University of Alberta

Estimating Attribute-Based Reliability in Cognitive Diagnostic Assessment

by

Jiawen Zhou

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

Spring 2010

Edmonton, Alberta

Examining Committee

Dr. Mark J. Gierl, Educational Psychology

Dr. Jacqueline P. Leighton, Educational Psychology

Dr. Ying Cui, Educational Psychology

Dr. Michael Carbonaro, Educational Psychology

Dr. Rebecca Gokiert, Extension

Dr. Eunice E. Jang, Education, University of Toronto

To

Brian Xiaoqi Shen

Abstract

Cognitive diagnostic assessment (CDA) is a testing format that employs a cognitive model to, first, develop or identify items measuring specific knowledge and skills and, then, use this model to direct psychometric analyses of examinees' item response patterns to promote diagnostic inferences. The attribute hierarchy method (AHM, Leighton, Gierl, & Hunka, 2004) is a psychometric procedure for classifying examinees' test item responses into a set of structured attribute patterns associated with different components from a cognitive model of task performance. Attribute reliability is a fundamental concept in cognitive diagnostic assessment because it refers to the consistency of the decisions made in diagnostic test about examinees' mastery of specific attributes. In this study, an adapted attribute-based reliability estimate was evaluated in comparison of the standard Cronbach's alpha using simulated data. Factors expected to influence attribute reliability estimates, including test length, sample size, model structure, and model-data fit level, were also studied. Results of this study revealed that the performances of the two attribute-based reliability estimation indices are comparable; however, the adapted index is conceptually more meaningful. Test length, model structure, and model-data fit were shown to impact attribute reliability estimates differentially. Implications to researchers and practitioners were given based on the simulation results. Limitations of the present study and future directions were also discussed.

Acknowledgement

I would like to acknowledge a number of people who have helped me through my doctoral program. I would like to first thank my supervisor, Dr. Mark Gierl. Mark has been a great mentor and provided professional opportunities for me to grow up as a qualified graduate. He included me in his research projects, guided me on writing research reports, and helped me gain valuable practical experiences. Without his supportive mentorship, I would not have been able to achieve goals as professional development. I would also like to express my gratitude to Dr. Jacqueline Leighton and Dr. Ying Cui who were in my supervisory committee. Jacqueline's classes deepened my understanding of the link between measurement and cognitive psychology. Ying helped me on mathematica skills which were employed for data simulation and analysis in my study. I will always be grateful for their kindness and support.

My sincere thanks go to Dr. Mike Carbonaro, Dr. Rebecca Gokiert, and Dr. Eunice Jang for serving as my committee members and providing valuable comments. My sincere thanks also go to my graduate school professors, especially Dr. Todd Rogers, who guided me to the educational measurement area. I would also like to thank my colleagues in the Centre for Research in Applied Measurement and Evaluation, particularly Adele Tan, Changjiang Wang, Ken Cor, Cecilia Alves, Xian Wang, and Oksana Babenko, for their helpful

discussion. Special gratitude is extended to my friends, Jing Wu and Bihua Xiang, for their encouragement and company that made my program more memorable.

Finally, I would like to thank my family for their support throughout my pursuit of a PhD. I am grateful the motivation from my husband, Gangxiang Shen, and my son, Brian Xiaoqi Shen. I am also grateful to my parents, Longxing Zhou and Jiaqian Lin, and my brother, Jianwen Zhou, for their understanding and unconditional support. Every step I made would not have been possible without them.

Table of contents

List of Tables

List of Figures

Chapter I: Introduction

Context of the Study

The development and progression of modern technology is infiltrating all research areas and, in turn, generating more demands for multidisciplinary and interdisciplinary research. Educational measurement is no exception. One key example of interdisciplinary research in educational measurement is the fusion of psychological principles with measurement practices (Snow & Lohman, 1989; Pelleino, Baxter, & Glaser, 1999, Mislevy, 2006).

Cognitive psychology reflects a psychological perspective that focuses on the realms of human internal mental processes such as perception, attention, thought, memory, and problem solving. The psychological perspective together with the substantive perspective determine the nature of elements in test performance and, thus, directly affects practice, such as test construction, test score interpretation, and the diagnostic feedback provided to examinees (Mislevy, 2006). In test construction, a cognitive theory of how examinees develop competence in a content domain yields clues about the types of item features that would elicit psychological evidence for claims of examinees' thinking processes and cognitive proficiencies (National Research Council, 2001). Accordingly, test users would be able to make valid test score interpretations in relation to examinees' cognitive strengths and weaknesses based on a cognitive theory. Then, based on the test score interpretation drawn from examinee

performance on the test items, cognitively diagnostic feedback could be provided to the examinees and their teachers to enhance learning and instruction.

Although cognitive psychology is exerting its influence on testing practice, its impact to-date has been overlooked (Leighton, Gierl, & Hunka, 2004). The significance of understanding the psychology underlying examinees' performance has been neglected in most contemporary large-scale testing programs. In contrast, much more attention has been paid to statistical models and psychometric techniques for scaling and scoring examinees' performance (Glaser, 2000; Leighton et al., 2004; Nichols, 1994). Consequently, information provided by most current large-scale tests for teachers, examinees, and parents is very limited about why some examinees perform poorly and how instructional conditions can be modified to improve teaching and learning (National Research Council, 2001).

The call for the integration of cognitive psychology and educational measurement began in earnest two decades ago (Snow & Lohman, 1989). Increasingly, researchers and practitioners are calling for the union of cognitive psychology and educational measurement to enhance learning and instruction (Embretson, 1998; Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001; Gierl, Cui, & Hunka, 2008; Leighton et al., 2004; Gierl & Zhou, 2008; Mislevy, 2006; Mislevy & Riconscente, 2006; National Research Council, 2001; Nichols, 1994; Sheehan, 1997; and Tatsuoka, 1995). Pellegrino et al. (1999) claimed that:

> *…it is the pattern of performance over a set of items or tasks explicitly constructed to discriminate between alternative profiles of knowledge that should be the focus of assessment. The latter can be used to determine the level of a given student's understanding and competence within a subject-matter domain. Such information is interpretative and diagnostic, highly informative, and potentially prescriptive*. (p.335)

That is, a test has the potential for identifying examinees' problem-solving strengths and weaknesses when it is created from cognitive models which provide a contemporary representation of the knowledge structures and processing skills that are believed to underlie conceptual understanding in a particular domain. Results of the assessment could also be integrated into the teaching and learning process because this form of assessment supports specific inferences about the examinees' problem-solving skills that could be linked with specific instructional methods designed to improve these cognitive skills.

The incremental pressure to adapt assessments to be informative about examinees' cognitive strengths and weaknesses promotes changes in educational measurement. The movement of integrating cognitive psychology and educational measurement has led to innovations and change in relevant research areas such as the development of cognitive diagnostic assessment (as discussed in Leighton & Gierl, 2007a).

Cognitive Models and Cognitive Diagnostic Assessment

Human problem-solving skills represent knowledge structures and information-processing performance. The knowledge structure includes declarative and procedural information while the processing skills indicate the management of the transformation of the information (Lohman, 2000). Cognitive diagnostic assessment (CDA) is a form of assessment designed to measure examinees' specific knowledge structures and processing skills and to provide information about examinees' cognitive strengths and weaknesses, particularly when the assessments are created from cognitive models that provide a contemporary representation of the knowledge structures and processing skills that are believed to underlie conceptual understanding in a particular domain (Gierl, Cui, & Zhou, 2009). Specifically, in order to make inferences about human problem solving, a cognitive model that represents an explicit interpretative framework of human knowledge structures and processing skills is first employed to develop or identify items that measure specific knowledge and skills and then used to direct the psychometric analyses of the examinees' item response patterns to promote diagnostic inferences.

In educational measurement, a cognitive model refers to a "simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills examinees at different levels of learning have acquired and to facilitate the explanation and prediction of

students' performance" (Leighton & Gierl, 2007a, p. 6). A cognitive model can be used to identify examinees' cognitive proficiencies because it provides the framework necessary to link cognitively-based inferences with specific problem-solving performance. Consequently, the strength of developing test items and analyzing testing data according to a cognitive model stems from the detailed information that can be obtained about the knowledge structures and processing skills that produce examinees' test scores. In other words, CDAs have the potential for identifying examinees' problem-solving strengths and weaknesses. CDAs can also help pinpoint why examinees perform as they do and how examinees' opportunities to learn can be maximized because the information is specifically tailored for each examinee to reflect the examinee's thinking and high-order cognitive processes associated with meaningful learning (Leighton & Gierl, 2007a). CDAs provide one way for linking theories of cognition and learning with instruction because they support specific inferences about the examinees' problem-solving strengths and weaknesses that, in turn, could be linked with the most effective and timely instructional methods designed to improve these cognitive skills. With such information, erroneous strategies can be corrected and misconceptions can be altered.

According to Messick (1989), substantively understanding test performances in terms of the mental processes examinees use to answer and/or solve test items is a core feature of construct validity theory. Particularly, he emphasized the verification of the domain processes to be revealed in

assessment tasks. As discussed above, CDA is an assessment format that focuses on highlighting examinees' mental processes as they are engaged in problem solving and then using this information for improving examinees' opportunity to learn (Leighton & Gierl, 2007a). Therefore, CDAs can help enhance test construct validity in terms of understanding examinees' mental processes.

Attribute Hierarchy Method

In an attempt to uncover the diagnostic information that may be embedded in examinees' item response data and address the challenge posed by Pellegrino et al. (1999), psychometric procedures have been developed to support test-score inference based on cognitive models of task performance. These cognitive diagnostic models contain parameters that link item features to examinees' response patterns so inferences about declarative, procedural, and strategic knowledge can be made. Some early examples include the *linear logistic test model* (Fischer, 1973), the *multicomponent latent trait model* (Embretson, 1980), and the *rule space model* (Tatsuoka, 1983). More recent examples include the *unified model* (Dibello, Stout, & Roussos, 1995), the *DINA models* (de la Torre & Douglas, 2004), and the *NIDA models* (Junker & Sijtsma, 2001).

In 2004, Leighton, Gierl, and Hunka introduced a procedure for CDA called the *attribute hierarchy method* (AHM). The AHM, a method that evolved from Tatsuoka's rule space model (see Gierl, 2007), refers to a cognitively-based

psychometric method which classifies examinees' test item responses into structured attribute patterns according to a cognitive model of task performance, with the assumption that test performance is associated with a specific set of hierarchically-organized cognitive components called attributes. An attribute is a description of declarative or procedural knowledge required to perform a task in a specific domain. These attributes are structured using a hierarchy so the ordering of the cognitive skills is specified. As a result, the *attribute hierarchy* serves as an explicit construct-centered cognitive model because it represents the psychological ordering among the cognitive attributes required to solve test items. This model, in turn, provides an explicit, fine-grained interpretative framework for designing test items and for linking examinees' test performance to specific inferences about psychological skill acquisition. AHM developments have been documented in the educational and psychological measurement literature, including psychometric advances (e.g., Leighton et al., 2004; Gierl, Leighton, & Hunka, 2007; Gierl, Cui, & Hunka, 2008; Cui, Leighton, Gierl, & Hunka, 2006) and practical applications (e.g., Gierl, Wang, & Zhou, 2007; Wang & Gierl, 2007). The AHM has also been used to study differential item functioning (Gierl, Zheng, & Cui, 2008) and to serve diagnostic adaptive testing (Gierl & Zhou, 2008).

Attribute-Based Reliability Estimation

To-date, however, the AHM has not been applied in an operational diagnostic testing situation because the reliability for attribute-based scoring has not been well developed. Reliability estimation, in educational measurement, considers how the scores resulting from a measurement procedure would be expected to vary across replications of times and test forms (i.e., parallel forms estimate of reliability); replications of times (i.e., test-retest estimate of reliability); and replications of test items (i.e., internal consistency estimate of reliability which contains the Spearman-Brown type procedure; Flanagan, Rulon, and Guttman procedure; KR-20 and KR-21 procedures; and Cronbach's coefficient alpha) (Haertel, 2006). That is, the concern of reliability estimation is to quantify the precision of test scores over repeated administrations. It is critical to determine the extent to which any single score produced by a measurement procedure is likely to depart from the average score over many replications of times, test forms, and/or test items because the variation reflects the reliability of a measurement procedure. Stated another way, a measurement procedure is reliable when it can produce test scores close to each other as it is administered over replications of times, test forms, and/or test items. Conversely, the greater the variation among the scores of a measurement, the less reliable the instrument.

Reliability estimation is critical in diagnostic assessment because diagnostic assessment provides information on examinees' mastery of specific

attributes. A functional diagnostic assessment should yield reliable feedback on an examinee's attribute mastery. In the AHM, attribute reliability refers to the consistency of the decisions made in a diagnostic test about examinees' mastery of specific attributes across multiple observations. Attribute reliability is a fundamental concept in CDA because score reports must provide users with a comprehensive yet succinct summary of the outcomes from testing, including score precision. Standard 5.10 in the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) makes this point clear:

> *When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used. (p. 65)*

A criterion-referenced score interpretation which compares an examinee's performance with a preset standard for acceptable achievement regardless of other examinees' performance is appropriate for use in a cognitive diagnostic assessment. The interpretation identifies an examinee's mastery of specific knowledge and skills assessed. Attribute-based reliability should be calculated, therefore, to examine the consistency of observed response pattern classification.

Purpose of Current Study

The purpose of the present study was to develop and comparatively evaluate a new measure of attribute reliability for cognitive diagnostic assessments designed to provide examinees with information about their cognitive strengths and weaknesses. Two reliability estimation procedures were considered: the standard form of Cronbach's alpha (Cronbach, 1951) and an adapted form in which each item is weighted to take account of those examinees who correctly answer the item but do not possess the attributes needed to correctly answer the item. The study explored which factors influenced each method and if the influence was the same for both methods. The research questions addressed in this study included:

1. Is attribute reliability as determined by adapted and standard Cronbach's alpha influenced by different cognitive model structures, specifically, a linear model versus a divergent model?
2. What is the minimum number of items required to measure each attribute to achieve adequate attribute reliability as determined by adapted and standard Cronbach's alpha?
3. Are the two indices influenced by sample size?
4. Is attribute reliability as determined by adapted and standard Cronbach's alpha influenced by discrepancy of examinees' observed response patterns from expected response patterns?

Organization of the Dissertation

The current study consists of five chapters. Chapter I, the present chapter, described the context of the study, provided a brief introduction to cognitive models and the AHM, and then presented the purpose of the study. Chapter II contains the theoretical framework of the study. It includes a detailed introduction of AHM and an introduction to the procedures for estimating attribute reliability. Details on the research deign, data simulation, and data analyses of the present study are elaborated in Chapter III. The results are then reported in Chapter IV. Chapter V provides a summary and discussion of the results. The conclusions drawn in light of the findings and limitations of the study are then presented. Implications for practice and suggestions for future research conclude this chapter.

Chapter II: Review of Attribute Hierarchy Method for Cognitive Diagnostic Assessment

In Chapter I, it was shown that the theoretical background of the present study lies in the integration of cognitive psychology and educational measurement. In this study, attribute-based reliability for the attribute hierarchy method (AHM) will be introduced and evaluated.

This chapter is divided into three sections. Section 1 gives an overview of cognitive models in diagnostic testing and describes why they are important in educational measurement and, more specifically, in cognitive diagnostic assessment (CDA). Some currently existing cognitive diagnostic models and their features will be reviewed. Section 2 provides a review of the AHM, including the cognitive model representation and its psychometric components. Section 3 introduces the procedures examined for estimating attribute reliability.

Cognitive Models and Educational Measurement

Currently, most large-scale tests are designed using test specifications (Leighton & Gierl, 2007a). Test specifications are represented, most commonly, in a two-way matrix which serves as a blueprint to link intended and actual test score inferences. Specifically, a blueprint describes the content and cognitive skills required for generating a set of relevant items that represents the defined achievement domain. The rows of the blueprint matrix typically represent content coverage while the columns of the matrix indicate cognitive processes to

be measured on the test. Often, there are two reporting categories: knowledge and skills. As relevant items are generated from the test specifications, they are designed to be representative samples of each matrix cell which is the cross combination of rows and columns (Bloom, 1956; Gierl, 1997; Webb, 2006). A sample blueprint of Alberta 2007 Grade 3 Mathematics Achievement Test is presented in Table 1. In this blueprint, the mathematics content to be assessed in the test is listed in each row and the cognitive processes are listed in the columns. For example, the Provincial Mathematics Achievement Test is designed to cover four content areas: *Number*, *Patterns and Relations*, *Shape and Space*, and *Statistics and Probability*. For each content category, the cognitive processes to be measured are classified as knowledge or skills. For Grade 3 examinees, knowledge of Mathematics is about recalling facts, concepts, and terminology; knowing number facts; recognizing place value; knowing procedures for computations; knowing procedures for constructing and measuring; knowing how to use a calculator/computer; and knowing mental computation and estimation strategies. The skills include representing basic mathematical concepts in concrete, pictorial, and/or symbolic modes; applying a mathematical concept in both familiar and new situations; creating new problem situations that exemplify a concept; justifying answers; judging reasonableness of answers; communicating why and when certain strategies are appropriate; applying basic mathematical concepts to solve problems; demonstrating and applying relationships among numbers, operations, number forms, and modes of

representation; explaining relationships among geometric forms; and using a variety of problem-solving strategies. The essential assumption of using test specifications for item development is that examinees would use the skills outlined in the specifications to solve items. The number of items for each cell in the blueprint is specified so test developers can prepare a test containing relevant items that together represent each content area and cognition interaction. That is, the items included in the test are assumed to be relevant and representative of the test specifications.

However, tests constructed from test specifications are not capable of yielding diagnostic claims about examinees' strengths and weaknesses because the grain size of the specifications only defines examinees' knowledge and skills at a general level of detail. That is, items are not designed to measure particular skills because the test specifications only reflect a general description of the domain of achievement. Moreover, studies are rarely conducted to verify that the thinking processes examinees use to answer the test mirror the expectations laid out in the test specifications (Leighton & Gierl, 2007a). Some of the studies that do exist reveal that the cognitive skills reflected in the test specifications do not, in fact, measure important aspects of examinees' thinking processes (Gierl, 1997; Hamilton, Nussbaum, & Snow, 1997; Poggio, Clayton, Glasnapp, Poggio, Haack, & Thomas, 2005). The absence of empirical evidence that the content and skills outlined in the test specifications are identical to what is being applied by examinees becomes an obstacle for providing detailed diagnostic claims

about examinees' cognitive strengths and weaknesses. Thus, large-scale tests which are developed using models of test specifications are regarded as instruments measuring only generalized knowledge and skills.

Currently, a total score which represents only a coarse evaluation of examinees' overall knowledge and skills measured on tests is provided in most large-scale testing programs. This limitation is mainly caused by the absence of any cognitive basis in the test development process. Fortunately, testing organizations have recognized the problem and are starting to address this issue. For example, one of the goals of the Learner Assessment Branch at Alberta Education, the organization responsible for the Alberta Provincial Achievement Tests, is to provide individual examinees with feedback about their performance on the achievement tests (personal communication with psychometricians in Alberta Education). This goal represents Alberta Education's ongoing effort towards linking large-scale tests with teaching and instruction. Also, the Education Quality and Accountability Office, the agency in Ontario responsible for the assessments conducted in Ontario, uses Cognitive Labs (Zucker, Sassman, & Case, 2004) as an activity in the development of educational assessments.

To develop cognitive diagnostic assessments that measure knowledge structures and processing skills so examinees receive information about their cognitive strengths and weaknesses, a cognitive model is required. Leighton and Gierl (2007a, p. 6) defined the term cognitive model in educational and

psychological measurement as: "… simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills examinees at different levels of learning have acquired and to facilitate the explanation and prediction of examinees' performance."

A cognitive model reflects the knowledge structures and processing skills examinees apply to solve educational tasks in a specific achievement domain. The knowledge structure contains factual and procedural information while the processing skills include the transformations and strategies required to manipulate this information (Lohman, 2000). The knowledge structures and processing skills are connected with test performance and test score interpretations through a cognitive model of task performance, which, in turn, provides a detailed framework for identifying and understanding how examinees use their cognitive skills to produce their responses and yield subsequent test scores.

A cognitive model of task performance must be specified at a fine-grain size to represent the detailed knowledge structures and thinking process that underlie examinees' task performance. Cognitive models are able to guide diagnostic inferences if they are specified at a small grain size because only in this manner can the cognitive processes that underlie test performance be magnified and, hence, understood. Each component in a cognitive model, specified at a small grain size, should reflect an identifiable problem-solving skill.

The components required to produce a satisfactory response should then be combined. In this way, test performance can be connected with score interpretations that allow the understanding of how examinees' cognitive skills were used to produce their responses and, subsequently, test scores. Often, due to the lack of existing cognitive models, items in the test are reviewed post hoc to extract the cognitive attributes measured by the items. This is called a post-hoc or retrofitting approach. However, the grain size of cognitive skills obtained using a post-hoc or retrofitting approach is not fine enough because items with these specific cognitive characteristics are unlikely to exist in a test developed without a cognitive model.

