University of Alberta

The Hierarchy Consistency Index: A Person-fit Statistic for the Attribute Hierarchy Method

by

Ying Cui

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

In

Measurement, Evaluation, and Cognition

Department of Educational Psychology

Edmonton, Alberta

Fall 2007

Your file Votre référence
ISBN: 978-0-494-32946-7
Our file Notre référence
ISBN: 978-0-494-32946-7

Abstract

The attribute hierarchy method (AHM) (Leighton, Gierl, & Hunka, 2004), which is based on the assumption that test items can be described by a set of hierarchically-ordered attributes, is designed explicitly to integrate cognitive models with a psychometric technique to model students' test performances and estimate their mastery of domain knowledge and cognitive skills. The AHM, by incorporating the assumption of attribute dependency, brings an important cognitive property into cognitive diagnostic models. However, the validity of this method depends critically on the accuracy and adequacy of the attribute hierarchy in representing students' response processes. This study introduces a person-fit statistic, called the hierarchy consistency index (HCI_i), to help assess the degree to which an observed student response vector is consistent with the attribute hierarchy, ultimately enhancing the validity of diagnostic feedback produced with the AHM. In order to statistically test the significance of an observed HCI_i, a simulation approach was used for setting critical values to determine whether the HCI_i value is sufficiently high to show a statistical fit of the observed response vector to the attribute hierarchy.

Simulation studies were conducted for two purposes. The first purpose was to evaluate the effectiveness of the HCI_i in assessing the misfit of a student response vector to the attribute hierarchy. The second purpose was to use statistical approaches to identify the critical values for testing both person fit and overall model fit using the HCI_i. Simulation results revealed that the HCI_i performed well in determining the degree to which an observed response vector was consistent with the attribute hierarchy across different simulation conditions. Results also indicated that critical values identified for

examining person fit were overly liberal, suggesting that the use of this statistical approach was not practically feasible. In addition, critical values for examining overall model fit varied across different hierarchies. Based on the simulation results and the author's practical experience with the HCI_i, criteria for interpreting the HCI_i were recommended.

Acknowledgements

I am very grateful to many people who were supportive and helpful throughout my doctoral work. My first thanks must go to my supervisor, Dr. Jacqueline Leighton. Her intelligence, kindness, enthusiasm, and support make all the difference in my academic career. I consider myself very lucky to have her as my mentor who has guided me to make wise decisions in my study, research, and career. I wish to extend my gratitude to Dr. Mark Gierl, Dr. Steve Hunka, and Dr. Todd Rogers for their generous support and insightful suggestions. Special thanks also go to other members of my dissertation committee: Dr. Michael Dawson and Dr. Bo Zhang who generally gave their time and expertise to improve my work.

None of my work would have been possible without the constant support from my family and friends. Particularly, my husband, Yinggan Zheng, has shown unconditional love and incredible patience to help me go through many difficult periods.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

By estimating a person's location on an underlying latent continuum, traditional assessments have been effective for selecting students who are most likely to succeed in a particular educational institution or program (Mislevy, 1995). Traditional assessments are typically constructed based on logical taxonomies and content specifications but lack explicit cognitive models of the structures and cognitive processes that underlie student performance (Snow & Mandinach, 1991). As a result, test scores from traditional assessments are tied to content areas rather than the student's cognitive processes measured by test items.

Test theories used for interpreting scores from traditional assessments are designed to optimize the estimate of a student's single score on an underlying latent scale – the true score scale in classical test theory (CTT) or the latent trait scale in item response theory (IRT). A single aggregate score produced using CTT and IRT provides general information about students' locations on a continuum. However, it fails to provide specific information to teachers about their students' cognitive strengths and weaknesses which may, in turn, help teachers make instructional decisions intended to help students succeed in educational settings (Nichols, 1994).

