toward a construct of critical reading for the new SAT.* Paper presented at the 2004 annual meeting of the National Council on Measurement in Education, San Diego, CA.

VanderVeen, A., Huff, K., Gierl, M. J., McNamara, D. S., Louwerse, M., & Graesser, A. C. (2007). Developing and validating instructionally relevant reading competency profiles measured by the Critical Reading sections of the SAT. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 137-171). Mahwah, NJ: Erlbaum.

Veenman, M. V. J., & Beishuizen, J. J. (2004). Intellectual and metacognitive skills of novices while studying texts under conditions of text difficulty and time constraint. *Learning and Instruction, 14,* 621-640.

Vos, S. H., Gunter, T. C., Schriefers, H., & Friederici, A. D. (2001). Syntactic parsing and working memory: The effects of syntactic complexity, reading span, and concurrent load. *Language & Cognitive Processes, 16,* 65-103.

Wang, C., & Gierl, M. J. (2007, April). *Investigating the cognitive processes underlying student performance on the SAT® Critical Reading subtest: An application of the attribute hierarchy method.* Paper presented at the annual meeting of the National

Council on Measurement in Education, Chicago, IL.

Wang, C., Gierl, M. J., & Leighton, J. P. (2006, April). *Investigating the cognitive attributes underlying student performance on a foreign language reading test: An application of the attribute hierarchy method.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Watermann, R., & Klieme, E. (2002). Reporting results of large-scale assessment in psychologically and educationally meaningful terms: Construct validation and proficiency scaling in TIMSS. *European Journal of Psychological Assessment, 18*, 190-203.