The impact of cognitive models of task performance on educational and psychological measurement is far reaching. One of the benefits of a cognitive model is that it is a viable guide for item and test development after the model is evaluated and validated. With the description of specific, fine-grain cognitive skills, test developers can create items according to the structural organization of the cognitive components in a cognitive model. By doing so, the test developer achieves control over the particular cognitive skills each item measures. That is, the assessment principles used in test construction are much more precise allowing items to be validly and efficiently created during the development cycle.

Another benefit of cognitive models is their facility to yield detailed cognitive diagnostic feedback to the examinees about their problem-solving

strengths and weaknesses. These models provide an explicit framework necessary to link cognitively-based inference with explicit, fine-grained test score interpretations (Gierl & Leighton, 2007; Leighton & Gierl, 2007a). A cognitive model of task performance also helps track the underlying knowledge requirements and thinking processes for solving tasks. Therefore, examinees will become aware of their strengths and weaknesses through the diagnostic inferences provided and, consequently, they will be able to improve their learning.

One other benefit of cognitive models is the potential for linking cognition theory with learning and instruction. Instructional principles are decided on the basis of how examinees reason and solve problems. The diagnostic inferences associated with examinees' knowledge and thinking processes will help instructors identify examinees' strengths and weaknesses and adjust, if necessary, their instructional strategies. Cognitive models provide one means to report examinees' cognitive skills on tasks of interest which could be used to associate their test score with instructional procedures designed to improve the examinees' skills (National Research Council, 2001; Pellegrino, 2002; Pellegrino, Baxter, Glaser, 1999).

Among a variety of procedures (e.g., judgmental and logical analyses, generalizability studies, and analyses of group differences; Messick, 1989) that may be used to generate a cognitive model, verbal report methods are

appropriate for the study of human information processing (Leighton & Gierl, 2007b). Researchers can develop a cognitive model by administering tasks to a sample of examinees that represent the intended population, having them think aloud as they respond these tasks, and then conducting protocol or verbal analysis with the corresponding verbal data (Chi, 1997; Ericsson & Simon, 1993; Leighton, 2004; Leighton & Gierl, 2007b; Taylor & Dionne, 2000).

Only a valid cognitive model can be used to empirically confirm the thinking processes individuals use to answer or solve classes of test items. An existing cognitive model can be validated using the same method as it is generated, for example, by verbal reports. A cognitive model can also be evaluated by checking how examinees' observed response data fit expected response patterns derived by the cognitive model. A valid cognitive model of task performance can be used to empirically confirm the thinking processes individuals use to answer or solve classes of test items (Leighton & Gierl, 2007b).

Obviously, a cognitive model plays a crucial role in CDA because it is designed to specify knowledge requirements and processing skills that underlie examinees' task performance when answering an item in a particular domain. To uncover the diagnostic information that is embedded in examinees' item response data, many cognitive diagnostic models have been proposed (e.g., de la Torre & Douglas, 2004; Dibello, Stout, & Roussos, 1995; Embretson, 1980; Fischer, 1973, 1983; Leighton, et al., 2004; Junker & Sijtsma, 2001; Tatsuoka,

1983, 1995). From a psychometric modeling perspective, most cognitive diagnostic models share a common property: they model the probability of yielding a correct response to an item as a function of examinees' attribute mastery associated with different knowledge and skills, regardless of the forms that the models might take. In the following sections, six cognitive diagnostic models will be briefly reviewed to provide information regarding the breadth these models cover in educational measurement.

An Overview of Existing Cognitive Diagnostic Models

Linear Logistic Latent Trait Model

Fischer's (1973, 1983) linear logistic latent trait model (LLTM), which is an extension of the IRT Rasch model, was regarded as the first approach to bring cognitive variables into psychometric models (Stout, 2002). The LLTM intends to account for the difficulty of test items with respect to a set of underlying cognitive skills or attributes which an examinee is hypothetically required to possess for solving items. The IRT item difficulty parameters are rewritten as a linear combination of the difficulties of *K* cognitive attributes. The item response probability of the LLTM can be expressed as:

$$p\left(x_{ij}=1\middle|\theta_i,\eta_k,c\right)=\frac{exp\,(\theta_i-(\sum_{k=1}^{K}q_{jk}\eta_k+c))}{1+exp\,(\theta_i-(\sum_{k=1}^{K}q_{jk}\eta_k+c))}, \qquad \text{(Equation 1)}$$

where

x_{ij} = the response of examinee i to item j,

θ_i = the ability of examinee i,

q_{jk} = the hypothetical minimum number of times that attribute k has to be used in solving item j,

η_k = the difficulty of attribute k, and

c = the normalization constant.

In the LLTM, an examinee's ability is modeled as a unidimensional parameter, θ_i. Since only one ability parameter is specified for each examinee, the LLTM can not be used to evaluate examinees in terms of individual attributes. In addition, as recognized by Embretson (1984, 1991), the cognitive attributes are "compensatory" in the LLTM, indicating that high ability on one attribute can compensate for low ability on other attributes. However, cognitive attributes are often not compensatory in nature. For example, if comprehension of text and algebraic manipulation are both required skills for solving a math problem, high ability on comprehension of text cannot compensate the lack of algebraic skills.

Multicomponent Latent Trait Model

To overcome the shortcomings of the LLTM, Embretson (1984) introduced a non-compensatory model called the multicomponent latent trait model (MLTM). Subtask responses are used to measure cognitive attributes underlying test items in the MLTM. The probability of satisfactory performance

on a test item is expressed as the product of probabilities of satisfying performances on subtasks of the item, each of which follows a separate one-parameter unidimensional IRT model,

$$p(x_{ij} = 1|\theta_i, b_j) = \prod_{k=1}^{K} p(x_{ijk} = 1|\theta_{ik}, b_{jk}) = \prod_{k=1}^{K} \frac{exp\,(\theta_{ik} - b_k)}{1 + exp\,(\theta_{ik} - b_k)},$$

(Equation 2)

where

θ_i = the vector of K subtask abilities for examinee i,

b_j = the vector of K subtask difficulties for item j,

x_{ijk} = the response of examinee i to subtask k for item j,

θ_{ik} = the ability of examinee i on subtask k, and

b_k = the difficulty of subtask k.

By using the multiplicative form of the probabilities for solving each subtask correctly, the MLTM captures the non-compensatory nature of cognitive attributes. Moreover, an examinee's ability parameters for subtasks can be estimated in situations in which several cognitive subtasks are required simultaneously to solve each of the test items correctly. However, a major limitation related to the MLTM is that this approach requires examinees' responses to subtasks of each item, which cannot be directly obtained from

multiple-choice items. As a result, the practicability of the MLTM for cognitive diagnosis is, to some extent, restricted.

Rule Space Model

One widely recognized cognitive diagnostic model is Tatsuoka's (1983, 1991, 1995) rule space model, which is currently used with the Preliminary Scholastic Assessment Test (PSAT). As Stout (2002) pointed out, the rule space model is "a major pioneering milestone, both from the psychometric and the formative assessment perspectives" (p. 508). Generally speaking, the rule space model contains two sequential parts. The first part of this model is to define an attribute-by-item incidence matrix (Q matrix) of order K by J, and to derive the universal set of knowledge states from the incidence matrix. The Q matrix is a predefined binary matrix consisting of 1s and 0s, where the 1s in the j-th column identify which of the K attributes are necessary for successful performance on item j. For example, a hypothetical Q matrix is shown as follows:

$$Q_{5,12} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad \text{(Matrix 1)}$$

This matrix consists of five rows and twelve columns, with each row corresponding to an attribute and each column corresponding to an item. The first column of this Q matrix shows that item 1 is measuring attribute 1. The second column indicates that item 2 is measuring attributes 1 and 2. The rest of

columns can be interpreted in the same manner. In the rule space model, an examinee is assumed to have mastered all the attributes that an item is measuring in order to answer the item correctly. Therefore, in order to answer item 2 correctly, the examinee must have mastered attributes 1 and 2. The Q matrix is typically obtained from a task analysis conducted by test developers or content experts by reviewing test items and identifying the attributes that underlie the items. Once the Q matrix is established, knowledge states can be derived and related to examinees' observable response patterns by using Boolean description functions (Tatsuoka, 1991; Varadi & Tatsuoka, 1992). In the rule space model, each cognitive attribute is dichotomized as mastered or non-mastered. As a result, knowledge states, used to describe examinees' profiles of cognitive skills, are represented by a list of mastered/non-mastered attributes.

The second part of the rule space model is to classify each observed response pattern into one of the knowledge states obtained from the analysis of the first part of the model (i.e., specification of the Q matrix). The rule space model uses a two-dimensional Cartesian coordinate system, characterized by theta (θ, the ability level from the IRT model) and zeta (ζ, an index measuring atypicality of response patterns), and a Bayesian decision rule for minimizing errors to facilitate inferences about examinees' knowledge states. By creating knowledge states from the Q matrix and then classifying observed item responses into one of the knowledge states, a link is established between examinee cognition and psychometric applications.

The Unified Model

Based on the idea of Tatsuoka's rule space model, Dibello et al. (1995) proposed a new cognitive diagnostic model called the unified model, which "brings together the discrete, deterministic aspects of cognition favoured by cognitive scientists, and continuous, stochastic aspects of test response behaviour that underlie item response theory" (Dibello et al., 1995, p. 361). The unified model adds to the rule space approach a cognitively-based IRT model, which is modeled with respect to discrete cognitive states and a continuous latent ability (Dibello, et al., 1995). In the unified model, each examinee is characterized by a dichotomous vector a_i representing the examinee's attribute mastery profile and a latent "residual" ability θ_i which is not captured by the Q matrix. Dibello et al. identified four possible sources of response behaviour that could lead to the variation in observed response patterns from those predicted by or derived from the Q matrix. These sources are: 1) the use of a different strategy from that presumed by the Q matrix, 2) the incompleteness of the Q matrix for attributes, 3) the positivity of attribute for the item (corresponding to the possibility that an examinee who possesses an attribute may fail to apply it correctly to an item and an examinee who lacks the attribute may still answer the item correctly by possessing partial knowledge), and 4) the possibility that an examinee makes a random error. The unified model incorporates these four sources of variation in the following equation for the item response probability:

$$p(x_{ij} = 1|\theta_i, a_i) = (1 - p)\left\{d_j \prod_{k=1}^{K} \pi_{jk}^{a_{ik}} r_{jk}^{(1-a_{ik})} p_j(\theta_i + \Delta c_j) + (1 - d_j) p_j(\theta_i)\right\}, \quad \text{(Equation 3)}$$

where

p = probability of making a random error,

d_j = probability of using attributes specified in the Q matrix to solve item j,

a_{ik} = the k^{th} element of vector a_i,

c_j = completeness index of attributes required for item j,

π_{jk} = P(Attribute k applied correctly to item $j|a_{ik} = 1$),

r_{jk} = P(Attribute k applied correctly to item $j|a_{ik} = 0$),

Δ = 2, and

$p_j(x)$ = one parameter logistic model with difficulty b_j.

Comparable to Embretson's MLTM, the unified model captures the non-compensatory nature of cognitive attributes as the probability of satisfying performance on an item derived from the Q matrix is presented as a product of the probabilities of applying each attribute correctly. Moreover, the explicit expression of the item response probabilistic function makes the likelihood-based classification procedures straightforward. However, the unified model

encounters an identifiability problem given that the item response data are essentially not rich enough to make all the item parameters identifiable. In an effort to solve the identifiability problem, Hartz (2002) reparameterized the unified model so that it can produce statistically identifiable and well interpretable parameters.

The DINA and NIDA Model

There are many other cognitive diagnostic models based upon the Q matrix in the literature, such as the deterministic input noisy and gate model (DINA) (de la Torre & Douglas, 2004; Doignon & Falmagne, 1999; Haertel, 1989; Junker & Sijstma, 2001; Macready & Dayton, 1977; Tatsuoka, 2002) and the noisy input deterministic and gate model (NIDA) (Junker & Sijstma, 2001). The DINA model partitions examinees into two classes for each item, those who have mastered all the attributes required by an item ($\xi_{ij} = 1$) and those who have not ($\xi_{ij} = 0$). It models the probability of a correct response to an item with two parameters: the probability that an examinee fails to answer the item correctly when the examinee has mastered all required attributes (s_j, the "slipping" parameter) and the probability that an examinee gets the correct answer when the examinee does not possess all of the required attributes (g_j, the "guessing" parameter). The item response probability can be written as:

$$p\left(x_{ij} = 1 \middle| \xi_{ij}, s_j, g_j\right) = (1 - s_j)^{\xi_{ij}} g_j{}^{(1-\xi_{ij})}, \qquad \text{(Equation 4)}$$

where x_{ij} is the response of examinee i to item j.

The NIDA model extends the DINA model by defining a slipping parameter s_k and a guessing parameter g_k for each attribute, independent of the item. That is, for all the items that require attribute k, the slipping parameter s_k and the guessing parameter g_k for attribute k are constant across these items. The NIDA model gives the probability of a correct response as:

$$p(x_{ij} = 1 | a_i, s, g) = \prod_{k=1}^{K} [(1 - s_k)^{a_{ik}} g_k{}^{(1-a_{ik})}]^{q_{jk}}, \quad \text{(Equation 5)}$$

where

a_i = the vector of the attribute profile for examinee i,

s = the vector of attribute slipping parameters,

g = the vector of attribute guessing parameters,

q_{jk} = the element of the Q matrix in the j^{th} row and k^{th} column, and

a_{ik} = the k^{th} element of vector a_i.

All of the cognitive diagnostic models discussed in this section require the specification of the Q matrix, based on which test developers can use to design test items according to a presumed set of attributes. The Q matrix, however, does not provide information in terms of the relationships among attributes. The attributes might be independent of each other in the sense that the mastery of each attribute does not depend on the possession of any other attributes in the Q matrix. However, cognitive research suggests that cognitive skills do not

operate in isolation but indeed function as a network of interrelated processes (e.g., Anderson, 1996; Kuhn, 2001; Mislevy, Steinberg, & Almond, 2003). As a result, it is necessary to build the relationships or dependencies among attributes into cognitive diagnostic models and, in turn, integrate this information into statistical pattern classification procedures.

In 2004, Leighton, Gierl, and Hunka introduced a procedure for cognitive diagnostic assessment called the *attribute hierarchy method* (AHM). The AHM brings an important cognitive property, *attribute dependency*, into cognitive modeling methodologies because the AHM is based on the assumption that test performance is associated with a specific set of hierarchically-organized cognitive components called attributes. This method is reviewed in the next section.

The Attribute Hierarchy Method

The attribute hierarchy method (AHM; Gierl, Leighton, & Hunka, 2007; Gierl, Cui, & Hunka, 2008; Leighton, Gierl, & Hunka, 2004) refers to a cognitively-based psychometric method that classifies examinees' test item responses into a set of structured attribute patterns with reference to a cognitive model of task performance. In the AHM, a cognitive attribute is defined as a description of the procedural or declarative knowledge required to perform a task in a specific domain of achievement (Leighton et al., 2004). Attributes are considered to be hierarchically related and therefore can be ordered into a hierarchy based upon their logical and/or psychological properties (Leighton & Gierl, 2007b). The

hierarchical structure of the AHM reflects an important characteristic of human cognition because cognitive skills operate dependently (Anderson, 1996; Mislevy, Steinberg, & Almond, 2003). A cognitive attribute hierarchy functions in the framework of the AHM by providing a manifest description of cognitive attributes required to solve test items and the relationships among these cognitive attributes (Leighton et al., 2004).

Once a cognitive attribute hierarchy for a specific content domain is identified and validated, test developers can create items and tests according to the cognitive characteristics that embody each attribute and the hierarchical organization of these attributes. Consequently, the test developer achieves control over the specific attribute each item measures and the cognitive features of the test composed by these items. The AHM also offers a more convenient way of providing cognitive feedback to examinees. This feedback is provided by mapping observed examinee response patterns onto expected examinee response patterns derived from the attribute hierarchy. An examinee with a certain observed response pattern is expected to master the attributes implied by the corresponding expected response pattern. As a cognitively-based psychometric approach, the AHM consists of two major components: the cognitive model representation component and the psychometric component.

Cognitive Model Representation Component of the AHM

The cognitive model representation component of the AHM refers to the

specification of the cognitive attribute hierarchy. A cognitive attribute hierarchy specifies the cognitive attributes to be measured by a test and the interrelationships among these attributes. There are two types of cognitive model structures, linear and divergent, as presented in Figure 1. The hypothesized linear model presented in Figure 1a contains all three attributes aligned in a single branch. This type of model could be used to characterize problem-solving when the knowledge and skills are ordered in a linear manner. Knowledge and skills that are restricted within a domain of basic logical application, for example, could be characterized in a linear mode. In other words, the attributes in a linear model measure a single construct at varying difficulty levels.

The second cognitive model is a more complex divergent hierarchy, as presented in Figure 1b. The hypothesized divergent hierarchy contains two independent branches which share a common prerequisite—attribute A1. Aside from attribute A1, the first branch includes two additional attributes, A2 and A3, and the second branch includes attribute A4. This type of model could be used to characterize problem-solving when the knowledge and skills differ as a function of the concepts and content within a domain. Examples of divergent model can be found in the study of Gierl, Wang, and Zhou (2008). The four cognitive hierarchies they used to describe examinee performance on the SAT algebra subtest are presented in Figure 2. The descriptions of attributes involved in the four cognitive hierarchies are presented in Appendix 1. Attributes in each

of the four hierarchies are organized in divergent manner, indicating the knowledge and skills they measure are unrestricted within one domain. However, these attributes are categorized in one cognitive hierarchy because the knowledge and skills they measure have some characteristics in common. Taken together, these two model structures, linear and divergent, represent different types of cognitive structures that could characterize examinee performance on a diagnostic test.

To specify the cognitive model components and their organization, the formal representations of the hierarchy, four different sequential matrices are developed: the adjacency, reachability, incidence, and reduced Q matrices (Gierl, Leighton, & Hunka, 2000; Leighton et al., 2004; Tatsuoka, 1995). The attribute hierarchy example presented in Figure 1b is used to identify and describe the different matrices in the AHM.

In the AHM, a binary adjacency matrix (A) of order (K, K), where K is the number of attributes, specifies the direct relationships among attributes. In the adjacency matrix, the diagonal elements are denoted as 0s while the off-diagonal elements are 1s or 0s depending on the relationship between two attributes. 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. It can be expressed as follows:

$$a_{jk} = \begin{cases} 1, & if\ attribute\ j\ is\ the\ prerequisite\ of\ attribute\ k \\ 0, & otherwise \end{cases}. \quad \text{(Equation 6)}$$

For example, the adjacency matrix for the hierarchy example shown in Figure 1b is:

$$A_{4,4} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{(Matrix 2)}$$

Row 1 of matrix 2 indicates that attribute A1 is a direct prerequisite to attribute A2 and A4 (i.e., $a_{12} = 1$ and $a_{14} = 1$) but not the direct prerequisite of attribute A3 (i.e., $a_{13} = 0$). Similarly, attribute A2 is the direct prerequisite of attribute A3 as $a_{23} = 1$. Attribute A3 and A4 are direct prerequisites of no other attributes because all elements are 0 in row 3 and row 4 of the matrix.

The adjacency matrix only expresses the direct relationship between attributes. To specify the direct and indirect relationship among attributes, a reachability matrix is used. The reachability matrix (R) of order (K, K), where K is the number of attributes, specifies both the direct and indirect relationships among attributes. The reachability matrix can be derived from the adjacency matrix by performing Boolean addition and multiplication. Boolean addition is defined by $1 + 1 = 1,\ 1 + 0 = 1,\ 0 + 1 = 1,\ and\ 0 + 0 = 0$. Boolean multiplication is defined by $0 * 0 = 0,\ 1 * 0 = 0,\ 0 * 1 = 0,\ and\ 1 * 1 = 1$. The R matrix is calculated using $R = (A + I)^n$, where n is the integer required for R to reach invariance, $n = 1, 2, \ldots, k$, given A is the adjacency matrix, and I is an identity matrix of order (K, K). For the divergent hierarchy example presented above, $A + I$ is equal to:

$$A + I = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{(Matrix 3)}$$

When $A + I$ is multiplied by itself repeatedly using Boolean algebra until the product become invariant, the resulting matrix is the reachability matrix. For example, to calculate the reachability matrix for the hierarchy shown in Figure 1b, $(A + I)^n$ is calculated for $n = 1, 2, 3, 4$ separately. Because the resulting matrices are same for $n = 3$ and $n = 4$, $(A + I)^3$ is, therefore, the R matrix for the hypothetical divergent hierarchy:

$$R_{4,4} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{(Matrix 4)}$$

The $j^{t\square}$ row of the R matrix specifies all the attributes, including the $j^{t\square}$ attribute, that the $j^{t\square}$ attribute can reach through direct or indirect connections. In the hypothetical reachability matrix, row 1 indicates that attribute A1 can reach itself and all other attributes through direct or indirect relations as all elements on row 1 are 1s. Row 2 indicates that attribute A2 can reach itself and attribute A3 (i.e., $r_{22} = 1 \; and \; r_{23} = 1$); row 3 and row 4 indicate that each of attributes A3 and A4 is a direct prerequisite of itself but neither is a direct or indirect prerequisite of any of the other attributes (i.e., only $r_{33} = 1 \; and \; r_{44} = 1$ in the corresponding row). In the AHM, the reachability matrix is used to select a subset of items from the potential pool of items, which corresponds to the dependencies of the attribute hierarchy.