Frustrated by the presence of these two limitations with traditional assessment approaches, measurement specialists have become increasingly interested in the development of new diagnostic assessments that are aimed at uncovering the cognitive processes used by students to respond to test items, determining the nature of poor performance, and classifying the poor performance in terms of an accepted typology of malfunctions (Scriven, 1999). As Nichols (1994) stated:

> These new assessments make explicit the test developer's substantive assumptions regarding processes and knowledge structures a performer in a test domain would use, how the processes and knowledge structures develop, and how more competent performers differ from less competent performers. (p. 578)

New diagnostic assessments should enable researchers and educators to make inferences about the knowledge and processing skills that students use when solving test items. A well-designed diagnostic assessment can measure the different knowledge and skills required to solve test items in a domain of interest, thereby providing a profile of students' mastery and non-mastery of cognitive skills. The value of diagnostic assessment lies in its ability to reveal each student's specific set of cognitive strengths and weaknesses and help design effective diagnostic interventions for individual students.

Chapter 1 of this thesis reviews some currently existing cognitive diagnostic models in the literature. Eight cognitive diagnostic models are presented. Of these eight models, the attribute hierarchy method introduced by Leighton, Gierl, and Hunka (2004) was chosen as the foundation for this research because it brings an important cognitive property, attribute dependency, into cognitive modeling methodologies. Chapter 2 presents a detailed description of the logic and procedures of the attribute hierarchy method and discusses the importance of examining the accuracy and the adequacy of cognitive models in representing students' knowledge structure and skills in the test domain. Chapter 3 introduces a person-fit statistic called the hierarchy consistency index (HCI_i), which is explicitly designed to examine the degree to which a student response vector is consistent with the cognitive model used with the AHM. A simulation approach is then used for setting critical values to determine whether the HCI_i value is sufficiently high to show a statistical fit of the observed response vector to the attribute hierarchy. Chapter 4 presents simulation studies conducted to evaluate the effectiveness of the HCI_i

in determining the degree to which a student response vector is consistent with the cognitive model. Simulated data are also used to investigate whether general guidelines can be developed for identifying good, moderate, and poor fitting student response vectors across different cognitive models by using the HCI_i. Chapter 5 provides a brief summary of the methods and the results from this study, followed by a discussion of how to use substantive analyses to complement the statistical results produced by the HCI_i. The directions for future research are outlined at the end of the chapter.

Cognitive Diagnostic Models: An Overview

Over the past two decades, many cognitive diagnostic models (CDMs) have been proposed (e.g., Dibello, Stout, & Roussos 1995; Fischer, 1973, 1983; Leighton, et. al., 2004; Mislevy, Almond, Yan, & Steinberg, 1999; Tatsuoka, 1983, 1984, 1990, 1995; Whitely, 1980). CDMs serve two purposes: 1) to aid in the development of diagnostic assessments, and 2) to estimate students' profiles associated with different cognitive skills. From a psychometric modeling perspective, most CDMs share a common feature: they model the probability of a correct response to an item as a function of students' attribute profiles associated with different cognitive skills, although the probabilistic models might take different forms. In the following sections, eight CDMs will be briefly reviewed to provide the reader with information regarding the breadth of these models in educational measurement.

Linear Logistic Latent Trait Model

Fischer's (1973, 1983) linear logistic latent trait model (LLTM), which is an extension of the IRT Rasch model, is considered to be the first approach to bring cognitive variables into psychometric models (Stout, 2002). The LLTM is intended to account for the difficulty of test items in terms of a set of underlying cognitive skills, or attributes, hypothetically needed 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:

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

where

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

θ_i = the ability of student 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, student ability is modeled as a unidimensional parameter, θ_i. Since only one ability parameter is specified for each student, the LLTM can not evaluate students with respect to the 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

In an effort to overcome the shortcomings of the LLTM, Embretson (formerly known as Whitely, 1980) proposed a noncompensatory model called the multicomponent latent trait model (MLTM). The MLTM uses subtask responses to measure cognitive attributes underlying test items. In the MLTM, the probability of successful performance on a test item is expressed as the product of probabilities of successful performances on subtasks of the item, each of which follows a separate one-parameter unidimensional IRT model,

$$p(x_{ij} = 1|\boldsymbol{\theta}_{\mathrm{i}}, \mathbf{b}_{\mathrm{j}}) = \prod_{k=1}^{K} p(x_{ijk} = 1|\theta_{ik}, b_{jk}) = \prod_{k=1}^{K} \frac{\exp(\theta_{ik} - b_k)}{1 + \exp(\theta_{ik} - b_k)},$$

where

$\boldsymbol{\theta}_{\mathrm{i}}$ = the vector of K subtask abilities for student i,

$\mathbf{b}_{\mathrm{j}}$ = the vector of K subtask difficulties for item j,

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

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

b_k = the difficulty of subtask k.

By using the multiplicative form of the probabilities for performing each subtask correctly, the MLTM captures the noncompensatory nature of cognitive attributes. Moreover, a student'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 limitation with the MLTM is that this approach requires students' responses to subtasks of each item, which cannot be directly obtained from multiple-choice items. As a result, the usefulness of the MLTM for cognitive diagnosis is, to a large degree, limited.

Bayes Net Approach for Cognitive Diagnosis

The Bayes net approach has been applied to cognitive diagnosis by Mislevy and his colleagues (Mislevy, 1994; Mislevy, et. al., 1999; Mislevy, Steinberg, & Almond, 2003). This approach combines an evidence-centered design and Bayesian inference networks to aid in the development and interpretation of diagnostic assessments. Evidence-centered design consists of three models: the student model, the evidence model, and the task model. The student model specifies the knowledge and skills that should be used to characterize individual students. The evidence model describes the evidence variables required to make inferences about students' knowledge and skills (i.e., the observable item scores), and models the relationship of these evidence variables to student model variables. The task model describes the features of a task that are useful to extract the evidence specified in the evidence model in order to make inferences about students' knowledge and skills. These three conceptual design models are then mathematically translated into probabilistic models using Bayesian inference networks.

The first step of the Bayes net approach to cognitive diagnosis is to define a prior distribution of each student's multidimensional skill vector. An assumption that all students share a common prior distribution is made. The prior distribution could range from vague to precise depending on the strength of prior theory and experience about the nature of the targeted knowledge and skills. The posterior distribution is referred to as the updated distribution of a student's skill vector based on the evidence of a student's performance. Once the posterior distribution is estimated, summaries of the posterior means and variances can be used to make inferences about students' knowledge and skills.

Rule Space Model

Another important cognitive diagnostic model is Tatsuoka's (1983, 1984, 1990, 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). Broadly speaking, the rule space model is composed of 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:

$$Q_{7,11} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

This matrix consists of 7 rows and 11 columns, with each row corresponding to an attribute and each column corresponding to an item. The first column of this matrix shows that item 1 is measuring attributes 1 and 2. The second column indicates that item 2 is measuring attributes 1, 2, and 3. The rest of columns can be interpreted in the same manner. In the rule space model, a student must have mastered all the attributes that an item is measuring in order to answer the item correctly. Therefore, in order to answer item 1 correctly, the student must have mastered attributes 1 and 2. In the rule space model, 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 students' observable response pattern by using Boolean description functions (Tatsuoka, 1991; Varadi & Tatsuoka, 1992). In the rule space model, each cognitive attribute is dichotomized as mastered or nonmastered. As a result, knowledge states, used to describe students' profiles of cognitive skills, are represented by a list of mastered/nonmastered 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 students' 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 student cognition and psychometric applications.

The Unified Model

Inspired by 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 behavior 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 in terms of discrete cognitive states and a continuous latent ability (Dibello et al., 1995). In the unified model, each student is characterized by a dichotomous vector $\boldsymbol{\alpha}_{\mathbf{i}}$ representing the student'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 an attribute for the item (corresponding to the possibility that a student who possesses an attribute may fail to apply it correctly to an item and a student who lacks the attribute may still answer the item correctly by possessing partial knowledge), and (4) the possibility that a student makes a random error. The unified model incorporates these four sources of variation into the following equation for the item response probability:

$$p(x_{ij}=1|\theta_i,\boldsymbol{\alpha}_{\mathbf{i}})=(1-p)\{d_j\prod_{k=1}^{K}\pi_{jk}^{\alpha_{ik}}r_{jk}^{(1-\alpha_{ik})}p_j(\theta_i+\Delta c_j)+(1-d_j)p_j(\theta_i)\},$$

where

$p =$ probability of making a random error,

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

$\alpha_{ik} =$ the k^{th} element of vector $\boldsymbol{\alpha}_{\mathbf{i}}$,

$c_j =$ completeness index of attributes required for item j,

$\pi_{jk} = P(\text{Attribute } k \text{ applied correctly to item } j | \alpha_{\text{ik}} = 1)$,