In order to maximize the control over the attributes each item measures, prior to the development of test items, the attribute hierarchy should be identified to represent the hierarchical relationship among attributes (Leighton et al., 2004). That is, the attribute hierarchy should be used to guide test and item development by which test items can be designed to assess the specific attributes taking into consideration of the inter-attribute hierarchical relationships.

The potential pool of items includes all combinations of attributes and represents the set of potential items when the attributes are independent of one other. That is, the adjacency matrix is a matrix of order (K, K) with all elements equal to zero and the reachability matrix is an identity matrix of order (K, K), where K is the number of attributes. The attributes in the potential pool of items are described by the incidence matrix (Q). The Q matrix is of order (K, P), where K is the number of attributes and P, which equals to $2^K - 1$, is the number of potential items. Each column of the Q matrix represents one item. The 1s in the column identify which attributes are required for successful performance on this item. The columns of the Q matrix are created by converting the integers ranging from 1 to $2^K - 1$ to their binary form. The Q matrix for the hypothetical divergent hierarchy is shown below:

$$Q_{4,15} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \qquad \text{(Matrix 5)}$$

For this hypothetical divergent hierarchy, the number of items (columns) in the Q matrix is $2^4 - 1 = 15$, therefore the Q matrix is of order $(4, 15)$. Column 1 of the Q matrix represents item 1 which indicates that only attribute 1 is required for examinees to correctly answer this item. Column 15 identifies that all four attributes are required for solving item 15. The rest of the columns can be interpreted in the same way.

In the AHM, the attributes are related hierarchically as some attributes are prerequisites of other attributes. Correspondingly, an item that probes an attribute must at the same time probe its prerequisite attribute(s). As a result, when the attributes share dependencies, the size of the potential item pool can be significantly reduced by imposing the constraints of the attribute hierarchy as embodied in the reachability matrix. For example, in the Q matrix presented above, item 2 (column 2 of the matrix: [0100]) indicates that it is developed to measure attribute A2. However, the hierarchy indicates that attribute A2 has attribute A1 as its prerequisite. Hence, the item that measures attribute A2 must measure attribute A1 as well. The representation of that item is [1100], which is identical to the third column of the Q matrix (item 3). For this reason, item 2 can be removed as the hierarchy imposes dependencies among the attributes. Logically speaking, the removal of items in this manner results in a reduced Q matrix (Q_r) which represents the dependencies among attributes.

In practice, the Q_r matrix is derived by determining which columns of the R matrix are logically included in columns of the Q matrix, using Boolean

inclusion (Tatsuoka, 1991). For example, column 3 of the reachability matrix specifies that any item that probes attribute 3 must also assess attribute 1 and 2. If the item does not measure these two additional attributes, the item does not match the attribute hierarchy and, therefore, would be removed. The Q_r matrix is of order (K, I) where K is the number of attributes and I is the reduced number of items resulting from the constraints imposed by the hierarchy. For the hypothetical divergent hierarchy, the Q_r matrix is as follows:

$$Q_{r_{4,6}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \quad \text{(Matrix 6)}$$

The Q_r matrix is of order $(4, 6)$. Thus, out of a potential pool of 15 items, if the attribute hierarchy is valid, only six items are logically meaningful to measure the domain of interest according to the hypothetical divergent hierarchy shown in Figure 1b.

The Q_r matrix represents the cognitive specifications or blueprint for the test as it describes attribute-by-item level combinations in the hierarchy. Therefore, the Q_r matrix can be used as a guide to develop and interpret items that measure specific attributes outlined in the hierarchy. In order to systematically evaluate each component in the cognitive model and in turn provide diagnostic inferences, items must be developed to directly mirror each attribute combination in the Q_r matrix. The Q_r matrix includes the total number of single items that is needed. For the hypothetical example, the Q_r matrix is of order $(4, 6)$, indicating that six items must be developed to measure the four

attributes in the hypothesized hierarchy. In the Q_r matrix, the first column indicates that an item must be created to measure attribute A1 while column 2 of the Q_r matrix indicates that an item must be created to measure both attributes A1 and A2. The remaining columns are interpreted in the same manner.

Given the reachability matrix and the Q_r matrix, the attribute patterns and expected response patterns for a group of examinees can then be generated. Attribute pattern refers to the combination of attributes that is consistent with the attribute hierarchy. The attribute pattern matrix is the transpose of a Q_r matrix with one additional row. This row, containing all 0s for each attribute, indicates that an examinee has mastered none of the attributes designated in the hierarchy. The attribute pattern matrix of the hypothesized divergent hierarchy is presented below:

$$AP_{7,4} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad \text{(Matrix 7)}$$

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 assessing the cognitive attributes that the examinee has fully mastered, but fails the items which evaluate the cognitive attributes that the examinee has not mastered. That is to

say, expected response patterns are those response patterns that can be clearly explained by the presence or absence of the attributes without any errors or "slips."

The expected response matrix (E) is created, again using Boolean inclusion (Tatsuoka, 1991), where the algorithm compares each row of the expected attribute pattern matrix to the columns of the Q_r matrix. The E matrix is of order (J, I), where J is the number of expected responses and I is the reduced number of items resulting from the constraints imposed by the hierarchy. The rows of the E matrix are those responses that would be produced by an examinee who possesses the attributes as defined and structured in the attribute hierarchy and presented by the columns of the Q_r matrix. The columns of the E matrix are the items that probe specific attribute combinations. When an examinee's attributes match those attributes measured by an item, a correct answer is expected. The expected response pattern matrix of the hypothesized divergent hierarchy is as follows:

$$E_{7,6} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \qquad \text{(Matrix 8)}$$

Given the assumption that the hypothesized divergent hierarchy is true, an examinee, who has an expected attribute pattern of $[1000]$, as presented in the second row of the attribute pattern matrix for the hypothetical example, is expected to have mastered only attribute 1. Hence, the examinee should

correctly answer item 1 but incorrectly answer the rest of the items (i.e., the second row of the expected response pattern matrix $[100000]$). Conversely, if an examinee has an observed response pattern $[110110]$, as presented in the sixth row of the expected response pattern matrix, then it can be inferred that the examinee has mastered attributes 1, 2, and 4 but has not mastered attribute 3 (i.e., the sixth row of the attribute pattern matrix $[1101]$). The remaining rows of attribute pattern and expected response pattern can be interpreted in the same manner.

Noticeably, an examinee who possesses attributes 1 and 2 (i.e., $[1100]$) can produce the expected response pattern $[110000]$ and obtain a total score of 2. An examinee who masters attributes 1 and 4 (i.e., $[1001]$) is expected to correctly solve items 1 and 4 (i.e., $[100100]$) and has a total score 2. That is, examinees with an equal total score do not necessarily master the same attribute patterns. Hence, an examinee's total score can not be consistently coupled with a single attribute pattern. Apparently, total scores are not adequate to diagnose examinees' cognitive strengths and weaknesses. The attribute pattern and the expected response pattern establish the correspondence between an examinee's test performance and the examinee's expected attribute pattern, and yields 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.

Psychometric Component of the AHM

After attribute hierarchies are specified, psychometric procedures are required to apply the AHM in test analysis. These procedures include approaches for observed response pattern classification (Gierl, Cui, & Hunka, 2008) and calculation of hierarchy consistency indices (Cui, Leighton, Gierl, & Hunka, 2006). To provide an overall review of attribute hierarchy method, these two procedures will be reviewed next.

Classification of Observed Response Patterns

In a real testing situation, discrepancies will occur between the observed response patterns and expected response patterns because slips are inevitable. A slip is the discordance of the expected response pattern based on the hierarchy compared to the observed response pattern. For example, the examinees may have the required attributes for an item, but due to carelessness or a mistake in writing on the answer sheet, they get the item wrong. Conversely, some examinees may not master the required attributes, but by guessing or by applying partial knowledge, they could get the item correct.

An examinee's observed response pattern is judged relative to the expected response patterns in the E matrix under the assumption that the cognitive model is true. The attribute probability is the probability that an examinee possesses these specific attribute combinations. Hence, the purpose of calculating attribute probability is to identify the attribute combinations that the examinee is likely to possess, given their observed response pattern. These

probabilities provide examinees with specific information about their attribute-level mastery as part of the test reporting process.

In the AHM, three classification methods have been developed to date (Cui, Leighton, Gierl, & Hunka, 2006; Leighton et al., 2004; Gierl, Cui, Hunka, 2008). Among three methods, two are IRT-based procedures (Method A and Method B) while the third is the artificial neural network approach. In these two methods, the probability of a correct response to individual items is first calculated for each expected response pattern using an IRT model. The three-parameter logistic IRT model is given by:

$$p(x_{ij} = 1|\theta_{i,}\, a_j, b_j, c_j) = c_j + \frac{1-c_j}{1+e^{-1.7a_j(\theta_i - b_j)}}, \qquad \text{(Equation 7)}$$

where a_j is the item discrimination parameter for item j, b_j is the item difficulty parameter for item j, c_j is the pseudo-guessing parameter for item j, and θ_i is the ability parameter for examinee i. The two-parameter logistic IRT model is a special case of the three-parameter model in which the c_i parameter is set to 0.0. The one-parameter model also called Rasch model is another form of the logistic IRT model in which all the items are assumed to have equal discrimination power and no guessing. Item parameters can be estimated based on the expected response patterns using BILOG 3.11 (Mislevy & Block, 1990). According to the psychometric features of a test, one of the three models is used to meet specific requirements.

Once item parameters and the theta value associated with each expected response pattern are estimated, the IRT probability of a correct response to each

item can be calculated for each expected response pattern. 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:

$$P_{ijExpected}(\theta_j) = \prod_{k \in S_{i0}} P_{jk}(\theta_j) \prod_{m \in S_{i1}} [1 - P_{jm}(\theta_j)], \qquad \text{(Equation 8)}$$

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

For illustration, Table 2 presents the information for classifying the observed response pattern [101000] 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 for illustration purposes only). The observed response pattern has different numbers of slips when compared with the expected response patterns. However, as the expected response pattern [111000] corresponds to the largest likelihood, the observed response pattern [101000] should be classified this expected response pattern with one 1 - > 0 slip. Because the expected response pattern [111000] corresponds to the attribute pattern [1110], diagnostic information can be provided to the examinees with such an observed response pattern that they have mastered attributes 1, 2, and 3, but need more work on attribute 4.

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 with the observed response pattern. For the expected response patterns that are not logically included in the observed response pattern, the likelihood of the slips of the form 1 -> 0 is computed as:

$$P_{ijExpected}(\theta_j) = \prod_{m \in S_{i1}} [1 - P_{jm}(\theta_j)], \qquad \text{(Equation 9)}$$

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 3 presents the information for classifying the observed response pattern [101000] 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 set 0.20 as the criterion, then the observed response pattern [101000] can be classified to the expected response pattern [111000], 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.

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 indicates, 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, even with such small values, the expected response pattern is most likely because its likelihood is highest among all the expected response patterns. Another weakness common to both methods A and B is that they rely on IRT-models and assumptions about the distribution of examinees, which are restrictive.

To address these problems, an artificial neural network approach can be used to estimate the probability that examinees possess specific attributes, given their observed item response pattern (Gierl, Cui, & Hunka, 2008). A neural network is a type of parallel-processing architecture that transforms a stimulus received by the input unit to a signal for the output unit through a series of mid-level hidden units. The input layer units produce a weighted linear combination of the inputs which are then transformed to non-linear weighted sums that are passed to every hidden layer unit. The hidden layer units, in turn, produce a weighted linear combination of the inputs which are transformed to non-linear weighted sums that are passed to every output layer unit. The network serves as a powerful pattern recognition technique because it can map any relationship between input and output.

The input (called exemplars in the terminology of neural network) to train the neural network is the set of expected response vectors produced from the AHM analysis. For each expected response vector there is a specific combination of examinee attributes (i.e., the transpose of the Q_r matrix). The examinee attribute patterns, as the expected response vectors, are meaningful because they are derived from the cognitive model and provide a description of each attribute pattern that should be associated with each expected response pattern. The relationship between the expected response vectors with their associated attribute vectors is established by presenting each pattern to the network repeatedly until it learns the associations (Gierl, Cui, & Hunka, 2008).

The final result is a set of weight matrices, one for cells in the hidden layer and one for the cells in the output layer, which can be used to transform any response vector to its associated attribute vector.

The functional relationship for mapping the examinees' observed response patterns onto the expected response pattern so that their attribute probabilities can be computed is given as follows. Let

$$F(z) = \frac{1}{1+e^{-z}}, \qquad \text{(Equation 10)}$$

and

$$a_k = \sum_{j=1}^{q} v_{kj}\, F(-\sum_{i=1}^{p} w_{ji}\, x_i), \qquad \text{(Equation 11)}$$

then the output for unit k, M_k^*, is given as

$$M_k^* = F(a_k), \qquad \text{(Equation 12)}$$

where q is the total number of hidden units, v_{jk} is the weight of hidden unit j for output unit k, p is the total number of input units, w_{ji} is the weight of input unit i for hidden unit j, and x_i is the input received from input unit i. The output values, scaled from 0 to 1, can be interpreted as probabilities. Using this approach, individual attribute probabilities can be computed for each observed response pattern thereby providing examinees with specific judgments about whether an examinee has mastered a certain attribute or not. Table 4 presents the classification information produced by the neural network approach for the observed response pattern [101000]. Based on the information, we can

conclude that examinees with this observed response pattern likely have mastered attributes 1, 2, and 3, as the probabilities for these three attributes are high, and that they likely have not mastered attribute 4, as the probability for this attribute is quite low.

Evaluation of Hierarchical Consistency

Cui, Leighton, Gierl, and Hunka (2006) proposed a model-data fit index for the AHM, the hierarchy consistency index (HCI_j), to evaluate the consistency between the expected and observed response patterns. The HCI_j examines the degree to which observed examinee response patterns generated from a large group of examinees is consistent with the expected response patterns generated from the attribute hierarchy. Given k attributes and i items, the element q_{ki} of the Q_r matrix indicates if attribute k is required to solve the $i^{t\square}$ item. It can be represented as

$$q_{ki} = \begin{cases} 1 & attribute\ k\ required\ by\ item\ i \\ 0 & otherwise \end{cases} \qquad \text{(Equation 13)}$$

Attribute mastery occurs when examinees correctly answer the items requiring the attribute. Thus, the HCI for examinee j is specified as

$$HCI_j = 1 - \frac{2\sum_{i \in S_{correct_j}} \sum_{g \in S_i} X_{j_i}(1 - X_{j_g})}{N_{C_j}}, \qquad \text{(Equation 14)}$$

where $S_{correct_j}$ includes items that are correctly answered by examinee j, X_{ji} is examinee j's score (0 or 1) to item i, X_{j_g} is examinee j's score (0 or 1) to item g,

S_i includes only those items that have attributes that are logically included in the attributes of item i, and N_{cj} is the total number of comparisons for correct-answered items by examinee j.

The term $\sum_{i \in S_{correct_j}} \sum_{g \in S_i} X_{ji}(1 - X_{j_g})$ in the numerator of the HCI_j represents the number of misfits between examinee j's item response vector and the Q_r matrix. If examinee j correctly answers item i, $X_{j_i} = 1$, then the examinee is also expected to answer item g that belongs to S_i correctly, namely, $X_{j_g} = 1(g \in S_i)$, where S_i includes items that require the subset of attributes measured by item i. If the examinee fails to correctly answer item g, $X_{j_g} = 0$, then $X_{j_i}(1 - X_{j_g}) = 1$ and it is a misfit between examinee j's observed response pattern and the expected response patterns specified by the attribute hierarchy. Thus, $\sum_{i \in S_{correct_j}} \sum_{g \in S_i} X_{ji}(1 - X_{j_g})$ is equal to the total number of misfits. The denominator of the HCI_j, N_{C_j}, contains the total number of comparisons for items that are correctly answered by examinee j. When the numerator of the HCI_j is set to equal the total number of misfits multiplied by 2, the HCI_j has the property of ranging from -1 to +1, which makes it relatively straightforward to interpret. The value of the HCI_j ranges from a perfect misfit of -1 to a perfect fit of 1. As the examinee's observed response pattern matches the hierarchy perfectly, the numerator is 0 and the HCI_j has a value of 1. When the examinee's observed response maximally misfits the hierarchy, the numerator is $(2 * N_{cj})$ and the HCI_j will have a value of -1. Therefore, HCI_j 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. Moreover, the mean and standard deviation of the HCI_j can be used as indicators of the overall model-data fit. Cui (2007) demonstrated that HCI_j values above 0.70 indicate good model-data fit.

Estimation of Attribute Reliability

Reliability Estimation

To date, the AHM has not been applied in an operational diagnostic testing situation because the reliability for attribute-based scoring must first be established before the AHM can be put into practice. In conventional testing theory, the concern of reliability estimation is to quantify the precision of test scores. The issue of reliability considers how the scores resulting from a measurement procedure would be expected to vary across replications of times and test forms (i.e., parallel forms estimate of reliability), replications of times (i.e., test-retest estimate of reliability), and replications of test items (i.e., internal consistency estimate of reliability which contains the Spearman-Brown procedure, Flanagan, Rulon, and Guttman procedures, KR-20 and 21 procedures, and Cronbach's α) (Haertel, 2006). It is critical to determine the extent to which any single score of a measurement procedure is likely to depart from the average score over many replications of times, test forms, and/or test items, as that is the way to represent the precision of test score report. The greater the variation

among the scores, the less reliable is the instrument. In addition, to any test form, reliability is a topic that has important implications because only if a test is reliable can the scores be validly interpreted.

Test score reliability must be conceived relative to particular testing purposes and contexts (Haertel, 2006). In the environment of a CDA, attribute-based reliability is considered, therefore, to examine the consistency of observed response pattern classification. Attribute reliability is a fundamental concept in cognitive diagnostic assessment because score reports yielded by CDA must provide users with a comprehensive yet succinct summary of the outcomes from testing, including score precision.

Attribute Reliability

Attribute reliability refers to the consistency of the decisions made in a diagnostic test about examinees' mastery of specific attributes. One method to estimate attribute reliability is by calculating Cronbach's (1951) coefficient alpha, which can be interpreted as the ratio of true score variance to observed score variance on the items that are probing each attribute. The reliability estimation for attribute k using standard Cronbach's α formula is given as

$$\alpha_{AHM_k} = \frac{n_k}{n_k - 1}\left[1 - \frac{\sum_{i \in S_k} \sigma^2_{X_i}}{\sigma^2_{\sum_{i \in S_k} X_i}}\right], \qquad \text{(Equation 15)}$$

where n_k is the number of items that are probing attribute k in the Q_r (i.e., the number of elements in S_k), S_k includes items that measure attribute k in the Q_r, $\sigma^2_{X_i}$ is the variance of the observed scores on item i, $\sum_{i \in S_k} \sigma^2_{X_i}$ is the sum of the

variance of the observed score on the items that are measuring attribute k, and $\sigma^2_{\sum_{i \in S_k} X_i}$ is the variance of the observed total scores.

In the AHM, an item can be designed to measure a combination of attributes. Consequently, for items that measure more than one attribute, each attribute only contributes to a part of the total item-level variance. The index, therefore, incorporates the concept of *attribute dependency* into the reliability calculation. The consideration of attribute dependency is important because theories of learning and performance which dictate the conditions over which the construct is expected to change or remain unchanged play a central role in attributing consistency or inconsistency in test takers' performance (Nichols & Smith, 1998). Consider attribute 1 in Figure 1b. Attribute 1 is the prerequisite for all other attributes in this hierarchy because an examinee must possess attribute 1 in order to correctly respond to items measuring any other attribute in the hierarchy. Similarly, if an examinee correctly answers items that directly probe attribute 2, then it can be inferred that the examinee has also mastered attribute 1 because of their structural relationship in the hierarchy. More generally, attribute 1 is measured directly or indirectly by all test items included in the set of items for the AHM model displayed in Figure 1b, and, thus, to calculate the reliability of attribute 1, all test items must be included. Now consider attribute 3 in Figure 1b. Attribute 3 does not serve as the prerequisite of any other attribute. If an examinee produces a correct answer to items that require attribute 4, for instance, then we could not discern if the examinee had

mastered attribute 3 because attributes 3 and 4 are independent. Hence, only items that directly probe attribute 3 can be included in the reliability estimate for attribute 3. Attribute dependency also implies that prerequisite attributes in the initial nodes of the hierarchy, such as attribute 1, are expected to have higher reliability estimates compared to attributes in the final nodes of the hierarchy, such as attributes 3 or 4 in Figure 1b because of the dependencies among the attributes that, in turn, affect the number of items that measure each attribute, either directly or indirectly.

Different from the first method of estimating attribute reliability which treats the contribution of each attribute towards an examinee's item-level performance equally, the second method to estimate attribute reliability is to isolate the contribution of each attribute to an examinee's item-level performance, in which the item score is weighted by the subtraction of two conditional probabilities. The first probability is associated with attribute mastery (i.e., an examinee who has mastered the attribute(s) can answer the item correctly) and the second probability is associated with attribute non-mastery (i.e., an examinee who has not mastered the attribute(s) can answer the item correctly). The weighted scores for items that measure the attribute are used in the reliability calculation.