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

$\Delta = 2$, and

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

Analogous to Embretson's MLTM, the unified model captures the noncompensatory nature of cognitive attributes in the sense that the probability of successful performance on an item using the Q matrix strategy is expressed 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 attempt 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; C. Tatsuoka, 2002) and the noisy input deterministic and gate model (NIDA) (Junker & Sijstma, 2001). The DINA model partitions students 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 a student fails to answer the item correctly when the student has mastered all the required attributes (s_j, the "slipping" parameter) and the probability that a student gets the correct answer when the student does not possess all of the required attributes (g_j, the "guessing" parameter). The item response probability can be written as:

$$p(x_{ij} = 1 | \xi_{ij}, s_j, g_j) = (1 - s_j)^{\xi_{ij}} g_j^{(1-\xi_{ij})},$$

where x_{ij} is the response of student 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:

$$p(x_{ij} = 1 | \mathbf{\alpha_i}, \mathbf{s}, \mathbf{g}) = \prod_{k=1}^{K} [(1 - s_k)^{\alpha_{ik}} g_k^{(1-\alpha_{ik})}]^{q_{jk}},$$

where

$\mathbf{\alpha_i} =$ the vector of the attribute profile for student i,

$\mathbf{s} =$ the vector of attribute slipping parameters,

$\mathbf{g} =$ the vector of attribute guessing parameters,

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

$\alpha_{ik} =$ the k^{th} element of vector $\mathbf{\alpha_i}$.

The Attribute Hierarchy Method

All of the cognitive diagnostic models just described require the specification of the Q matrix, which requires researchers to describe test items using a presumed set of attributes. However, the Q matrix does not provide 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 independently but function as a network of interrelated processes (e.g., Kuhn, 2001; Vosniadou & Brewer, 1992). As a result, it is necessary to build the relationships or dependencies among attributes into cognitive diagnostic models and integrate this information into the statistical pattern classification procedures.

The attribute hierarchy method (AHM) (Leighton et. al., 2004; also see Gierl, Leighton, & Hunka, 2000), which is based on the assumption that test items can be described by a set of hierarchically ordered attributes, is a cognitive diagnostic model designed explicitly to model related cognitive skills underlying academic problem solving. In the AHM, attributes are considered to be hierarchically related and therefore can be ordered into a hierarchy based upon their logical and/or psychological properties. The attribute hierarchy can be used as a basis for the development of test items. After the test items are administered to students, vectors of binary responses (1 or 0) that take into account the dependencies of the attribute hierarchy are produced. In turn, a student's response vector is then used to estimate the student's probability of mastery and nonmastery of the attributes illustrated in the attribute hierarchy.

Purpose of Current Study

The AHM (Leighton, et. al., 2004) is designed explicitly to integrate cognitive models with a psychometric technique to model students' test performances and estimate their mastery of domain knowledge and cognitive skills. The AHM, by incorporating the assumption of attribute dependency, brings an important cognitive property into cognitive diagnostic models based on the Q matrix. The validity of this new diagnostic model depends critically on the accuracy and the adequacy of the attribute hierarchy. For example, if the cognitive attributes summarized in the attribute hierarchy do not correspond to any real aspects of the cognitive processes used by each student, then any diagnoses of the student produced with the attribute hierarchy will be meaningless. In addition, the inclusion of superfluous attributes in the attribute hierarchy may lead to a high misclassification rate due to the unnecessarily high complexity of the model in terms of the large number of deceptive knowledge states around students' true states. Furthermore, a model that fails to include some of the important attributes will not provide sufficient diagnostic information to permit test users to develop and implement interventions designed to maintain students' cognitive strengths and address students' cognitive weaknesses.