Let W_{ik} denote the weight for item i in the calculation of attribute k. A W_{ik} value of 1 indicates that performance on item i is completely determined by attribute k. Hence, the variance of the responses on item i should be used in the

calculation of the reliability for attribute k. Conversely, if W_{ik} has a value of 0, indicating that the mastery of attribute k could not increase the probability of solving item i correctly, then item i should not be used to calculate the reliability of attribute k. W_{ik} can be calculated as

$$W_{ik} = p(X_i = 1|A_k = 1) - p(X_i = 1|A_k = 0), \qquad \text{(Equation 16)}$$

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

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

$$p(X_i = 1|A_k = 1) = \frac{p(A_k=1, X_i=1)}{p(A_k=1)}, \qquad \text{(Equation 17)}$$

where $p(A_k = 1, X_i = 1)$ is the joint probability that an examinee has attribute k and correctly answers item i, and $p(A_k = 1)$ is the marginal probability that an examinee has attribute k. To obtain $p(A_k = 1, X_i = 1)$ and $p(A_k = 1)$, the attribute patterns, the expected response patterns, and the estimate of the population probabilities for each of the expected response patterns must be specified.

The term $p(X_i = 1|A_k = 0)$ should be 0 because examinees are not expected to answer item i correctly since they lack attribute k required by item i. However, in an actual testing situation, it is possible that examinees can still answer the item correctly by guessing or by applying partial knowledge to reach

their solution, particularly when the multiple-choice item format is used. Therefore, $p(X_i = 1|A_k = 0)$ can also be fixed at a specific value (e.g., 0.20) that reflects a "pseudo-guessing" parameter.

The weights for attribute included in the hypothesized divergent hierarchy present in Figure 1b, for example, are as follows:

$$\begin{bmatrix} 0.37 & 0.24 & 0.12 & 0.15 & 0.10 & 0.02 \\ 0.00 & 0.50 & 0.25 & 0.00 & 0.21 & 0.05 \\ 0.00 & 0.00 & 0.84 & 0.00 & 0.00 & 0.16 \\ 0.00 & 0.00 & 0.00 & 0.55 & 0.37 & 0.08 \end{bmatrix} \quad \text{(Matrix 9)}$$

The first row of the Matrix 9 indicates that the weights of items 1 to 6 (columns 1 to 6) on attribute 1 are 0.37, 0.24, 0.12, 0.15, 0.10, and 0.02, respectively. The remaining rows can be interpreted in the same manner.

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

$$\alpha_{AHM_k} = \frac{n_k}{n_k - 1}\left[1 - \frac{\sum_{i \in S_k} W_{ik}^2 \sigma_{X_i}^2}{\sigma^2_{\sum_{i \in S_k} W_{ik} X_i}}\right], \quad \text{(Equation 18)}$$

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

observed score on the items that are measuring attribute k, and $\sigma^2_{\sum_{i \in S_k} W_{ik} X_i}$ is the variance of the weighted observed total scores.

The Spearman-Brown formula can also be used to evaluate the effect of changes to test length on the attribute reliability coefficient. The attribute-based Spearman-Brown formula is specified as

$$\alpha_{AHM-SB_k} = \frac{n_k \alpha_{AHM_k}}{1+(n_k-1)\alpha_{AHM_k}}, \qquad \text{(Equation 19)}$$

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

Chapter III: Methodology

The purpose of the present study was to evaluate reliability when the attribute hierarchy method (AHM) is used for cognitive diagnostic assessment (CDA). CDAs are designed to measure specific knowledge structures and processing skills to provide examinees with information about their cognitive strengths and weaknesses. The new adapted Cronbach' alpha, presented in Chapter 2, was compared to the standard form of Cronbach's alpha in terms of attribute reliability estimation. Specifically, the study was designed to answer the following four research questions:

1. Is attribute reliability as determined by adapted and standard Cronbach's alpha influenced by different cognitive model structures, specifically, a linear model versus a divergent model?
2. What is the minimum number of items required to measure each attribute to achieve adequate attribute reliability as determined by adapted and standard Cronbach's alpha?
3. Are the two indices influenced by sample size?
4. Is attribute reliability as determined by adapted and standard Cronbach's alpha influenced by discrepancy of examinees' observed response patterns from expected response patterns?

To answer research questions 1 to 4, a simulation study was conducted to evaluate and illustrate the two attribute reliability indices. Four factors expected to affect attribute reliability were manipulated in the simulation study: model

structure (linear and divergent types), test length (2, 4, 6, and 8 items per attribute), sample size (250, 500, 750, and 1000), and slip percentage (10, 15, 20, and 25%). The manipulated factors and the data generation procedure are described next.

Simulation Study

Factors to be Manipulated

A simulation study was conducted to evaluate attribute reliability by manipulating four factors that could directly affect CDA outcomes. The first factor is the structure of the cognitive model. Two models were evaluated, a linear and a divergent hierarchy. The number of attributes in each model was fixed at six to make the comparison between different model structures straightforward. The first cognitive model is a simple linear hierarchy. This model, as presented in Figure 3, contains all six attributes aligned in a single branch. This type of model could be used to characterize problem-solving when the knowledge and skills are ordered in a linear manner. Knowledge and skills that are restricted within a domain of basic logical application, for example, could be characterized in a linear model. In other words, the attributes in a linear model measure a single construct at varying difficulty levels. The second cognitive model was a more complex divergent hierarchy, as presented in Figure 4. This model contains two independent branches which share a common prerequisite, attribute 1. The first branch includes two additional attributes, 2 and 3, while the second branch includes a self-contained sub-hierarchy with

attributes 4 through 6. Two independent branches form the sub-hierarchy: attributes 4, 5 and attributes 4, 6. This type of model could be used to characterize problem-solving when the knowledge and skills differ as a function of the concepts and content within a domain. Examples of divergent model can be found in the study of Gierl et al. (2008). The four cognitive hierarchies they used to describe examinee performance on the SAT algebra subtest are presented in Figure 4. Attributes in each of the four hierarchies are organized in a divergent manner, indicating the knowledge and skills they measure are unrestricted within one domain. However, these attributes are categorized in one cognitive hierarchy because the knowledge and skills they measure have some characteristics in common. Taken together, these two model structures, linear and divergent, represent different types of cognitive structures that could characterize examinee performance on a diagnostic test.

The second factor was the number of items measuring each of the six attributes. Four different items sets were evaluated—two, four, six, and eight item sets—because test length (i.e., number of items contained in a test) is one factor known to affect reliability. The two item set yields a diagnostic test with 12 items, as each of the six attributes is measured by two items. Similarly, the four item set produces a diagnostic test with 24 items, the six-item set a test with 36 items, and the eight-item form a test with 48 items. Stated in another way, tests with different lengths were simulated in this study; a short 12-item test (2 item set), a short-to-moderate 24-item test (4 item set), a moderate 36-

item test (6 item set), and a long 48-item test (8 item set).

The third factor was the sample size involved in each simulated condition. Sample size was manipulated because the range of scores may differ in different size samples. If the variance does change, then the change will likely influence the estimates of attribute reliability. Therefore, four sample sizes were considered: 250, 500, 750, and 1000.

The fourth factor was the percentage of slips involved in the simulated responses. These slips represent the differences between the expected responses prescribed by the cognitive model in the E matrix and the actual responses produced by examinees. The discrepancies between expected and observed responses are common in practice because expected response patterns represent theoretically ideal responses. As an examinee solves an item, certain sources of random error might impact the examinee's test performance and result in a discrepancy between the expected and observed responses. Four slip levels were evaluated to produce different percentages of model-data misfit—10%, 15%, 20%, and 25%, meaning the difference between the expected and actual responses ranged from 10 to 25%.

Taken together, 256 conditions were assessed in the simulation study [i.e., (2) formula conditions * [(2) cognitive models * (4) items sets * (4) sample sizes * (4) slip conditions] = 256]. Each condition was simulated once as the possible variation involved in replication is on simulated response data. Attribute reliability estimation will not vary with multiple replications of each

condition. Therefore, observed response data was simulated with no replication.

Data Simulation

To generate the observed response data, a three-step process was used. First, the matrices of the AHM for each hierarchy—adjacency (A), reachability (R), incidence (Q), reduced incidence (Q_r), and expected response matrix (E)—were derived. The A matrix for the linear hierarchy shown in Figure 3 is:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{(Matrix 9)}$$

The R matrix for the linear hierarchy is:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{(Matrix 10)}$$

The Q matrix for the linear hierarchy is:

$$\begin{bmatrix} 1010101010101010101010101010101010101010101010101010101010101010 \\ 0110011001100110011001100110011001100110011001100110011001100110 \\ 0001111000011110000111100001111000011110000111100001111000011110 \\ 0000000111111110000000011111111000000001111111100000000111111110 \\ 0000000000000001111111111111111000000000000000011111111111111110 \\ 0000000000000000000000000000000111111111111111111111111111111110 \end{bmatrix}$$

(Matrix 11)

The Q_r matrix for the linear hierarchy is as follows:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{(Matrix 12)}$$

Lastly, the 2-item set expected response matrix for the linear model is:

$$\begin{bmatrix} 000000000000 \\ 110000000000 \\ 111100000000 \\ 111111000000 \\ 111111110000 \\ 111111111100 \\ 111111111111 \end{bmatrix}$$

(Matrix 13)

The 4-item set expected response matrix for the linear model is:

$$\begin{bmatrix} 000000000000000000000000 \\ 111100000000000000000000 \\ 111111110000000000000000 \\ 111111111110000000000000 \\ 111111111111111100000000 \\ 111111111111111111110000 \\ 111111111111111111111111 \end{bmatrix}$$

(Matrix 14)

The 6-item set expected response matrix for the linear model is:

$$\begin{bmatrix} 000000000000000000000000000000000000 \\ 111111000000000000000000000000000000 \\ 111111111111000000000000000000000000 \\ 111111111111111111000000000000000000 \\ 111111111111111111111111000000000000 \\ 111111111111111111111111111111000000 \\ 111111111111111111111111111111111111 \end{bmatrix}$$

(Matrix 15)

The 8-item set expected response matrix for the linear model is:

$$\begin{bmatrix} 000000000000000000000000000000000000000000000000 \\ 111111110000000000000000000000000000000000000000 \\ 111111111111111100000000000000000000000000000000 \\ 111111111111111111111111000000000000000000000000 \\ 111111111111111111111111111111110000000000000000 \\ 111111111111111111111111111111111111000000000000 \\ 111111111111111111111111111111111111111111111111 \end{bmatrix}$$

(Matrix 16)

The A matrix for the divergent hierarchy as shown in Figure 4 is:

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$ Matrix 17)

The R matrix for the divergent hierarchy is:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$ (Matrix 18)

The Q matrix for the divergent hierarchy is:

$$\begin{bmatrix} 1010101010101010101010101010101010101010101010101010101010101010101 \\ 0110011001100110011001100110011001100110011001100110011001100110011 \\ 0001111000011110000111100001111000011110000111100001111000011110000111 \\ 0000000111111110000000011111111000000001111111100000000111111110000000011111111 \\ 0000000000000000111111111111111100000000000000001111111111111111 \\ 0000000000000000000000000000000011111111111111111111111111111111 \end{bmatrix}$$

(Matrix 19)

The Q_r matrix for the divergent hierarchy is as follows:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$ (Matrix 20)

The 2-item set expected response matrix for the divergent model is:

$$\begin{bmatrix}
000000000000\\
110000000000\\
111100000000\\
111111000000\\
110000110000\\
111100110000\\
111111110000\\
110000111100\\
111100111100\\
111111111100\\
110000110011\\
111100110011\\
111111110011\\
110000111111\\
111100111111\\
111111111111
\end{bmatrix}$$

(Matrix 21)

The 4-item set expected response matrix for the divergent model is:

$$\begin{bmatrix}
000000000000000000000000\\
111100000000000000000000\\
111111110000000000000000\\
111111111111000000000000\\
111100000000111100000000\\
111111110000111100000000\\
111111111111111100000000\\
111100000000111111110000\\
111111110000111111110000\\
111111111111111111110000\\
111100000000111100001111\\
111111110000111100001111\\
111111111111111100001111\\
111100000000111111111111\\
111111110000111111111111\\
111111111111111111111111
\end{bmatrix}$$

(Matrix 22)

The 6-item set expected response matrix for the divergent model is:

$$\begin{bmatrix}
000000000000000000000000000000000000\\
111111000000000000000000000000000000\\
111111111111000000000000000000000000\\
111111111111111111000000000000000000\\
111111000000000000111111000000000000\\
111111111111000000111111000000000000\\
111111111111111111111111000000000000\\
111111000000000000111111111111000000\\
111111111111000000111111111111000000\\
111111111111111111111111111111000000\\
111111000000000000111111000000111111\\
111111111111000000111111000000111111\\
111111111111111111111111000000111111\\
111111000000000000111111111111111111\\
111111111111000000111111111111111111\\
111111111111111111111111111111111111
\end{bmatrix}$$

(Matrix 23)

The 8-item set expected response matrix for the divergent model is:

$$\begin{bmatrix}
00000000000000000000000000000000000000000000\\
11111111000000000000000000000000000000000000\\
11111111111111110000000000000000000000000000\\
11111111111111111111111100000000000000000000\\
11111111000000000000000011111111000000000000\\
11111111111111110000000011111111000000000000\\
11111111111111111111111111111111000000000000\\
11111111000000000000000011111111111111100000000\\
11111111111111110000000011111111111111100000000\\
11111111111111111111111111111111111111100000000\\
11111111000000000000000011111111000000011111111\\
11111111111111110000000011111111000000011111111\\
11111111111111111111111111111111000000011111111\\
11111111000000000000000011111111111111111111111\\
11111111111111110000000011111111111111111111111\\
11111111111111111111111111111111111111111111111
\end{bmatrix}$$

(Matrix 24)

Second, expected response data were generated for both linear and divergent hierarchies with four different sample size conditions. The expected response pattern matrix is used as the basis for generating expected response

data. The ability estimates for the expected response vectors were produced using maximum likelihood estimation. For both the linear and divergent hierarchies, expected response data were generated with the constraint that the ability estimates associated with the expected response vectors be normally distributed because the normal distribution reflects the common characteristic of ability estimates for population in most operational testing situations.

Third, slips were added to the expected response data to simulate the observed response patterns. The values of randomly added slips are 10%, 15%, 20%, and 25%. These slips are based on item probabilities calculated from each expected response vector using the 2-parameter logistic item response theory model,

$$p_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}} \qquad i = 1, 2, \ldots, n \qquad \text{(Equation 20)}$$

where $p_i(\theta)$ is the probabilty that a random examinee with ability θ answers item i correctly, a_i is the item i discrimination parameter, b_i is the item i difficulty parameter, θ is the ability level, n is the number of items in the test, e is a transcendental number whose value is 2.718, and the factor D is a scaling factor introduced to make the logistic funtion as close as possible to the normal ogive function. It has been shown that when $D = 1.7$, values of $p_i(\theta)$ for the two-parameter normal ogive and the two-parameter logistic models differ in absolute value by less than 0.01 for all values of θ. To simulate responses that most likely reflect the data in practice, the chance that a simulated examinee correctly answers an item by guessing was not considered in this study.

Therefore, a modified two-parameter logistic item response theory model was used in this study.

The item parameters used in the simulation study for both linear and divergent models are presented in Table 5. The discrimination parameter (a-parameter) is set at 1.0 across all items involved in both the linear and divergent models to represent items with acceptable discriminating power. The difficulty parameter (b-parameter) ranges from -2.50 to 2.50 with an increment of 1.0 for items measured in the linear model. This range indicates items measuring these six attributes in the linear model cover difficulty level from easy to hard. The difficulty parameters for items measured in the divergent model range from -2.50 to 2.50 with increment of 2.50. Again, the range covers items that are easy, moderate, and difficult.

Two types of slips were generated. First, slips were created for the subset of items expected to be answered incorrectly according to the attribute hierarchy (i.e., slips of the form 0 to 1). The percentage of these slips was specified as the item probability. Second, slips were created for the subset of items expected to be answered correctly according to the attribute hierarchy (i.e., slips of the form 1 to 0). The percentage of these slips was specified as one minus the item probability.

As an example, consider the divergent model in Figure 4 where two items are used to measure each attribute and 1,000 examinees. According to the frequency of normally distributed expected response vectors of the 1,000

examinees, 85 examinees are expected to have attribute pattern [100000] thereby producing the response pattern [110000000000]. The response probabilities for these items were computed using item parameters and the ability level associated with this attribute pattern. If one only considers the probabilities for item 1 ($a_1 = 1.0,\ b_1 = -2.5$) and item 3 ($a_2 = 1.0,\ b_2 = 0.0$) in the expected response pattern ($\theta = -1.485$), then the values of probabilities are 0.85 and 0.07, respectively. According to the distribution of expected response pattern, the 85 examinees are expected to answer the first item correctly. However, the probability of a correct response calculated from the 2-parameter logistic IRT model is 0.85 for item 1, indicating that although examinees have mastered the attributes required by the item, they still have 1 – 85% = 15% chance of producing an incorrect response. Therefore, 85 x 15% ≈ 13 response vectors associated with attribute pattern [100000] were randomly selected in the simulated data and changed from 1 to 0 for item 1. The 85 examinees are also expected to answer item 3 incorrectly. Using the item parameters and ability level estimate, the probability is 0.07 for item 3. This value suggests that although examinees are not expected to answer the item correctly based on their attribute mastery, they still have 7% chance of producing a correct response. As a result, 85 x 7% ≈ 6 slips of the form 0 to 1 were randomly introduced into the simulated data for item 3. Using this data generation procedure, four different slip percentages—10, 15, 20, and 25%—

were introduced into the expected response data producing observed responses that differed from the expected responses in the form of either 1 to 0 or 0 to 1.

Chapter IV: Results

The simulation study results are presented in Tables 6 to 13. Each of these tables contains two sub-tables, a and b. Sub-table a lists the results calculated using the adapted Cronbach's alpha coefficient for the AHM framework while sub-table b lists the results calculated using the standard Cronbach's alpha coefficient. Among these tables, Tables 6 to 9 list results for the linear model with four different sample size conditions, respectively, and tables 10 to 13 contain results for the divergent model with four different sample size conditions. Each table contains the reliability level for each of the six attributes as a function of the number of items measuring each attribute (2, 4, 6, 8) and the percentage of slips (10%, 15%, 20%, 25%). These results are presented in the order of increasing sample size.

Linear Cognitive Model with 250 Sample Size Condition for Adapted Formula

Table 6a contains the reliability estimates for the 250 sample size condition estimated by adapted Cronbach's alpha coefficient for the linear cognitive model. With two items measuring each attribute, the reliability estimate was highest for attribute 1 and lowest for attribute 6 in all slip conditions and decreased as the percentage of slips increased for each attribute. For example, with 10% error, the attribute reliability estimates differed from 0.83 for attribute 1 to 0.37 for attribute 6. Reliability estimates ranged from 0.81 for attribute 1 to 0.34 for attribute 6 with 15% error. With 20% error, the

outcome ranged from 0.78 for attribute 1 to 0.27 for attribute 6. With 25% error, reliability estimates varied from 0.74 for attribute 1 to 0.28 for attribute 6.

The same pattern was observed when the number of items per attribute was four. With four items measuring each attribute, the reliability estimate was again highest for attribute 1 and lowest for attribute 6 in all slip conditions. However, test length and model-data fit both clearly affected reliability as the outcomes in the four item condition were higher than the two item condition and decreased as the percentage of slips increased. With 10% error, reliability estimates varied from 0.91 for attribute 1 to 0.58 for attribute 6. With 15% error, the outcomes differed from 0.90 for attribute 1 to 0.43 for attribute 6. The reliability estimates ranged from 0.88 for attribute 1 to 0.37 for attribute 6 with 20% error. With 25% error, the reliability estimates changed from 0.86 for attribute 1 to 0.26 for attribute 6.

With six items per attribute, the reliability estimate was highest for attribute 1 and lowest for attribute 6, as in the previous item-set conditions. The importance of test length and model-data fit was apparent as the overall reliability outcomes in the six item condition were higher than the two or four item conditions and the outcomes in larger slip conditions were lower compared to those in smaller slip conditions. The reliability estimates ranged from 0.94 for attribute 1 to 0.70 for attribute 6 with 10% error. With 15% error, the reliability estimates varied from 0.93 for attribute 1 to 0.58 for attribute 6. With 20% error, the outcomes differed from 0.92 for attribute 1 to 0.53 for attribute 6.

The reliability estimates ranged from 0.90 for attribute 1 to 0.52 for attribute 6 with 25% error.

With eight items per attribute, reliability was observed to be highest for attribute 1 and lowest for attribute 6, as in the previous item-set conditions. Again, the importance of test length and model-data fit was apparent as the overall reliability outcomes in the eight item condition were higher than the two, four, or six item conditions and the outcomes in larger slip conditions were lower compared to those in smaller slip conditions. With 10% error, reliability ranged from 0.96 for attribute 1 to 0.76 for attribute 6. Reliability varied from 0.95 for attribute 1 to 0.64 for attribute 6 with 15% error while it changed from 0.94 for attribute 1 to 0.55 for attribute 6 with 20% error. With 25% error, reliability differed from 0.93 for attribute 1 to 0.55 for attribute 6.

To summarize, the same reliability estimates pattern was observed for different item-set conditions. That is, with two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attribute 6 across all slip conditions. Test length clearly affected reliability estimates. For example, the reliability estimates for attribute 1 with 10% slip for the two, four, six, and eight items conditions were 0.83, 0.91, 0.94, and 0.96, respectively; and for attribute 6, the corresponding estimates were 0.37, 0.58, 0.70, and 0.76. Similar patterns were observed for all the other attributes. Also, the reliability estimates decreased as the percentage of slips increased. For instance, the outcomes for two items measuring attribute

1 were 0.83, 0.81, 0.78, and 0.74 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Linear Cognitive Model with 250 Sample Size Condition for Standard Formula

Table 6b contains the reliability estimates for the 250 sample size condition estimated by Standard Cronbach's alpha coefficient for the linear cognitive model. As two items measuring each attribute, the reliability estimate was highest for attribute 1 and lowest for attribute 6 across all slip conditions and decreased as the percentage of slips increased for each attribute. For example, with 10% error, the attribute reliability estimates ranged from 0.85 for attribute 1 to 0.37 for attribute 6. The reliability varied from 0.83 for attribute 1 to 0.34 for attribute 6 with 15% error. With 20% error, the outcomes differed from 0.80 for attribute 1 to 0.27 for attribute 6. With 25% error, reliability ranged from 0.75 for attribute 1 to 0.28 for attribute 6.

The same pattern was observed for the other item-set conditions. That is, with two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attribute 6 across all slip conditions. Test length clearly affected reliability estimates. For example, the reliability estimates for attribute 1 with 10% slip for the two, four, six, and eight items conditions were 0.85, 0.92, 0.95, and 0.96, respectively; and for attribute 6, the corresponding estimates were 0.37, 0.58, 0.70, and 0.76. Similar patterns were observed for all the other attributes. In addition, the reliability estimates for each attribute decreased as the percentage of slips

increased in same item set condition. For instance, the outcomes for two items measuring attribute 2 were 0.84, 0.82, 0.77, and 0.71 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Linear Cognitive Model with 500 Sample Size Condition for Adapted Formula

Table 7a contains the reliability estimates for the 500 sample size condition estimated by adapted Cronbach's alpha coefficient for the linear cognitive model. With two items measuring each attribute, the reliability estimate for attribute 1 was highest and for attribute 6 was lowest in all slip conditions and decreased as the percentage of slips increased for each attribute. For example, with 10% error, the attribute reliability ranged from 0.84 for attribute 1 to 0.47 for attribute 6. The reliability altered from 0.81 for attribute 1 to 0.33 for attribute 6 with 15% error. With 20% error, the reliability differed from 0.79 for attribute 1 to 0.30 for attribute 6. With 25% error, the reliability ranged from 0.75 for attribute 1 to 0.26 for attribute 6.

Similar pattern was observed for the other item-set conditions. That is, with two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attribute 6 across all slip conditions. Test length clearly affected reliability estimates. For example, the reliability estimates for attribute 1 with 10% slip for the two, four, six, and eight items conditions were 0.84, 0.92, 0.94, and 0.96, respectively; and for attribute 6, the corresponding estimates were 0.47, 0.57, 0.71, and 0.75. Similar patterns were observed for all the other attributes. In addition, the

reliability estimates decreased as the percentage of slips increased. For instance, the outcomes for two items measuring attribute 2 were 0.82, 0.78, 0.74, and 0.69 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Linear Cognitive Model with 500 Sample Size Condition for Standard Formula

Table 7b contains the reliability estimates for the 500 sample size condition estimated by Standard Cronbach's alpha coefficient for the linear cognitive model. With two items measuring each attribute, the reliability estimate was highest for attribute 1 and lowest for attribute 6 in all slip conditions and decreased as the percentage of slips increased for each attribute. For example, with 10% error, the attribute reliability varied from 0.86 for attribute 1 to 0.47 for attribute 6. With 15% error, the reliability differed from 0.83 for attribute 1 to 0.33 for attribute 6. With 20% error, the reliability altered from 0.80 for attribute 1 to 0.30 for attribute 6. With 25% error, the reliability ranged from 0.76 for attribute 1 to 0.26 for attribute 6.

The same pattern was observed for the other item-set conditions. That is, with two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attribute 6 across all slip conditions. Test length is apparently a factor that affects reliability estimates. For example, the reliability estimates for attribute 1 with 10% slip for the two, four, six, and eight items conditions were 0.86, 0.93, 0.95, and 0.96, respectively; and for attribute 6, the corresponding estimates were 0.47, 0.57, 0.71, and 0.75. Similar patterns were observed for all the other attributes. In

addition, the reliability estimates for each attribute decreased as the percentage of slips increased in the same item set condition. For instance, the outcomes for four items measuring attribute 2 were 0.92, 0.90, 0.87, and 0.84 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Linear Cognitive Model with 750 Sample Size Condition for Adapted Formula

Table 8a contains the reliability estimates for the 750 sample size condition estimated by adapted Cronbach's alpha coefficient for the linear cognitive model. With two items measuring each attribute, the reliability outcome for attribute 1 was highest and for attribute 6 was lowest in all slip conditions and decreased as the percentage of slips increased for each attribute. For instance, with 10% error, the attribute reliability ranged from 0.84 for attribute 1 to 0.40 for attribute 6. The reliability differed from 0.81 for attribute 1 to 0.30 for attribute 6 with 15% error. With 20% error, the reliability altered from 0.78 for attribute 1 to 0.18 for attribute 6. With 25% error, the reliability varied from 0.75 for attribute 1 to 0.20 for attribute 6.

The same pattern was observed for the other item-set conditions. That is, with two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attribute 6 across all slip conditions. Test length is apparently a factor that affects reliability estimates. For example, the reliability estimates for attribute 1 with 10% slip for the two, four, six, and eight items conditions were 0.84, 0.92, 0.94, and 0.96, respectively; and for attribute 6, the corresponding estimates were 0.40, 0.59,

0.68, and 0.76. Similar patterns were observed for all the other attributes. In addition, the reliability estimates decreased as the percentage of slips increased. For instance, the outcomes for six items measuring attribute 4 were 0.89, 0.86, 0.81, and 0.76 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Linear Cognitive Model with 750 Sample Size Condition for Standard Formula

Table 8b contains the reliability estimates for the 750 sample size condition estimated by Standard Cronbach's alpha coefficient for the linear cognitive model. With two items measuring each attribute, the reliability estimate was highest for attribute 1 and lowest for attribute 6 in all slip conditions and decreased as the percentage of slips increased for each attribute. With 10% error, the attribute reliability ranged from 0.85 for attribute 1 to 0.40 for attribute 6. With 15% error, the reliability differed from 0.83 for attribute 1 to 0.30 for attribute 6. The reliability altered from 0.80 for attribute 1 to 0.18 for attribute 6 with 20% error. With 25% error, the reliability ranged from 0.76 for attribute 1 to 0.20 for attribute 6.

The same pattern was observed for the other item-set conditions. That is, with two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attribute 6 across all slip conditions. Test length clearly affected reliability estimates. For example, the reliability estimates for attribute 1 with 20% slip for the two, four, six, and eight items conditions were 0.80, 0.89, 0.92, and 0.94, respectively; and for attribute 6, the corresponding estimates were 0.18, 0.35, 0.48, and 0.57.

Similar patterns were observed for all the other attributes. In addition, the reliability estimates for each attribute decreased as the percentage of slips increased in the same item set condition. For instance, the outcomes for six items measuring attribute 5 were 0.86, 0.79, 0.72, and 0.63 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Linear Cognitive Model with 1000 Sample Size Condition for Adapted Formula

Table 9a contains the reliability estimates for the 1000 sample size condition estimated by adapted Cronbach's alpha coefficient for the linear cognitive model. With two items measuring each attribute, the reliability estimate was highest for attribute 1 and lowest for attribute 6 in all slip conditions and decreased as the percentage of slips increased for each attribute. With 10% error, the attribute reliability ranged from 0.84 for attribute 1 to 0.49 for attribute 6. The reliability differed from 0.81 for attribute 1 to 0.25 for attribute 6 with 15% error. With 20% error, the reliability varied from 0.78 for attribute 1 to 0.23 for attribute 6. With 25% error, the reliability altered from 0.75 for attribute 1 to 0.18 for attribute 6.

Similar patterns were observed for the other item-set conditions. That is, with two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attribute 6 across all slip conditions. Test length is apparently a factor that affects reliability estimates. For example, the reliability estimates for attribute 1 with 15% slip for the two, four, six, and eight items conditions were 0.81, 0.90, 0.93, and 0.95,

respectively; and for attribute 6, the corresponding estimates were 0.25, 0.45, 0.57, and 0.64. Similar patterns were observed for all the other attributes. In addition, the reliability estimates decreased as the percentage of slips increased. For instance, the outcomes for four items measuring attribute 3 were 0.88, 0.85, 0.82, and 0.77 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Linear Cognitive Model with 1000 Sample Size Condition for Standard Formula

Table 9b contains the reliability estimates for the 1000 sample size condition estimated by Standard Cronbach's alpha coefficient for the linear cognitive model. With two items measuring each attribute, the reliability estimate was highest for attribute 1 and lowest for attribute 6 in all slip conditions and decreased as the percentage of slips increased for each attribute. With 10% error, the attribute reliability ranged from 0.86 for attribute 1 to 0.49 for attribute 6. With 15% error, the reliability differed from 0.83 for attribute 1 to 0.25 for attribute 6. The reliability altered from 0.80 for attribute 1 to 0.23 for attribute 6 with 20% error. With 25% error, the reliability varied from 0.75 for attribute 1 to 0.18 for attribute 6.

The same pattern was observed for the other item-set conditions. With two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attribute 6 across all slip conditions. Test length is apparently a factor that affects reliability estimates. For example, the reliability estimates for attribute 1 with 20% slip for the two, four, six, and eight items conditions were 0.80, 0.89, 0.92, and 0.94, respectively;

and for attribute 6, the corresponding estimates were 0.23, 0.39, 0.48, and 0.57. Similar patterns were observed for all the other attributes. Moreover, the reliability estimates for each attribute decreased as the percentage of slips increased in the same item set condition. For instance, the outcomes for four items measuring attribute 2 were 0.92, 0.90, 0.87, and 0.84 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Summary for Conditions of the Linear Cognitive Model

It was concluded that for the linear model the attribute reliability estimate derived from both the adapted and standard Cronbach's alpha coefficients was highest for attribute 1 and lowest for attribute 6 across all item sets and slip conditions. The test length and model-data fit were apparently the factors that influenced the attribute reliability estimates. The longer the test and the smaller the number of the slips, the higher the reliability estimates. However, as the sample size increased across each slip and item set condition, attribute reliability estimates were quite similar to one another.

Noticeably, for attribute reliability estimates in the same slip, sample size, and formula condition, differences between consecutive estimates became progressively smaller as the number of items on each attribute increased. For example, the change between the two and four item sets in the condition of 250 sample size, 10% slip percentage, and the adapted formula was 0.08, between the four and six item sets, 0.03, and between the six and eight item sets, 0.02,

for attribute 1. Similar patterns were observed for all other attributes, which reflects the behavior of the Spearman-Brown formula.

The reliability estimates derived from the adapted and standard Cronbach's alpha were also compared. Table 14 presents correlation coefficients between attribute reliability estimates for the linear model derived by the adapted and standard Cronbach's alpha coefficients for each sample size and slip percentage condition. These correlation coefficients range from 0.99 to 1.00, indicating essentially perfect fit between reliability result sets estimated by the two formulas. However, for each sample size-slip percentage-item set condition, the reliability estimates derived by the standard Cronbach's alpha coefficient were slightly higher than those estimated by the adapted Cronbach's alpha coefficient. Table 15 presents the root mean square deviation between the attribute reliability estimates for the linear model derived by the adapted and standard Cronbach's alpha coefficient for each sample size, slip percentage, and item set condition. These values range from 0.01 to 0.06, indicating negligible differences between the reliability results estimated by the two formulas.

Divergent Cognitive Model with 250 Sample Size Condition for Adapted Formula

Table 10a contains the reliability estimates for the 250 sample size condition estimated by adapted Cronbach's alpha coefficient for the divergent cognitive model. Recall, the divergent model is more complex than the linear model because it contains two independent branches which share a common prerequisite, attribute 1. The first branch includes two attributes, 2 and 3, while

the second branch forms two sub-hierarchies consisting of attributes 4 and 5 and 4 and 6. With two items measuring each attribute, the reliability estimate was highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy (i.e., attributes 3, 5, and 6) in all slip conditions and decreased as the percentage of slips increased for each attribute. For instance, with 10% error, the reliability estimate for branch one (i.e., attributes 1, 2, and 3) ranged from 0.81 for attribute 1 to 0.69 for attribute 3. For branch two (i.e., attributes 4, 5, and 6), the reliability was 0.76 for attribute 4 and 0.71 and 0.63 for attributes 5 and 6, respectively. With 15% error, the reliability for branch one ranged from 0.82 for attribute 1 to 0.56 for attribute 3. For branch two, the outcome was 0.74 for attribute 4 and 0.57 and 0.53 for attributes 5 and 6, respectively. With 20% error, the reliability for branch one altered from 0.80 for attribute 1 to 0.45 for attribute 3. For branch two, the reliability was 0.68 for attribute 4 and 0.52 and 0.47 for attributes 5 and 6, respectively. With 25% error, the reliability for branch one varied from 0.80 for attribute 1 to 0.43 for attribute 3. For branch two, the reliability was 0.67 for attribute 4 and 0.45 and 0.42 for attributes 5 and 6, respectively.

The same pattern was observed for the other item-set conditions. With two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy across all slip conditions. As with the linear model analysis, test length is apparently a factor that affects reliability estimates. For

example, the reliability estimates for attribute 1 with 10% slip for the two, four, six, and eight items conditions were 0.81, 0.90, 0.93, and 0.95, respectively; and for attribute 6, the corresponding estimates were 0.63, 0.81, 0.86, and 0.90. Similar patterns were observed for all the other attributes. Moreover, the reliability estimates for each attribute decreased as the percentage of slips increased in the same item set condition. For instance, the outcomes for four items measuring attribute 5 were 0.82, 0.74, 0.61, and 0.55 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Divergent Cognitive Model with 250 Sample Size Condition for Standard Formula

Table 10b contains the reliability estimates for the 250 sample size condition estimated by Standard Cronbach's alpha coefficient for the divergent cognitive model. With two items measuring each attribute, the reliability estimate was highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy (i.e., attributes 3, 5, and 6) in all slip conditions. With 10% error, the reliability for branch one (i.e., attributes 1, 2, and 3) ranged from 0.81 for attribute 1 to 0.69 for attribute 3. For branch two (i.e., attributes 4, 5, and 6), the reliability was 0.77 for attribute 4 and 0.71 and 0.63 for attributes 5 and 6, respectively. With 15% error, the reliability for branch one varied from 0.82 for attribute 1 to 0.56 for attribute 3. For branch two, the reliability was 0.75 for attribute 4 and 0.57 and 0.53 for attributes 5 and 6, respectively. With 20% error, the reliability for branch one altered from 0.79 for attribute 1 to 0.45 for attribute 3. For branch two, the reliability was 0.69 for attribute 4 and 0.52 and

0.47 for attributes 5 and 6, respectively. With 25% error, the reliability for branch one differed from 0.79 for attribute 1 to 0.43 for attribute 3. For branch two, the reliability was 0.68 for attribute 4 and 0.45 and 0.42 for attributes 5 and 6, respectively.

The same pattern was observed for the other item-set conditions. With two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy across all slip conditions. As with the linear model analysis, test length is apparently a factor that affects reliability estimates. For example, the reliability estimates for attribute 1 with 15% slip for the two, four, six, and eight items conditions were 0.82, 0.90, 0.93, and 0.95, respectively; and for attribute 6, the corresponding estimates were 0.53, 0.73, 0.79, and 0.84. Similar patterns were observed for all the other attributes. Moreover, the reliability estimates for each attribute decreased as the percentage of slips increased in the same item set condition. For instance, the outcomes for six items measuring attribute 3 were 0.86, 0.78, 0.68, and 0.65 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Divergent Cognitive Model with 500 Sample Size Condition for Adapted Formula

Table 11a contains the reliability estimates for the 500 sample size condition estimate by adapted Cronbach's alpha coefficient for the divergent cognitive model. With two items measuring each attribute, the reliability was highest for attribute 1 and lowest for attributes in the final nodes of the

hierarchy (i.e., attributes 3, 5, and 6) in all slip conditions. For example, with 10% error, the reliability for branch one (i.e., attributes 1, 2, and 3) ranged from 0.80 for attribute 1 to 0.71 for attribute 3. For branch two (i.e., attributes 4, 5, and 6), the reliability was 0.76 for attribute 4 and 0.73 and 0.71 for attributes 5 and 6, respectively. With 15% error, the reliability for branch one altered from 0.81 for attribute 1 to 0.61 for attribute 3. For branch two, the reliability was 0.73 for attribute 4 and 0.57 and 0.58 for attributes 5 and 6, respectively. With 20% error, the reliability for branch one varied from 0.80 for attribute 1 to 0.25 for attribute 3. For branch two, the reliability was 0.69 for attribute 4 and 0.43 and 0.42 for attributes 5 and 6, respectively. With 25% error, the reliability for branch one differed from 0.80 for attribute 1 to 0.48 for attribute 3. For branch two, the reliability was 0.66 for attribute 4 and 0.36 and 0.37 for attributes 5 and 6, respectively.

A similar pattern was observed for the other item-set conditions. With two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy across all slip conditions. As with the linear model analysis, test length clearly affected reliability estimates. For example, the reliability estimates for attribute 1 with 25% slip for the two, four, six, and eight items conditions were 0.80, 0.89, 0.93, and 0.94, respectively; and for attribute 6, the corresponding estimates were 0.37, 0.57, 0.70, and 0.75. The same patterns were observed for all the other attributes. In addition, the reliability

estimates for each attribute decreased as the percentage of slips increased in the same item set condition. For instance, the outcomes for two items measuring attribute 4 were 0.76, 0.73, 0.69, and 0.66 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Divergent Cognitive Model with 500 Sample Size Condition for Standard Formula

Table 11b contains the reliability estimates for the 500 sample size condition estimated by Standard Cronbach's alpha coefficient for the divergent cognitive model. With two items measuring each attribute, the reliability was highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy (i.e., attributes 3, 5, and 6) in all slip conditions. With 10% error, the reliability for branch one (i.e., attributes 1, 2, and 3) ranged from 0.80 for attribute 1 to 0.71 for attribute 3. For branch two (i.e., attributes 4, 5, and 6), the reliability was 0.77 for attribute 4 and 0.73 and 0.71 for attributes 5 and 6, respectively. With 15% error, the reliability for branch one altered from 0.81 for attribute 1 to 0.61 for attribute 3. For branch two, the reliability was 0.74 for attribute 4 and 0.54 and 0.58 for attributes 5 and 6, respectively. With 20% error, the reliability for branch one varied from 0.80 for attribute 1 to 0.52 for attribute 3. For branch two, the reliability was 0.70 for attribute 4 and 0.43 and 0.42 for attributes 5 and 6, respectively. With 25% error, the reliability for branch one differed from 0.79 for attribute 1 to 0.48 for attribute 3. For branch two, the reliability was 0.68 for attribute 4 and 0.36 and 0.37 for attributes 5 and 6, respectively.

The same pattern was observed for the other item-set conditions. With two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy across all slip conditions. As with the linear model analysis, test length is apparently a factor that affects reliability estimates. For example, the reliability estimates for attribute 1 with 10% slip for the two, four, six, and eight items conditions were 0.80, 0.90, 0.93, and 0.95, respectively; and for attribute 6, the corresponding estimates were 0.71, 0.83, 0.87, and 0.90. Similar patterns were observed for all the other attributes. In addition, the reliability estimates for each attribute decreased as the percentage of slips increased in the same item set condition. For instance, the outcomes for four items measuring attribute 3 were 0.82, 0.72, 0.63, and 0.57 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Divergent Cognitive Model with 750 Sample Size Condition for Adapted Formula

Table 12a contains the reliability estimates for the 750 sample size condition estimate by adapted Cronbach's alpha coefficient for the divergent cognitive model. With two items measuring each attribute, the reliability was highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy (i.e., attributes 3, 5, and 6) in all slip conditions. With 10% error, the reliability for branch one (i.e., attributes 1, 2, and 3) ranged from 0.80 for attribute 1 to 0.69 for attribute 3. For branch two (i.e., attributes 4, 5, and 6), the reliability was 0.76 for attribute 4 and 0.68 and 0.71 for attributes 5 and 6,

respectively. With 15% error, the reliability for branch one varied from 0.80 for attribute 1 to 0.57 for attribute 3. For branch two, the reliability was 0.71 for attribute 4 and 0.48 and 0.56 for attributes 5 and 6, respectively. With 20% error, the reliability for branch one altered from 0.80 for attribute 1 to 0.46 for attribute 3. For branch two, the reliability was 0.67 for attribute 4 and 0.39 and 0.47 for attributes 5 and 6, respectively. With 25% error, the reliability for branch one ranged from 0.79 for attribute 1 to 0.49 for attribute 3. For branch two, the reliability was 0.65 for attribute 4 and 0.34 and 0.49 for attributes 5 and 6, respectively.