In order for the AHM to produce cognitively and statistically valid results, it is important for the attribute hierarchy to be supported by both psychological and statistical evidence which demonstrates that students' problem-solving behavior has been measured. To date, such evidence is limited. Consequently, methods for assessing the accuracy and adequacy of the attribute hierarchy in describing the cognitive processes used by students to solve test items must be developed. One method for doing this is to employ a person-fit statistic. Generally, methods for evaluating the misfit of a student response vector to the hypothesized item-score vector have been referred to as "person-fit" methods. Numerous person-fit statistics based on CTT and IRT have been proposed and investigated (e.g., Donlon & Fischer, 1968; Levine & Rubin, 1979; Meijer, 1994; Meijer & Sijtsma, 1995, 2001; Sijtsma, 1986; Sijtsma & Meijer, 1992; Tatsuoka & Tatsuoka, 1983; van der Flier, 1982; Wright & Stone, 1979). However, as will be discussed in Chapter 3, most of these methods are based on a single estimate of student ability on the true score scale or the latent trait scale without referring to the mastery and nonmastery of a set of attributes that underlie student performances. Therefore, it is inadequate to directly use these existing person-fit statistics with the AHM.

Consequently, the first purpose of the current study was to develop and validate a person-fit statistic called the hierarchy consistency index (HCI_i; Cui, Leighton, Gierl & Hunka, 2006). The HCI_i is designed to examine explicitly the degree to which a student response pattern is consistent with the attribute hierarchy. The second purpose of the current study was to conduct simulation studies to assess the effectiveness of the HCI_i in determining the degree to which a student response vectors fits the attribute hierarchy and to identify critical values for interpreting the HCI_i with different types of hierarchies, number of attributes, and sample size, and if the HCI_i was influenced by these factors.

Chapter 2: Review of the Attribute Hierarchy Method

The AHM is a cognitive diagnostic model designed to help develop cognitive diagnostic assessments and estimate students' profiles that reflect their mastery of a set of hierarchically ordered attributes. Based on the rule space approach (Tatsuoka, 1983, 1984, 1990, 1995), the AHM represents an important variation by explicitly modeling attribute dependency. In the AHM, attributes are assumed to be hierarchically related, and therefore, the attributes can be ordered based upon their logical and/or psychological properties. The AHM is composed of three sequential stages. In the first stage, the attribute hierarchy is defined to describe the knowledge structures and skill processes that students need to use in the test domain. This is a critical step because the validity of the cognitive model links directly to the accuracy of the inferences to be made about students with the AHM. In the second stage, the attribute hierarchy is used as a basis for developing test items to ensure that each component of the cognitive model has been measured adequately with test items. In the third stage, statistical classification procedures are used to classify each student into one of the knowledge states, derived from the cognitive model, thereby making specific inferences about students' cognitive strengths and weaknesses. In order to familiarize the reader with the AHM, this chapter provides a detailed description of the three stages of the AHM.

Stage 1: Defining the Attribute Hierarchy

The first step in using the AHM for cognitive diagnosis is to define the attribute hierarchy that serves as a cognitive model in the domain or for the task of interest. Leighton and Gierl (2007) identified three types of cognitive models that could be used in educational measurement, including cognitive models of domain mastery, cognitive models of test specification, and cognitive models of task performance. A model of domain mastery describes the population of knowledge and skills associated with competence in a test domain. Curriculum-based tests developed by teachers for the purpose of formative evaluation are considered as using the cognitive model of domain mastery. In order to thoroughly evaluate students' domain mastery, multiple tests need to be administered to students. The use of multiple tests can help teachers gain a detailed picture of what their students know and can do within the test domain. According to Leighton and Gierl (2007), however, cognitive models of domain mastery that are underlying these tests cannot provide strong support for making inferences about students' cognitive strengths and weaknesses given that they fail to clearly specify the cognitive processes underlying student performance. Although a student answers an item correctly, one cannot conclude that the student uses a correct strategy in solving the item. Therefore, cognitive models of domain mastery typically will indicate that a student can exhibit a certain test response.