Similar patterns were observed for the other item-set conditions. With two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy across all slip conditions. Apparently, test length is a factor that affects reliability estimates. For example, the reliability estimates for attribute 1 with 20% slip for the two, four, six, and eight items conditions were 0.80, 0.89, 0.93, and 0.94, respectively; and for attribute 6, the corresponding estimates were 0.47, 0.60, 0.70, and 0.77. Similar patterns were observed for all the other attributes. In addition, the reliability estimates for each attribute decreased as the percentage of slips increased in the same item set condition. For instance, the outcomes for eight items measuring attribute 3 were 0.89, 0.82, 0.75, and 0.74 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Divergent Cognitive Model with 750 Sample Size Condition for Standard Formula

Table 12b contains the reliability estimates for the 750 sample size condition estimated by Standard Cronbach's alpha coefficient for the divergent cognitive model. With two items measuring each attribute, the reliability was highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy (i.e., attributes 3, 5, and 6) in all slip conditions. With 10% error, the reliability for branch one (i.e., attributes 1, 2, and 3) ranged from 0.81 for attribute 1 to 0.69 for attribute 3. For branch two (i.e., attributes 4, 5, and 6), the reliability was 0.77 for attribute 4 and 0.68 and 0.71 for attributes 5 and 6, respectively. With 15% error, the reliability for branch one varied from 0.80 for attribute 1 to 0.57 for attribute 3. For branch two, the reliability was 0.73 for attribute 4 and 0.48 and 0.56 for attributes 5 and 6, respectively. With 20% error, the reliability for branch one altered from 0.79 for attribute 1 to 0.46 for attribute 3. For branch two, the reliability was 0.69 for attribute 4 and 0.39 and 0.47 for attributes 5 and 6, respectively. With 25% error, the reliability for branch one differed from 0.79 for attribute 1 to 0.49 for attribute 3. For branch two, the reliability was 0.67 for attribute 4 and 0.34 and 0.49 for attributes 5 and 6, respectively.

Similar patterns were observed for the other item-set conditions. With two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy across all slip conditions. Test length is a factor that

affects reliability estimates. For example, the reliability estimates for attribute 1 with 25% slip for the two, four, six, and eight items conditions were 0.79, 0.88, 0.92, and 0.94, respectively; and for attribute 6, the corresponding estimates were 0.49, 0.58, 0.68, and 0.75. Similar patterns were observed for all the other attributes. Moreover, the reliability estimates for each attribute decreased as the percentage of slips increased in the same item set condition. For instance, the outcomes for four items measuring attribute 3 were 0.81, 0.69, 0.60, and 0.56 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Divergent Cognitive Model with 1000 Sample Size Condition for Adapted Formula

Table 13a contains the reliability estimate for the 1000 sample size condition estimate by adapted Cronbach's alpha coefficient for the divergent cognitive model. With two items measuring each attribute, the reliability was highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy (i.e., attributes 3, 5, and 6) in all slip conditions. With 10% error, the reliability for branch one (i.e., attributes 1, 2, and 3) varied from 0.80 for attribute 1 to 0.67 for attribute 3. For branch two (i.e., attributes 4, 5, and 6), the reliability was 0.76 for attribute 4 and 0.67 and 0.71 for attributes 5 and 6, respectively. With 15% error, the reliability for branch one ranged from 0.80 for attribute 1 to 0.51 for attribute 3. For branch two, the reliability was 0.71 for attribute 4 and 0.52 and 0.58 for attributes 5 and 6, respectively. With 20% error, the reliability for branch one changed from 0.80 for attribute 1 to 0.44 for attribute 3. For branch two, the reliability was 0.68 for attribute 4 and 0.44 and

0.47 for attributes 5 and 6, respectively. With 25% error, the reliability for branch one altered from 0.79 for attribute 1 to 0.42 for attribute 3. For branch two, the reliability was 0.64 for attribute 4 and 0.40 and 0.42 for attributes 5 and 6, respectively.

The same pattern was observed for the other item-set conditions. With two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy across all slip conditions. Similar as with the linear model analysis, test length apparently affects reliability estimates. For example, the reliability estimates for attribute 1 with 15% slip for the two, four, six, and eight items conditions were 0.80, 0.89, 0.93, and 0.95, respectively; and for attribute 6, the corresponding estimates were 0.58, 0.70, 0.79, and 0.82. Similar patterns were observed for all the other attributes. In addition, the reliability estimates for each attribute decreased as the percentage of slips increased in the same item set condition. For instance, the outcomes for two items measuring attribute 3 were 0.80, 0.69, 0.60, and 0.56 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Divergent Cognitive Model with 1000 Sample Size Condition for Standard Formula

Table 13b contains the reliability estimate for the 1000 sample size condition estimated by Standard Cronbach's alpha coefficient for the divergent cognitive model. With two items measuring each attribute, the reliability was

highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy (i.e., attributes 3, 5, and 6) in all slip conditions. With 10% error, the reliability for branch one (i.e., attributes 1, 2, and 3) varied from 0.80 for attribute 1 to 0.67 for attribute 3. For branch two (i.e., attributes 4, 5, and 6), the reliability was 0.77 for attribute 4 and 0.67 and 0.71 for attributes 5 and 6, respectively. With 15% error, the reliability for branch one altered from 0.80 for attribute 1 to 0.51 for attribute 3. For branch two, the reliability was 0.73 for attribute 4 and 0.52 and 0.58 for attributes 5 and 6, respectively. With 20% error, the reliability for branch one changed from 0.80 for attribute 1 to 0.44 for attribute 3. For branch two, the reliability was 0.70 for attribute 4 and 0.44 and 0.47 for attributes 5 and 6, respectively. With 25% error, the reliability for branch one differed from 0.78 for attribute 1 to 0.42 for attribute 3. For branch two, the reliability was 0.66 for attribute 4 and 0.40 and 0.42 for attributes 5 and 6, respectively.

Similar patterns were observed for the other item-set conditions. With two, four, six, and eight items measuring each attribute, the reliability estimate was consistently highest for attribute 1 and lowest for attributes in the final nodes of the hierarchy across all slip conditions. Test length, again, is a factor that affects reliability estimates. For example, the reliability estimates for attribute 1 with 20% slip for the two, four, six, and eight items conditions were 0.80, 0.89, 0.92, and 0.94, respectively; and for attribute 6, the corresponding estimates were 0.47, 0.61, 0.71, and 0.76. The same patterns were observed for

all the other attributes. Moreover, the reliability estimates for each attribute decreased as the percentage of slips increased in the same item set condition. For instance, the outcomes for four items measuring attribute 4 were 0.88, 0.85, 0.83, and 0.80 for 10%, 15%, 20%, and 25% slip conditions, respectively.

Summary for Conditions of the Divergent Cognitive Model

It was concluded that for the divergent model, attribute reliability estimated by both the adapted and standard Cronbach's alpha coefficients was highest for attribute 1 and lowest for attributes in the final nodes, i.e., attributes A3, A5, and A6, across all item sets and slip conditions. The test length and model-data fit were apparently the factors that influenced the attribute reliability estimates. The longer the test and the smaller the slip, the higher the reliability estimates. However, as the sample size increased across each slip and item set condition, attribute reliability estimates were quite similar to one another.

Noticeably, for attribute reliability estimates calculated by either formula in the same slip and sample size condition, differences between consecutive estimates became progressively smaller as the number of items on each attribute increased. For example, the change between the two and four item sets in the condition of 1000 sample size, 10% slip percentage, and the adapted formula was 0.10, between the four and six item sets, 0.03, and between the six and eight item sets, 0.02, for attribute 1. Similar patterns were observed for all other attributes which reflect the behavior of the Spearman-Brown formula.

The reliability estimates derived from the adapted and standard Cronbach's alpha were compared. Table 16 presents correlation coefficients between attribute reliability estimates for the divergent model derived by the adapted and standard Cronbach's alpha coefficients for each sample size and slip percentage condition. These correlation coefficients range from 0.99 to 1.00, indicating essentially perfect fit between reliability result sets estimated by the two formulas. However, for each sample size-slip percentage-item set condition, the reliability estimates derived by the standard Cronbach's alpha coefficient were slightly higher than those estimated by the adapted Cronbach's alpha coefficient. Table 17 presents root mean square deviation between attribute reliability estimates for the divergent model derived by adapted and standard Cronbach's alpha coefficient for each sample size, slip percentage, and item set condition. These values ranged from 0.00 to 0.02, indicating negligible differences between the values of the two reliability estimates.

Summary of both the Linear and Divergent Models

Overall, the attribute reliability for both the linear and divergent models, estimated by either the adapted or the standard Cronbach's alpha formulas, is systematically affected by test length and slip percentage but not by sample size. By comparing the results between two model structures, the variation of the six reliability estimates calculated by both adapted and standard formula, is noticeably and consistently larger for the linear model compared to the

divergent model. This outcome can be accounted for by the attribute dependencies inherent in each hierarchy with different structures.

Another noticeable outcome is that although reliability estimates derived by the standard Cronbach's alpha coefficient for some conditions were slightly greater than those estimated by the adapted Cronbach's alpha coefficient, the reliability estimates for the majority of conditions differed not at all or by a very small amount. However, the ways in which the two formulas take into consideration examinees who obtain the correct answer but without possessing the required attributes differ in that the adapted formula fully accounted for these examinees while the standard formula did not. For example, in the linear model with six attributes, the weights were assigned to each attribute in a way that respected each item's contribution to the measurement of that attribute. In contrast, if attribute reliability is estimated by standard formula, each of the six attributes are considered to completely contribute to the performance of the item that measures examined attribute. Therefore, the adapted Cronbach's alpha formula is more conceptually meaningful than the standard formula.

Chapter V: Discussion and Conclusions

The increasing demand for providing diagnostic feedback about examinees' cognitive proficiency has encouraged measurement specialists to probe new methods of evaluating and interpreting examinees' performance. Cognitive diagnostic assessment is one means that can be used to support specific inferences about examinees' cognitive strengths and weaknesses. Cognitive diagnostic assessment is a test form that employs a cognitive model to develop or identify items measuring specific knowledge and skills and to direct the psychometric analyses of examinees' item response pattern to promote diagnostic inferences. The cognitive model plays a foundational role as it provides a representation of the knowledge structures and processing skills that are believed to underlie conceptual understanding in a particular domain.

In 2004, Leighton, Gierl, and Hunka proposed a procedure for cognitive diagnostic assessment, the attribute hierarchy method (AHM), as a way to link psychological principles with psychometric procedures. With the assumption that test performance is associated with a specific set of hierarchically-organized attributes, the AHM is a cognitively-based psychometric method that classifies examinees' test item responses into structured attribute patterns according to a cognitive model of task performance. These attributes are structured using a hierarchy in which the ordering of the cognitive skills is specified. As a result, the attribute hierarchy serves as an explicit construct-centered cognitive model because it represents the psychological ordering among the cognitive attributes

required to solve test items. This model, in turn, provides an explicit, fine-grained interpretative framework for designing test items and for linking examinees' test performance to specific inferences about psychological skill acquisition.

The issue of reliability estimation is critical in cognitive diagnostic assessment because it concerns precision of decisions about examinees' specific cognitive strengths and weaknesses. In the AHM, attribute reliability refers to the consistency of the decisions made in a diagnostic test about examinees' mastery of specific attributes across multiple items that measure each attribute. Attribute reliability is a fundamental concept in CDA because it reflects score precision that should be reported to test users. To date, however, the reliability for attribute-based scoring for AHM has not been established. The present study was designed to introduce and evaluate an analytic procedure for assessing attribute reliability for cognitive diagnostic assessments and to explore factors that influence attribute reliability.

This chapter includes five sections. First, the research questions are revisited together with a brief summary of the methods used for the present study. A summary and discussion of the results are then presented. The limitations of the study are discussed followed by the presentation of the conclusions drawn from the results. Lastly, the educational and practical implications from the study and recommendations for future research directions are outlined.

Restatement of Research Questions and Summary of Methods

The purpose of this study was to evaluate attribute reliability for cognitive diagnostic assessments so examinees could receive precise information about their cognitive problem-solving strengths and weaknesses. The customized feedback for each examinee would, on one hand, guide examinees' individual learning, and on the other hand, enhance teachers' instruction. The accuracy and consistency by which cognitive diagnostic assessments can classify examinees' observed response patterns is, therefore, key to providing diagnostic feedback about examinees' cognitive proficiencies. Hence, attribute-based reliability that considers the consistency of decision about examinees' cognitive proficiencies was conducted and evaluated in the present study. Two reliability estimation procedures were considered: the standard form of Cronbach's alpha (Cronbach, 1951) and an adapted form in which each item is weighted to take account of those examinees who correctly answer the item but do not possess the attributes needed to correctly answer the item. For each method, the factors expected to influence attribute reliability estimate were studied and explored. The research questions addressed in this study include:

1. Is attribute reliability as determined by adapted and standard Cronbach's alpha influenced by different cognitive model structures, specifically, a linear model versus a divergent model?

2. What is the minimum number of items required to measure each attribute to achieve adequate attribute reliability as determined by adapted and standard Cronbach's alpha?
3. Are the two indices influenced by sample size?
4. Is attribute reliability as determined by adapted and standard Cronbach's alpha influenced by discrepancy of examinees' observed response patterns from expected response patterns?

To answer the research questions, a simulation study was conducted. The four factors identified in the research questions were manipulated for each method. To answer the first question, two types of model structures were evaluated, a linear and a divergent hierarchy. The number of attributes in each model is fixed at six to make the comparison between different model structures more interpretable. To answer the second question, tests with different numbers of items were simulated to create tests with different lengths. Four levels were created: a short 12-item test, a short-to-moderate 24-item test, a moderate 36-item test, and a long 48-item test. To answer the third question, four different sample sizes (250, 500, 750, and 1000) were used for simulation to represent a relatively small scope of sample size that can occur in certain testing situations such as provincial achievement tests. To answer the fourth question, four levels of discrepancies between expected and observed responses were simulated: 10%, 15%, 20%, and 25% to represent different percentages of model-data misfit.

Results and Discussion

Impact of the Adaptation of Cronbach's alpha

Although reliability estimates derived by the standard Cronbach's alpha coefficient for some conditions were slightly greater than those estimated by the adapted Cronbach's alpha coefficient, the reliability estimates for the majority of conditions differed not at all or by a very small amount. However, the ways in which the two formulas take into consideration examinees who obtain the correct answer but without possessing the required attributes differ in that the adapted formula fully accounted for these examinees while the standard formula did not. For example, in the linear model with six attributes, the weights were assigned to each attribute in a way that respected each item's contribution to the measurement of that attribute. In contrast, if attribute reliability is estimated by standard formula, each of the six attributes are considered to completely contribute to the performance of the item that measures examined attribute. Therefore, the adapted Cronbach's alpha formula is more conceptually meaningful than the standard formula.

Impact of Model Structure

For both linear and divergent models, attribute reliability, estimated by both the adapted and standard Cronbach's alpha coefficients for all sample sizes, was highest for attribute 1 and lowest for attributes at the final nodes of the hierarchy (i.e., attribute 6 in the linear model and attributes 3, 5, and 6 in the divergent model) across all item sets and slip conditions. However, the variation

of the six reliability estimates, calculated by both adapted and standard formulas, is noticeably and consistently larger for the linear model compared to the divergent model. Therefore, model structure was a factor that influences attribute reliability estimates.

This outcome can be accounted for by the attribute dependencies inherent in each hierarchy. For instance, in the linear hierarchy with two items per attribute, attribute 1 was directly or indirectly measured by 12 items whereas attribute 6 was directly measured by only two items because of the model structure. For the divergent hierarchy with two items per attribute, attribute 1 was directly or indirectly measured by 12 items whereas attributes 3, 5, and 6 were directly measured by only two items. From this example it becomes clear that the divergent model contains attributes with fewer dependencies and, thus, each attribute is affected by a smaller number of items, either directly or indirectly. As a result, there is less variation among the reliability estimates with the divergent model across all study conditions when compared with the linear model.

Impact of Test Length

Test length showed a consistent positive effect on attribute reliability estimates in simulated conditions with different model structures, sample sizes, and slip percentages for both reliability calculation formulas. As the test length increased across each slip and sample size condition for both hierarchies, attribute reliability estimates calculated by both formulas increased. That is, the

longer the test, the higher the reliability estimates regardless of the other factors considered. Therefore, test length was clearly a factor that influences attribute reliability estimates. This finding reflects the Spearman-Brown finding where parallel items are added to increase reliability estimates.

Impact of Sample Size

The factor of sample size did not affect on attribute reliability estimates in simulated conditions with different item sets, model structures, slip percentages, and reliability calculation formulas. As the sample size increased, attribute reliability estimates for each item set-hierarchy-slip percentage-formula condition were quite similar to one another. For example, as the sample size ranges from 250 to 1,000, the reliability estimates for attributes 1 to 6 in linear model, calculated by adapted formula, were (0.81, 0.79, 0.74, 0.67, 0.45, and 0.34), (0.81, 0.78, 0.73, 0.65, 0.49, and 0.33), (0.81, 0.79, 0.73, 0.65, 0.48, and 0.30), and (0.81, 0.78, 0.73, 0.64, 0.49, and 0.25) in 2-item set condition with 15% slips. Therefore, sample size was not a factor that influences attribute reliability estimates. The lack of difference in variability is attributable to the selection of the samples from the sample population.

Impact of Discrepancy between Expected and Observed Responses

The factor of slip percentage showed a consistent negative impact on attribute reliability estimates. As the slip percentage increased across each item set and sample size condition for both hierarchies, the attribute reliability estimates calculated by both the adapted and standard formulas decreased.

That is to say, the higher the slip percentage, the lower the reliability estimates regardless of the considerations of other factors. Therefore, slip percentage was a factor that influences attribute reliability estimates.

Summary

Nunnally (1972) suggested that the reliability should be at least 0.75 and preferably at least 0.85 (p. 91). If Nunnally's minimum is considered, then the 24-, 36-, and 48-item linear model produced acceptable results for all attributes in the 10% slip condition, with one exception: attribute 6 in the 24- and 36-item conditions. The reliability results were also acceptable in the 36- and 48-item conditions with 15% error, again, with the exception of attribute 6. The divergent model was more robust as it produced acceptable results for all the attributes in the 24-, 36-, and 48-item conditions with 10% slip percentage and all the attributes in the 36- and 48-item conditions with 15% slip percentage. The reliability results were even acceptable in most of 48-item conditions with 20% and 25% error.

Conclusions

Given there was no difference between the reliability estimates yielded by the adapted and the standard Cronbach's alpha formulas and conceptual richness, it is concluded that the adapted formula is a more appropriate formula for estimating the internal consistency reliability of attributes in the AHM regardless of whether the model is linear or divergent.

In using this formula, the factors of test length and occurrence of slips need to be considered. The longer the test, the higher the attribute reliability estimates. The lower the slip percentage, the higher the attribute reliability estimates. Sample size had no impact.

Limitations of the Study

The present study is limited in several aspects. First, only a simulation study was conducted to evaluate attribute-based reliability. Not including a real data study limits the practicability and persuasiveness of the present study. A real data study was not investigated because tests in practice usually are not developed according to a cognitive model of task performance. A real data study should be considered by researchers and practitioners when tests developed directly from a cognitive model are available, with the variation of test length, variability of samples, and model structure, for example.

Second, test length, domain heterogeneity, test speedness, and sample variability are factors expected to influence reliability estimates. However, sample variability was not manipulated in the current simulation study. Although sample size for simulated response data was selected as a factor in this study, the sample variability was not involved in simulation. The lack of variability among samples could be one reason to explain why reliability estimates in different sample size conditions keep invariant.

Third, the two attribute reliability estimation indices evaluated and compared in the current study are different on how the contribution of attribute

towards examinees' item-level performance is treated, which means the discrimination power of each item on each attribute is considered in the adapted index but not in the standard one. However, in the current study, discrimination power of each simulated item was fixed across all conditions, which might be the factor that evens out the possible difference on attribute reliability estimates which was expected to be resulted from the two formulas.

Another limitation of the study relates to the structure of cognitive models simulated. The two models involved in the present study were linear and divergent. The number of attributes was fixed at six for both hierarchies to make the comparison between models interpretable. However, the structure of the divergent model demonstrated a relatively simple representation of attribute organization. This limits the generalizability of the present study. More complex divergent model structures need to be considered to increase the generalizability of the findings.

Implications and Future Directions

Educational and Practical Implications

The results of the present study have three practical implications. First, attribute reliability estimates can be used to enhance score reporting by creating confidence intervals around attribute-based scores. In the AHM, the probability that an examinee possesses specific attribute combinations can be estimated in the statistical pattern recognition stage. These attribute probabilities provide examinees with specific information about their cognitive strengths and

weaknesses. Also, these attribute probabilities can be used to create score reporting profiles. By creating a confidence interval around each probability using the standard error of measurement with the attribute reliability coefficient, the precision of these point-estimate probabilities would be enhanced. In this case, the standard error of measurement should reflect weights assigned to each attribute involved in a hierarchy. Moreover, the impact on reliability estimates by adding parallel items to the test can also be evaluated using the attribute-based Spearman-Brown formula, as discussed in Chapter II and IV.

Second, the attribute-based reliability index proposed and evaluated in the present study represents a method for estimating attribute reliability in cognitive diagnostic assessment. The index considers the hierarchical relationship among attributes in a cognitive model, therefore, this procedure can be applied to not only the attribute hierarchy method but also to all other attribute-based procedures such as the rule space model (Tatsuoka, 1983), the unified model (Dibello, Stout, & Roussos, 1995), the DINA models (de la Torre & Douglas, 2004), and the NIDA models (Junker & Sijtsma, 2001) for cognitive diagnostic assessment.

Third, the outcomes on attribute reliability estimation help operationalize CDA. One implementation of CDAs is in the area of formative, classroom-based assessment. These assessments can be developed by government or test agencies at a prior time. Such assessment could be administered periodically

during the ongoing teaching and learning process. The content examined on the test should be directly linked to the curriculum because the assessment outcomes are expected to guide following teaching and learning. In addition, assessment outcomes should also have the capability of providing support on specific decisions about students' homework or teachers' instructional planning. Therefore, an assessment with acceptable reliability index will provide precision decisions and, consequently, better use of CDA in practical testing situations.

Future Research Directions

Three major issues require further investigations. First, the present study was only conducted with simulation data. The feature of simulation studies limited the practicability of the current study. How attribute reliability would perform with real cognitive models and examinees' response data is unclear and, therefore, needs to be studied. In particular, discrimination power of items measuring different attributes should vary and be more practical with a real cognitive model, and, in turn, might result in difference on the performance of the two formulas investigating attribute reliability estimates. This line of research will contribute to further understanding of attribute-based reliability in the environment of cognitive diagnostic assessments with a more practical perspective. However, such study should be conducted only with a test that is developed initially from a corresponding cognitive model of task performance and administered to public.

Second, internal consistency estimate of reliability was considered in the current study but not the consistency of classification. In cognitive diagnostic assessment, attribute reliability refers to the consistency of decision made about examinee's mastery of specific attribute. Therefore, further study is required for developing an index of classification consistency for cognitive diagnostic assessment.

Third, the number of items measuring each attribute was set as the same for all attributes in a hierarchy in the present study. Given the property of attribute dependency, attributes that are positioned at the top of a hierarchy will be assessed more reliably than attributes located at the bottom of a hierarchy. That is, attributes that are prerequisites of other attributes require fewer items to probe to achieve acceptable reliability estimates. This outcome indicates that the number of items measuring each attribute is not necessarily same for all attributes in a hierarchy. Thus, investigations focused on the relationship between diagnostic model structure and its fit to the number of items probing each attribute when estimating attribute reliability should be conducted.

References

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

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

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *Journal of the Learning Sciences, 6*, 271-315.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16(3)*, 297-334.

Cui, Y. (2007). *The hierarchy consistency index: Development and analysis*. Unpublished Doctoral Dissertation. University of Alberta: Edmonton, Alberta, Canada.

Cui, Y., Leighton, J. P., Gierl, M. J., & Hunka, S. (2006). *A person-fit statistic for the attribute hierarchy method: The hierarchy consistency index.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69,* 333-353.

Doignon, J. P., & Falmagne, J. C. (1999). *Knowledge Spaces.* NY: Springer-Verlag.

DiBello, L., Stout, W., & Roussos, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. Chipman, & R. Brennen (Eds.), *Cognitively diagnostic assessment* (pp. 361-389). Hillsdale, NJ: Earlbaum.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175-186.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*(3), 495-515.

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

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

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*(1), 3-26.

Gierl, M. J. (1997). Comparing the cognitive representations of test developers and examinees on a mathematics achievement test using Bloom's taxonomy. *Journal of Educational Research, 91*, 26-32.

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

Gierl, M. J., 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 Practices, 20*, 26-36.

Gierl, M. J., Cui, Y., & Hunka, S. (2008). Using connectionist models to evaluate examinees' response patterns on tests. *Journal of Modern Applied Statistical Methods, 7(1),* 234-245.

Gierl, M. J., Cui, Y., & Zhou, J. (2009). Reliability and Attribute-Based Scoring in Cognitive Diagnostic Assessment. *Journal of educational Measurement, 46(3),* 293-313.

Gierl, M. J., & Leighton, J. P. (2007). Linking cognitively-based models and psychometric methods. In C. R. Rao & S. Sinharay (Eds.) *Handbook of statistics: Psychometrics, Volume 26* (pp. 1103–1106). North Holland, UK: Elsevier.

Gierl, M. J., Leighton, J. P., & Hunka, S. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice, 19*, 34-44.

Gierl, M. J., Leighton, J. P., & Hunka, S. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills.

In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications.* (pp. 242-274). Cambridge, UK: Cambridge University Press.

Gierl, M. J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT*©*. *Journal of Technology, Learning, and Assessment, 6* (6). Retrieved [date] from http://www.jtla.org.

Gierl, M. J., Zheng, Y., & Cui, Y. (2008). Using the attribute hierarchy method to identify and interpret the cognitive skills that produce group differences. *Journal of Educational Measurement, 45,* 65-89.

Gierl, M. J., & Zhou, J. (2008). Computer adaptive-attribute testing: A new approach to cognitive diagnostic assessment. *Zeitschift für Psychologie—Journal of Psychology, 216,* 29-39.

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

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*, 181-200.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measure, 26*, 333-352.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Washington, DC: American Council on Education.

Hartz, S. M. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: blending theory with practicality.* Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign, Department of Statistics.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258-272.

Kuhn, D. (2001). Why development does (and does not occur) occur: Evidence from the domain of inductive reasoning. In J. L. McClelland & R. Siegler (Eds.), *Mechanisms of Cognitive Development: Behavioral and Neural Perspectives* (pp. 221-249). Hilldale, NJ: Erlbaum.

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

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

Leighton, J. P., & Gierl, M. J. (2007b). Verbal reports as data for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. (pp. 146–172). Cambridge, UK: Cambridge University Press.

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

Lohman, D. F. (2000). Complex information processing and intelligence. In R.J. Sternberg (Ed.), *Handbook of intelligence* (pp. 285–340). NY: Cambridge University Press.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 33*, 279-416.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. Washington, DC: American Council on Education.

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

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: *Item analysis and test scoring with binary logistic test models [Computer Program]*. Moorseville, IN: Scientific Software.

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

Mislevy, R. J., Steinberg, L. S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-62.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, and R. Glaser (Eds.). Board on Testing and Assessment, Center for Education. Washington, DC: National Academy Press.

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

Nichols, P. D., & Smith, P. L. (1998). Contextualizing the interpretation of reliability data. *Educational Measurement: Issues and Practice, 17,* 24-36.

Nunally, J. C. (1972). Educational Measurement and Evaluation, McGraw-Hill, New York.

Pellegrino, J. W. (2002). *Understanding how students learn and inferring what they know: Implications for the design of curriculum, instruction, and assessment*. In M. J. Smith (Ed.), NSF K-12 Mathematics and Science Curriculum and Implementation Centers Conference Proceedings (pp. 76-92). Washington, DC: National Science Foundation and American Geological Institute.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practices. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* (pp. 307-353). Washington, DC: American Educational Research Association.

Poggio, A., Clayton, D. B., Glasnapp, D., Poggio, J., Haack, P., & Thomas, J. (2005). *Revisiting the item format question: Can the multiple choice format meet the demand for monitoring higher-order skills?* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Roussos, L., Dibello, L., Stout, W., Hartz, S., Henson, R., & Templin, J. (2007). The fusion model skills diagnostic system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment in education: Theory and applications* (pp. 275-318). Cambridge, UK: Cambridge University Press.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*(4), 333-352.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: American Council on Education, Macmillian.

Standards for Educational and Psychological Testing. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*(4), 485-518.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification methods. *Journal of the Royal Statistical Society Series C (Applied Statistics), 51*, 337-350.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345-354.

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

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

Taylor, K. L., & Dionne, J-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92*, 413-425.

Wang, C., & Gierl, M. J. (2007, April). *Investigating the cognitive processes underlying student performance on the SAT Critical Reading subtest.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Webb, N. L. (2006). Identifying content for student achievement tests. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of test development* (pp. 155-180). Mahwah, NJ: Erlbaum.

Zucker, S., Sassman, C., & Case, B. J. (2004). Cognitive labs. San Antonio, TX: Harcourt Assessment. Retrieved 10/11/2006 from https://harcourtassessment.com/hai/Images/resource/library/pdf/CognitiveLabs_Final.pdf

Table 1

Blueprint of 2007 Alberta Grade 3 Mathematics Achievement Test

General Outcomes	Reporting Category		Number (Percentage) of Questions
	Knowledge	Skills	
Number • Develop a number sense for whole numbers 0 to 1 000, and explore fractions (fifths and tenths) • Apply an arithmetic operation (addition, subtraction, multiplication, or division) on whole numbers, and illustrate its use in creating and solving problems • Use and justify an appropriate calculation strategy or technology to solve problems	8	9	17 (40%)
Patterns and Relations • Investigate, establish, and communicate rules for numerical and non-numerical patterns, including those found in the home, and use these rules to make predictions	2	4	6 (14%)
Shape and Space • Estimate, measure, and compare by using whole numbers and primarily standard units of measure • Describe, classify, construct, and relate 3-D objects and 2-D shapes • Use numbers and direction words to describe the relative positions of objects in one dimension using everyday contexts	4	8	12 (28%)
Statistics and Probability • Collect first- and second-hand data, display the results in more than one way, and interpret the data to make predictions • Use simple probability experiments designed by others in order to explain outcomes	3	5	8 (18%)
Number (Percentage) of Questions	17 (39%)	26 (61%)	43 (100%)

Table 2

An Illustration of Classification Method A: Classifying Observed Response Pattern (101000)

Examinee Attribute	Expected Response Pattern	Ability Level	No. of Slips	Likelihood
0000	000000	-2.357	2	0.0022
1000	100000	-0.792	1	0.1227
1100	110000	-0.156	2	0.0019
1110	**111000**	**0.428**	**1**	**0.2537**
1001	100100	-0.161	2	0.0018
1101	110110	0.637	4	0.0001
1111	111111	1.853	4	0.0001

Table 3

An Illustration of Classification Method B: Classifying Observed Response Pattern (101000)

Examinee Attribute	Expected Response Pattern	Ability Level	No. of Slips	Likelihood
0000	000000	-2.357	0	*
1000	100000	-0.792	0	*
1100	110000	-0.156	1	0.0219
1110	**111000**	**0.428**	**1**	**0.2483**
1001	100100	-0.161	1	0.0518
1101	110110	0.637	3	0.0001
1111	111111	1.853	4	0.0001

Table 4

An Illustration of Neural Network Classification: Classifying Observed Response Pattern (101000)

Observed Response Pattern	Attribute Probability			
	1	2	3	4
101000	0.99	0.95	0.96	0.06

Table 5

Item Parameters Estimated from the Expected Response Matrix for the Linear and Divergent Models

	Linear		*Divergent*	
Attribute	a	b	a	b
1	0.50	-2.50	1.00	-2.50
2	1.00	-1.50	2.00	0.00
3	1.50	-0.50	3.00	2.50
4	2.00	0.50	2.00	0.00
5	2.50	1.50	3.00	2.50
6	3.00	2.50	3.00	2.50

Table 6a

Attribute Reliability Estimated by Adapted Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Linear Cognitive Model as a Function of Test Length (Sample Size=250)

Item Set	Attribute	Slip Percentages			
		10%	15%	20%	25%
2	1	0.83	0.81	0.78	0.74
	2	0.81	0.79	0.74	0.68
	3	0.77	0.74	0.68	0.60
	4	0.71	0.67	0.61	0.51
	5	0.60	0.45	0.34	0.25
	6	0.37	0.34	0.27	0.28
4	1	0.91	0.90	0.88	0.86
	2	0.90	0.89	0.86	0.83
	3	0.88	0.86	0.81	0.77
	4	0.84	0.80	0.75	0.67
	5	0.77	0.69	0.62	0.52
	6	0.58	0.43	0.37	0.26
6	1	0.94	0.93	0.92	0.90
	2	0.93	0.92	0.90	0.88
	3	0.92	0.90	0.87	0.83
	4	0.89	0.86	0.82	0.75
	5	0.84	0.78	0.71	0.64
	6	0.70	0.58	0.53	0.52
8	1	0.96	0.95	0.94	0.93
	2	0.95	0.94	0.93	0.91
	3	0.94	0.92	0.90	0.87
	4	0.92	0.89	0.86	0.81
	5	0.87	0.82	0.76	0.68
	6	0.76	0.64	0.55	0.55

Table 6b

Attribute Reliability Estimated by Standard Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Linear Cognitive Model as a Function of Test Length (Sample Size=250)

		Slip Percentages			
Item Set	Attribute	10%	15%	20%	25%
2	1	0.85	0.83	0.80	0.75
	2	0.84	0.82	0.77	0.71
	3	0.82	0.79	0.73	0.64
	4	0.77	0.72	0.64	0.53
	5	0.66	0.51	0.40	0.29
	6	0.37	0.34	0.27	0.28
4	1	0.92	0.91	0.89	0.86
	2	0.92	0.90	0.87	0.84
	3	0.90	0.88	0.84	0.78
	4	0.87	0.83	0.78	0.69
	5	0.80	0.72	0.64	0.52
	6	0.58	0.43	0.37	0.26
6	1	0.95	0.94	0.92	0.90
	2	0.94	0.93	0.91	0.88
	3	0.94	0.92	0.89	0.85
	4	0.91	0.88	0.84	0.77
	5	0.86	0.80	0.73	0.65
	6	0.70	0.58	0.53	0.52
8	1	0.96	0.95	0.94	0.93
	2	0.96	0.95	0.93	0.91
	3	0.95	0.94	0.91	0.88
	4	0.93	0.91	0.87	0.82
	5	0.89	0.84	0.78	0.69
	6	0.76	0.64	0.55	0.55

Table 7a

Attribute Reliability Estimated by Adapted Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Linear Cognitive Model as a Function of Test Length (Sample Size=500)

Item Set	Attribute	Slip Percentages			
		10%	15%	20%	25%
2	1	0.84	0.81	0.79	0.75
	2	0.82	0.78	0.74	0.69
	3	0.78	0.73	0.68	0.61
	4	0.72	0.65	0.60	0.52
	5	0.61	0.49	0.45	0.37
	6	0.47	0.33	0.30	0.26
4	1	0.92	0.90	0.88	0.86
	2	0.91	0.89	0.86	0.83
	3	0.88	0.85	0.82	0.77
	4	0.85	0.79	0.74	0.67
	5	0.77	0.69	0.60	0.51
	6	0.57	0.40	0.36	0.34
6	1	0.94	0.93	0.92	0.90
	2	0.94	0.92	0.90	0.88
	3	0.92	0.90	0.87	0.84
	4	0.90	0.86	0.81	0.75
	5	0.85	0.78	0.70	0.63
	6	0.71	0.56	0.51	0.46
8	1	0.96	0.95	0.94	0.93
	2	0.95	0.94	0.93	0.91
	3	0.94	0.92	0.90	0.87
	4	0.92	0.89	0.86	0.81
	5	0.88	0.83	0.77	0.70
	6	0.75	0.64	0.58	0.54

Table 7b

Attribute Reliability Estimated by Standard Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Linear Cognitive Model as a Function of Test Length (Sample Size=500)

		Slip Percentages			
Item Set	Attribute	10%	15%	20%	25%
2	1	0.86	0.83	0.80	0.76
	2	0.85	0.81	0.78	0.73
	3	0.83	0.78	0.74	0.66
	4	0.78	0.72	0.66	0.56
	5	0.68	0.57	0.51	0.42
	6	0.47	0.33	0.30	0.26
4	1	0.93	0.91	0.89	0.86
	2	0.92	0.90	0.87	0.84
	3	0.91	0.88	0.84	0.79
	4	0.88	0.83	0.77	0.70
	5	0.81	0.72	0.63	0.52
	6	0.57	0.40	0.36	0.34
6	1	0.95	0.94	0.92	0.90
	2	0.95	0.93	0.91	0.89
	3	0.94	0.91	0.89	0.85
	4	0.92	0.88	0.84	0.77
	5	0.87	0.80	0.73	0.63
	6	0.71	0.56	0.51	0.46
8	1	0.96	0.95	0.94	0.93
	2	0.96	0.95	0.93	0.91
	3	0.95	0.93	0.91	0.88
	4	0.94	0.91	0.88	0.83
	5	0.90	0.85	0.79	0.71
	6	0.75	0.64	0.58	0.54

Table 8a

Attribute Reliability Estimated by Adapted Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Linear Cognitive Model as a Function of Test Length (Sample Size=750)

Item Set	Attribute	Slip Percentages			
		10%	15%	20%	25%
2	1	0.84	0.81	0.78	0.75
	2	0.82	0.79	0.74	0.70
	3	0.77	0.73	0.67	0.62
	4	0.71	0.65	0.58	0.50
	5	0.58	0.48	0.40	0.29
	6	0.40	0.30	0.18	0.20
4	1	0.92	0.90	0.88	0.86
	2	0.91	0.88	0.86	0.83
	3	0.88	0.85	0.82	0.77
	4	0.84	0.79	0.74	0.68
	5	0.77	0.68	0.59	0.51
	6	0.59	0.43	0.35	0.36
6	1	0.94	0.93	0.92	0.91
	2	0.94	0.92	0.90	0.88
	3	0.92	0.90	0.87	0.84
	4	0.89	0.86	0.81	0.76
	5	0.84	0.77	0.70	0.63
	6	0.68	0.55	0.48	0.43
8	1	0.96	0.95	0.94	0.93
	2	0.95	0.94	0.93	0.91
	3	0.94	0.92	0.90	0.87
	4	0.92	0.89	0.86	0.81
	5	0.88	0.82	0.76	0.69
	6	0.76	0.65	0.57	0.57

Table 8b

Attribute Reliability Estimated by Standard Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Linear Cognitive Model as a Function of Test Length (Sample Size=750)

		Slip Percentages			
Item Set	Attribute	10%	15%	20%	25%
2	1	0.85	0.83	0.80	0.76
	2	0.85	0.81	0.77	0.73
	3	0.82	0.77	0.72	0.65
	4	0.78	0.71	0.63	0.53
	5	0.66	0.55	0.44	0.31
	6	0.40	0.30	0.18	0.20
4	1	0.92	0.91	0.89	0.86
	2	0.92	0.90	0.87	0.84
	3	0.91	0.88	0.84	0.79
	4	0.87	0.83	0.78	0.70
	5	0.80	0.71	0.62	0.53
	6	0.59	0.43	0.35	0.36
6	1	0.95	0.94	0.92	0.91
	2	0.95	0.93	0.91	0.89
	3	0.94	0.91	0.89	0.85
	4	0.91	0.88	0.84	0.78
	5	0.86	0.79	0.72	0.63
	6	0.68	0.55	0.48	0.43
8	1	0.96	0.95	0.94	0.93
	2	0.96	0.95	0.93	0.91
	3	0.95	0.93	0.91	0.89
	4	0.94	0.91	0.88	0.83
	5	0.89	0.84	0.78	0.71
	6	0.76	0.65	0.57	0.57

Table 9a

Attribute Reliability Estimated by Adapted Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Linear Cognitive Model as a Function of Test Length (Sample Size=1000)

Item Set	Attribute	Slip Percentages			
		10%	15%	20%	25%
2	1	0.84	0.81	0.78	0.75
	2	0.82	0.78	0.74	0.69
	3	0.78	0.73	0.68	0.61
	4	0.71	0.64	0.57	0.48
	5	0.59	0.49	0.39	0.28
	6	0.49	0.25	0.23	0.18
4	1	0.92	0.90	0.88	0.86
	2	0.91	0.88	0.86	0.83
	3	0.88	0.85	0.82	0.77
	4	0.84	0.80	0.75	0.68
	5	0.77	0.69	0.60	0.52
	6	0.59	0.45	0.39	0.39
6	1	0.94	0.93	0.92	0.90
	2	0.94	0.92	0.90	0.88
	3	0.92	0.90	0.87	0.84
	4	0.89	0.86	0.82	0.76
	5	0.84	0.78	0.70	0.62
	6	0.71	0.57	0.48	0.45
8	1	0.96	0.95	0.94	0.93
	2	0.95	0.94	0.93	0.91
	3	0.94	0.92	0.90	0.87
	4	0.92	0.89	0.86	0.81
	5	0.88	0.82	0.76	0.70
	6	0.77	0.64	0.57	0.56

Table 9b

Attribute Reliability Estimated by Standard Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Linear Cognitive Model as a Function of Test Length (Sample Size=1000)

		Slip Percentages			
Item Set	Attribute	10%	15%	20%	25%
2	1	0.86	0.83	0.80	0.75
	2	0.85	0.81	0.77	0.71
	3	0.83	0.78	0.72	0.64
	4	0.78	0.70	0.62	0.51
	5	0.68	0.55	0.43	0.30
	6	0.49	0.25	0.23	0.18
4	1	0.93	0.91	0.89	0.86
	2	0.92	0.90	0.87	0.84
	3	0.91	0.88	0.84	0.79
	4	0.88	0.83	0.78	0.71
	5	0.81	0.72	0.64	0.54
	6	0.59	0.45	0.39	0.39
6	1	0.95	0.94	0.92	0.90
	2	0.95	0.93	0.91	0.89
	3	0.94	0.92	0.89	0.85
	4	0.92	0.88	0.84	0.78
	5	0.87	0.80	0.72	0.63
	6	0.71	0.57	0.48	0.45
8	1	0.96	0.95	0.94	0.93
	2	0.96	0.95	0.93	0.92
	3	0.95	0.94	0.91	0.89
	4	0.94	0.91	0.87	0.84
	5	0.90	0.84	0.78	0.72
	6	0.77	0.64	0.57	0.52