A model of test specification is commonly used in large-scale assessments designed to rank students on a continuum within a test domain. Cognitive models of test specification are often generated by test developers and content specialists. Although, test specification attempts to specify the knowledge and skills that students are supposed to use as they solve test items, substantial evidence is often missing for determining whether students actually use them. Therefore, cognitive models of test specification fail to provide an explicit description of the knowledge structures and cognitive processes that students use in solving test items. In addition, the substantial financial costs associated with the administration of large-scale assessments make it impossible to use multiple tests to extensively evaluate the population knowledge and skills. As a result, only a sample of knowledge and skills can be evaluated by large-scale assessments. Given the presence of these two limitations with cognitive models of test specification, the grain size of these models is often relatively large thereby weakening the specificity of inferences made about students' cognitive strengths and weaknesses.

A cognitive model of task performance is generated based on empirical studies that examine the knowledge and cognitive skills used by students as they solve test items. The collection of think-aloud verbal reports could play an important role in generating cognitive models of task performance. For example, a cognitive model of task performance can be generated by administering students a set of test items and having them think aloud as they solve these items. In doing so, a detailed description of students' cognitive steps or processes is obtained and consequently cognitive models of task performance are created with a small grain size. Other methods, such as experimental study and the evaluation of expert judges can also be used to generate cognitive models of task performance. According to Leighton and Gierl (in press), the grain size of a cognitive model is associated to the type of inferences made about student performance. Since cognitive models of task performance should illustrate the detailed knowledge and skills students actually use as they answer test items, assessments based on these models can be used to make specific inferences about students' strengths and weaknesses.

In the AHM, attributes are defined as basic cognitive processes or skills required to solve test items correctly (Leighton et al., 2004). The attribute hierarchy serves as a cognitive model that specifies the knowledge and skills required for students to answer each item correctly. In order to use the AHM to make inferences about student cognitive strengths and weaknesses, the attribute hierarchy must be generated at a relatively small grain size and also be supported by empirical evidence that demonstrates students actually use the model-specified knowledge and skills in solving problem-solving tasks or test items.

As pointed out by Leighton et al. (2004), methods from cognitive psychology, such as task and protocol analysis, play an important role in the identification of attributes and the formation of the attribute hierarchy in a domain. Many studies have been conducted to identify the attributes required for successful performance on test items and tasks. For example, in a language testing study, Buck and Tatsuoka (1998) identified the attribute set for a 35-item listening comprehension test by using two main sources: an extensive literature review to seek the theoretical and empirical evidence for the attributes that affect performance on listening tests and the results from a series of verbal protocol studies conducted by Buck (1990, 1991, 1994) for examining the second language listening processes.

Once identified, the attributes need to be organized into a hierarchy. This is a major difference between the rule space model and the AHM in that the hierarchy reflects different assumptions about the relationships among attributes. Gierl (2007) discussed extensively the differences between the rule space model and the AHM. In the rule space model, the attributes are not necessarily related to each other and could operate independently. For example, Tatsuoka, Birenbaum, Lewis, and Sheehan (1993) conducted a task analysis of the SAT mathematics test and produced 14 independent attributes, which accounted for 75% of the total variance of the IRT item difficulties using multiple regression. In the AHM, however, the attributes are assumed to be hierarchically related. As explained by Leighton et al. (2004), the assumption of attribute dependency is consistent with the conclusion that "cognitive skills do not operate in isolation but belong to a network of interrelated competencies (Kuhn, 2001; Vosniadou & Brewer, 1992)" (p. 209).

The ordering of the attributes into a hierarchy should be based on "empirical considerations (e.g., a series of well defined, ordered cognitive steps identified via protocol analysis) or theoretical considerations (e.g., a series of developmental sequences suggested by Piaget such as preoperational, concrete operational, and formal operational)" (Leighton et al., 2004, p. 209). Since the attribute hierarchy represents the underlying construct of test items, the validity of the AHM depends critically on the correct identification of the attribute hierarchy. Leighton et al. (2004) described four types of hierarchy structures – divergent, convergent, linear, and unstructured structures. Different types of hierarchy structures describe distinct orderings of cognitive competencies required to solve problems successfully in a specific domain. A divergent structure represents hierarchies with divergent branches. This type of hierarchy is commonly present for a test domain with multiple divergent cognitive competencies. A hierarchy with convergent or linear structure ends at a single point, representing a test domain that involves a single end state of mastery. An unstructured hierarchy is present for a test domain with competencies that are not related to one another.