Table 10a

Attribute Reliability Estimated by Adapted Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Divergent Cognitive Model as a Function of Test Length (Sample Size=250)

Item Set	Attribute	Slip Percentages			
		10%	15%	20%	25%
2	1	0.81	0.82	0.80	0.80
	2	0.71	0.67	0.59	0.61
	3	0.69	0.56	0.45	0.43
	4	0.76	0.74	0.68	0.67
	5	0.71	0.57	0.52	0.45
	6	0.63	0.53	0.47	0.42
4	1	0.90	0.91	0.89	0.89
	2	0.84	0.82	0.77	0.79
	3	0.81	0.72	0.61	0.56
	4	0.87	0.86	0.81	0.80
	5	0.82	0.74	0.61	0.55
	6	0.81	0.73	0.65	0.55
6	1	0.93	0.94	0.92	0.93
	2	0.89	0.87	0.83	0.88
	3	0.86	0.78	0.68	0.65
	4	0.91	0.90	0.87	0.86
	5	0.86	0.79	0.70	0.66
	6	0.86	0.79	0.69	0.65
8	1	0.95	0.95	0.94	0.94
	2	0.92	0.90	0.87	0.88
	3	0.89	0.84	0.75	0.71
	4	0.94	0.93	0.90	0.90
	5	0.89	0.83	0.75	0.71
	6	0.90	0.84	0.76	0.73

Table 10b

Attribute Reliability Estimated by Standard Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Divergent Cognitive Model as a Function of Test Length (Sample Size=250)

		Slip Percentages			
Item Set	Attribute	10%	15%	20%	25%
2	1	0.81	0.82	0.79	0.79
	2	0.75	0.70	0.61	0.61
	3	0.69	0.56	0.45	0.43
	4	0.77	0.75	0.69	0.68
	5	0.71	0.57	0.52	0.45
	6	0.63	0.53	0.47	0.42
4	1	0.90	0.90	0.88	0.88
	2	0.86	0.83	0.77	0.77
	3	0.81	0.72	0.61	0.56
	4	0.88	0.86	0.82	0.80
	5	0.82	0.74	0.61	0.55
	6	0.81	0.73	0.65	0.55
6	1	0.93	0.93	0.92	0.92
	2	0.90	0.88	0.83	0.83
	3	0.86	0.78	0.68	0.65
	4	0.92	0.90	0.87	0.86
	5	0.86	0.79	0.70	0.66
	6	0.86	0.79	0.69	0.65
8	1	0.95	0.95	0.94	0.94
	2	0.93	0.90	0.87	0.87
	3	0.89	0.84	0.75	0.71
	4	0.94	0.93	0.90	0.89
	5	0.89	0.83	0.75	0.71
	6	0.90	0.84	0.76	0.73

Table 11a

Attribute Reliability Estimated by Adapted Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Divergent Cognitive Model as a Function of Test Length (Sample Size=500)

Item Set	Attribute	Slip Percentages			
		10%	15%	20%	25%
2	1	0.80	0.81	0.80	0.80
	2	0.70	0.65	0.58	0.58
	3	0.71	0.61	0.52	0.48
	4	0.76	0.73	0.69	0.66
	5	0.73	0.54	0.43	0.36
	6	0.71	0.58	0.42	0.37
4	1	0.90	0.90	0.89	0.89
	2	0.84	0.80	0.77	0.78
	3	0.82	0.72	0.63	0.57
	4	0.87	0.85	0.82	0.81
	5	0.82	0.70	0.60	0.59
	6	0.83	0.70	0.60	0.57
6	1	0.93	0.93	0.93	0.93
	2	0.89	0.86	0.83	0.84
	3	0.87	0.78	0.70	0.68
	4	0.91	0.90	0.87	0.86
	5	0.87	0.78	0.70	0.69
	6	0.87	0.79	0.70	0.70
8	1	0.95	0.95	0.94	0.94
	2	0.92	0.89	0.87	0.88
	3	0.90	0.83	0.76	0.75
	4	0.94	0.92	0.90	0.90
	5	0.90	0.83	0.77	0.75
	6	0.90	0.83	0.77	0.75

Table 11b

Attribute Reliability Estimated by Standard Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Divergent Cognitive Model as a Function of Test Length (Sample Size=500)

		Slip Percentages			
Item Set	Attribute	10%	15%	20%	25%
2	1	0.80	0.81	0.80	0.79
	2	0.74	0.69	0.61	0.59
	3	0.71	0.61	0.52	0.48
	4	0.77	0.74	0.70	0.68
	5	0.73	0.54	0.43	0.36
	6	0.71	0.58	0.42	0.37
4	1	0.90	0.89	0.89	0.88
	2	0.86	0.81	0.77	0.76
	3	0.82	0.72	0.63	0.57
	4	0.88	0.85	0.82	0.81
	5	0.82	0.70	0.60	0.59
	6	0.83	0.70	0.60	0.57
6	1	0.93	0.93	0.92	0.92
	2	0.90	0.87	0.83	0.83
	3	0.87	0.78	0.70	0.68
	4	0.92	0.90	0.88	0.87
	5	0.87	0.78	0.70	0.69
	6	0.87	0.79	0.70	0.70
8	1	0.95	0.95	0.94	0.94
	2	0.93	0.90	0.87	0.87
	3	0.90	0.83	0.76	0.75
	4	0.94	0.92	0.90	0.90
	5	0.90	0.83	0.77	0.75
	6	0.90	0.83	0.77	0.75

Table 12a

Attribute Reliability Estimated by Adapted Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Divergent Cognitive Model as a Function of Test Length (Sample Size=750)

Item Set	Attribute	Slip Percentages			
		10%	15%	20%	25%
2	1	0.80	0.80	0.80	0.79
	2	0.71	0.63	0.61	0.61
	3	0.69	0.57	0.46	0.49
	4	0.76	0.71	0.67	0.65
	5	0.68	0.48	0.39	0.34
	6	0.71	0.56	0.47	0.49
4	1	0.90	0.89	0.89	0.89
	2	0.84	0.79	0.77	0.78
	3	0.81	0.69	0.60	0.56
	4	0.87	0.84	0.81	0.80
	5	0.81	0.71	0.60	0.59
	6	0.81	0.69	0.60	0.58
6	1	0.93	0.93	0.93	0.92
	2	0.89	0.86	0.84	0.84
	3	0.86	0.78	0.71	0.70
	4	0.92	0.89	0.87	0.86
	5	0.87	0.78	0.70	0.69
	6	0.87	0.79	0.70	0.68
8	1	0.95	0.95	0.94	0.94
	2	0.92	0.89	0.87	0.88
	3	0.89	0.82	0.75	0.74
	4	0.94	0.92	0.90	0.89
	5	0.90	0.83	0.76	0.75
	6	0.90	0.83	0.77	0.75

Table 12b

Attribute Reliability Estimated by Standard Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Divergent Cognitive Model as a Function of Test Length (Sample Size=750)

		Slip Percentages			
Item Set	Attribute	10%	15%	20%	25%
2	1	0.81	0.80	0.79	0.79
	2	0.75	0.67	0.62	0.61
	3	0.69	0.57	0.46	0.49
	4	0.77	0.73	0.69	0.67
	5	0.68	0.48	0.39	0.34
	6	0.71	0.56	0.47	0.49
4	1	0.90	0.89	0.89	0.88
	2	0.86	0.81	0.77	0.76
	3	0.81	0.69	0.60	0.56
	4	0.88	0.85	0.82	0.81
	5	0.81	0.71	0.60	0.59
	6	0.81	0.69	0.60	0.58
6	1	0.93	0.93	0.92	0.92
	2	0.90	0.87	0.84	0.83
	3	0.86	0.78	0.71	0.70
	4	0.92	0.90	0.88	0.86
	5	0.87	0.78	0.70	0.69
	6	0.87	0.79	0.70	0.68
8	1	0.95	0.95	0.94	0.94
	2	0.93	0.90	0.87	0.87
	3	0.89	0.82	0.75	0.74
	4	0.94	0.92	0.90	0.90
	5	0.90	0.83	0.76	0.75
	6	0.90	0.83	0.77	0.75

Table 13a

Attribute Reliability Estimated by Adapted Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Divergent Cognitive Model as a Function of Test Length (Sample Size=1000)

Item Set	Attribute	Slip Percentages			
		10%	15%	20%	25%
2	1	0.80	0.80	0.80	0.79
	2	0.70	0.64	0.59	0.60
	3	0.67	0.51	0.44	0.42
	4	0.76	0.71	0.68	0.64
	5	0.67	0.52	0.44	0.40
	6	0.71	0.58	0.47	0.42
4	1	0.90	0.89	0.89	0.89
	2	0.84	0.79	0.77	0.76
	3	0.80	0.69	0.60	0.56
	4	0.87	0.84	0.82	0.80
	5	0.82	0.71	0.63	0.58
	6	0.81	0.70	0.61	0.58
6	1	0.93	0.93	0.93	0.92
	2	0.89	0.86	0.84	0.84
	3	0.86	0.78	0.71	0.68
	4	0.91	0.89	0.88	0.86
	5	0.87	0.78	0.71	0.69
	6	0.87	0.79	0.71	0.68
8	1	0.95	0.95	0.94	0.94
	2	0.92	0.89	0.87	0.88
	3	0.90	0.82	0.76	0.74
	4	0.94	0.92	0.90	0.89
	5	0.90	0.83	0.77	0.74
	6	0.90	0.82	0.76	0.74

Table 13b

Attribute Reliability Estimated by Standard Cronbach's Alpha Coefficient Using Different Percentages of Slips in the Observed Response Patterns for a Divergent Cognitive Model as a Function of Test Length (Sample Size=1000)

		Slip Percentages			
Item Set	Attribute	10%	15%	20%	25%
2	1	0.80	0.80	0.80	0.78
	2	0.74	0.66	0.61	0.60
	3	0.67	0.51	0.44	0.42
	4	0.77	0.73	0.70	0.66
	5	0.67	0.52	0.44	0.40
	6	0.71	0.58	0.47	0.42
4	1	0.90	0.89	0.89	0.88
	2	0.86	0.80	0.77	0.75
	3	0.80	0.69	0.60	0.56
	4	0.88	0.85	0.83	0.80
	5	0.82	0.71	0.63	0.58
	6	0.81	0.70	0.61	0.58
6	1	0.93	0.93	0.92	0.92
	2	0.90	0.87	0.84	0.83
	3	0.86	0.78	0.71	0.68
	4	0.92	0.90	0.88	0.86
	5	0.87	0.78	0.71	0.69
	6	0.87	0.79	0.71	0.68
8	1	0.95	0.94	0.94	0.94
	2	0.93	0.90	0.87	0.86
	3	0.90	0.82	0.76	0.74
	4	0.94	0.92	0.91	0.89
	5	0.90	0.83	0.77	0.74
	6	0.90	0.82	0.76	0.74

Table 14

Correlation of Reliability Estimates on Linear Model between Adapted and Standard Formula

	Sample Size			
Slip Percentage	250	500	750	1000
10%	0.99	0.99	0.99	0.99
15%	1.00	0.99	0.99	1.00
20%	1.00	0.99	1.00	1.00
25%	1.00	1.00	1.00	1.00

Table 15

Root Mean Square Deviation of Reliability Estimates on the Linear Model between Adapted and Standard Formula

		Slip Percentage			
	Item set	10%	15%	20%	25%
250	2	0.05	0.04	0.04	0.03
	4	0.02	0.02	0.02	0.01
	6	0.02	0.02	0.02	0.02
	8	0.01	0.01	0.01	0.01
500	2	0.05	0.05	0.05	0.04
	4	0.03	0.02	0.02	0.02
	6	0.02	0.02	0.02	0.01
	8	0.01	0.01	0.01	0.01
750	2	0.05	0.05	0.04	0.02
	4	0.03	0.02	0.02	0.02
	6	0.02	0.02	0.02	0.01
	8	0.01	0.01	0.01	0.01
1000	2	0.06	0.04	0.04	0.02
	4	0.03	0.03	0.03	0.02
	6	0.02	0.02	0.02	0.01
	8	0.01	0.01	0.01	0.03

Table 16

Correlation of Reliability Estimates on Divergent Model between Adapted and Standard Formula

	Sample Size			
Slip Percentage	250	500	750	1000
10%	0.99	0.99	0.99	0.99
15%	1.00	1.00	1.00	1.00
20%	1.00	1.00	1.00	1.00
25%	1.00	1.00	1.00	1.00

Table 17

Root Mean Square Deviation of Reliability Estimates on the Divergent Model between Adapted and Standard Formula

		Slip Percentage			
	Item set	10%	15%	20%	25%
250	2	0.02	0.01	0.01	0.01
	4	0.01	0.01	0.00	0.01
	6	0.01	0.00	0.00	0.02
	8	0.00	0.00	0.00	0.01
500	2	0.02	0.02	0.02	0.01
	4	0.01	0.01	0.00	0.01
	6	0.01	0.00	0.00	0.01
	8	0.00	0.00	0.00	0.01
750	2	0.02	0.02	0.01	0.01
	4	0.01	0.01	0.00	0.01
	6	0.01	0.00	0.00	0.01
	8	0.00	0.00	0.00	0.00
1000	2	0.02	0.02	0.01	0.01
	4	0.01	0.01	0.00	0.01
	6	0.01	0.00	0.00	0.01
	8	0.00	0.00	0.00	0.01

Figure 1a. A linear cognitive model containing three attributes.

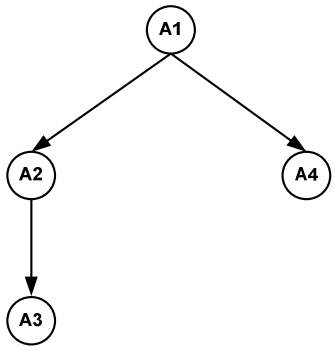
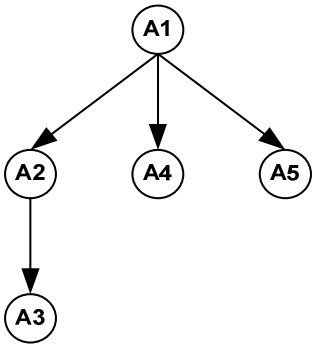
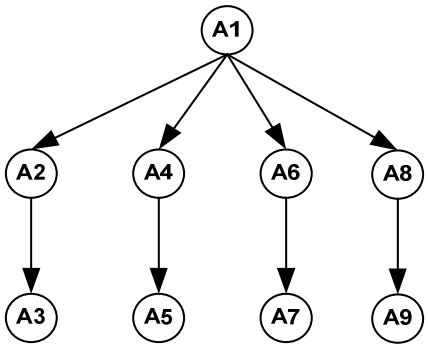


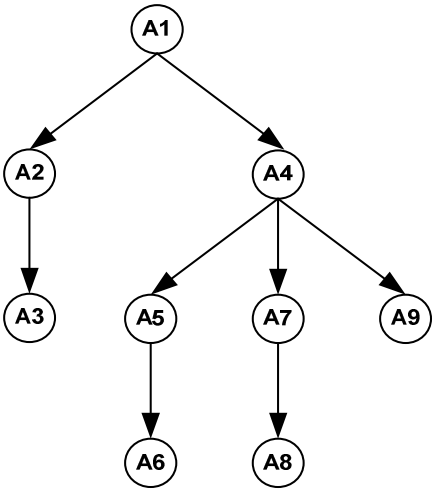
Figure 1b. A divergent cognitive model containing four attributes.



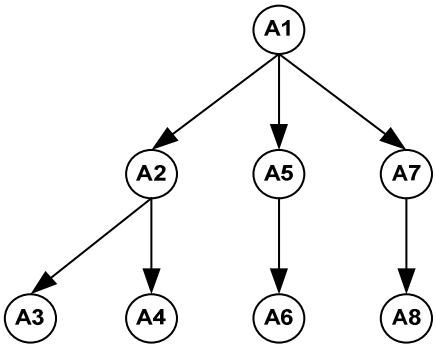
Hierarchy 1: Exponents and Algebra



Hierarchy 2: Basic Algebra



Hierarchy 3: Ratios and Algebra



Hierarchy 4: Equation and Inequality Solution, Algebraic Operation, Algebraic Substitution, and Exponents

Figure 2. Four cognitive hierarchies used to describe examinee performance on the SAT Algebra Subtest.

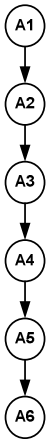


Figure 3. Linear Hierarchy

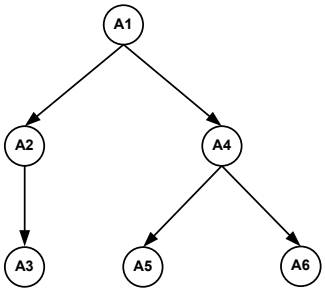


Figure 4. Divergent Hierarchy

Appendix 1

Part 1. Summary of the Attributes Required to Solve the Items in Hierarchy 1, Basic Algebra I

Attribute A1 includes the basic mathematical knowledge and skills required for setting up a single ratio by comparing two quantities.

Attribute A2 requires the mastery of the skills to order a geometric series. This attribute involves the knowledge about geometric series (e.g., the nature of the between-term ratio) and/or the consecutive numerical computation (e.g., multiplication and division).

Attribute A3 considers the skills for solving geometric series in an abstract pattern.

Attribute A4 includes the skills required for representing and executing multiple basic algebraic skills.

Attribute A5, termed fraction transformation, is also an attribute with multiple skills. This attribute requires a host of specific skills including representing and executing multiple advanced algebraic skills such as setting up a single ratio, skills for transforming fraction, and insights, such as when, where, and/or how to do the transformation.

Part 2. Summary of the Attributes Required to Solve the Items in Hierarchy 2, Basic Algebra II

Attribute A1, which includes the basic language knowledge enabling the student to understand the test item, and the most basic mathematical knowledge and skills, such as the property of absolute value and arithmetic operations.

Attribute A2 includes the basic knowledge of exponential and power addition operation.

Attribute A3 involves the knowledge of power multiplication and flexible application of multiple rules in exponential operation.

Attribute A4 requires the mastery of the skills to order a geometric series. This attribute involves the knowledge about geometric series (e.g., the nature of the between-term ratio) and/or the consecutive numerical computation (e.g., multiplication and division)—see also Hierarchy 1, Attribute A2.

Attribute A5 considers the skills for solving geometric series in an abstract pattern—see also Hierarchy 1, Attribute A3.

Attribute A6 requires the basic mathematical skills in solving for a linear equation (e.g., subtraction or division on both sides). This attribute also requires the management of the basic mathematical skills (i.e., Attribute A1) on both sides of a linear equation.

Attribute A7 requires the skills of setting up and solving for a quadratic equation, which generally involves the skills in solving a linear equation and additional skills (e.g., factoring).

Attribute A8 represents the skills of mapping a graph of a familiar function (e.g., a parabola) with its corresponding function. This attribute involves the knowledge about the graph of a familiar function and/or substituting points in the graph.

Attribute A9 deals with the abstract properties of functions, such as recognizing the graphical representation of the relationship between independent and dependent variables.

Part 3. Summary of the Attributes Required to Solve the Items in Hierarchy 3, Ratios and Algebra

Attribute A1 represents the most basic arithmetic operation skills (e.g., addition, subtraction, multiplication, and division of numbers).

Attribute A2 includes the knowledge about the properties of factors.

Attribute A3 involves the skills of applying the rules of factoring.

Attribute A4 includes the skills required for substituting values into algebraic expressions.

Attribute A5 represents the skills of mapping a graph of a familiar function (e.g., a parabola) with its corresponding function—see also Hierarchy 2, Attribute 8.

Attribute A6 deals with the abstract properties of functions, such as recognizing the graphical representation of the relationship between independent and dependent variables—see also Hierarchy 2, Attribute 9.

Attribute A7 requires the skills to substitute numbers into algebraic expressions.

Attribute A8 represents the skills of advanced substitution. Algebraic expressions, rather than numbers, need to be substituted into another algebraic expression.

Attribute A9 related to skills associated with rule understanding and application.

Part 4. Summary of the Attributes Required to Solve the Items in Hierarchy 4, Equation and Inequality Solutions, Algebraic Operations, Algebraic Substitution, and Exponents

Attribute A1, which includes the basic language knowledge enabling the student to understand the test item, and the most basic mathematical knowledge and skills, such as the property of absolute value and arithmetic operations—see also Hierarchy 2, Attribute 1.

Attribute A2 represents the most basic arithmetic operation skills (e.g., addition, subtraction, multiplication, and division of numbers)—see also Hierarchy 3, Attribute 1.

Attribute A3 involves the skills of solving quadratic inequality with two variables.

Attribute A4 represents the skills of solving multiple linear equations.

Attribute A5 considers the skills of substituting values into algebraic expressions—see also Hierarchy 3, Attribute A7.

Attribute A6 involves the skills of rule understanding and substitution—see also Hierarchy 3, Attribute A9.

Attribute A7 requires the basic knowledge of exponential and power addition operation—see also Hierarchy 2, Attribute A2.

Attribute A8 represents the knowledge of power multiplication and flexible application of multiple rules in exponential operation—see also Hierarchy 2, Attribute A3.