Stage 2: The Construction of Test Items and Student Knowledge States

In this stage, a series of matrices (e.g., the adjacency, reachability, incidence, and reduced *Q* matrices) initially introduced by Tatsuoka (1983, 1984, 1996) are derived from the attribute hierarchy to facilitate the development of test items and the construction of students' potential knowledge states in terms of their attribute profiles if the attribute hierarchy is true.

Representing the Attribute Hierarchy

Once identified, the attribute hierarchy can be mathematically represented by a binary adjacency matrix (A) of order $K \times K$, where K is the number of attributes. In the adjacency matrix, the direct relationship between each pair of attributes is specified. The element a_{ij} of the adjacency matrix indicates if attribute i is a direct prerequisite of attribute j. It can be expressed as follows:

$$a_{ij} = \begin{cases} 1 & \text{if attribute } i \text{ is the prerequisite of attribute } j \\ 0 & \text{otherwise} \end{cases}.$$

For example, consider the attribute hierarchy illustrated in Figure 1. Attribute 1 is the direct prerequisite of attribute 2. Attribute 2 is in turn the direct prerequisite of attribute 3 and 4. And attribute 4 is the direct prerequisite of attribute 5 and 6. This hierarchical configuration is represented in a 6 X 6 adjacency matrix, where the elements $a_{12}, a_{23}, a_{24}, a_{45}, a_{46}$ of the adjacency matrix are 1.

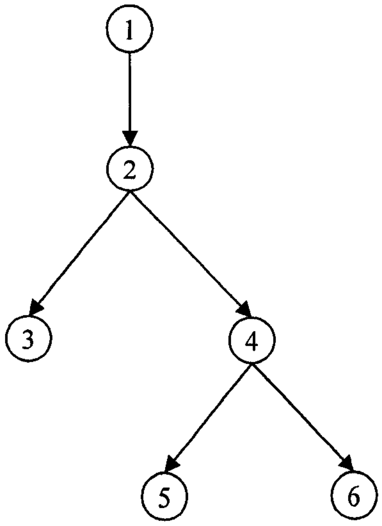

Figure 1. A Six-attribute Hierarchy

The adjacency matrix of the attribute hierarchy is given below:

$$A_{6,6}=\begin{bmatrix}0&1&0&0&0&0\\0&0&1&1&0&0\\0&0&0&0&0&0\\0&0&0&0&1&1\\0&0&0&0&0&0\\0&0&0&0&0&0\end{bmatrix}.$$

Row 1 of the adjacency matrix represents attribute 1 and the elements of row 1 show the attributes which are directly connected to attribute 1. Row 2 shows that attributes 3 and 4 are directly connected to attribute 2. The rest of rows can be interpreted in the same manner.

It should be noted that the adjacency matrix only expresses the direct relationship between attributes. To specify the direct and indirect relationship among attributes, a reachability matrix (R) of order $K \times K$ is used. To derive the reachability matrix from the adjacency matrix, Boolean addition and multiplication are performed on the adjacency matrix. Boolean addition is defined by 1 + 1 = 1, 1 + 0 = 1, 0 + 1 = 1, and 0 + 0 = 0. Boolean multiplication is defined by 0 x 0 = 0, 1 x 0 = 0, 0 x 1 = 0, and 1 x 1 = 1. The reachability matrix can be obtained using the equation $R = (A + I)^n$, where I is an identity matrix of order $K \times K$, and n is the integer between 1 and K that leads R to become invariant. That is, when $(A + I)$ is multiplied by itself repeatedly using Boolean algebra until the product become invariant, the obtained matrix is the reachability matrix. The 1s of the j^{th} row of the reachability matrix identify all the attributes for which attribute j is the direct or indirect prerequisite. For example, to calculate the reachability matrix for the attribute hierarchy in Figure 1, $(A_{6,6} + I)^n$ is calculated for $n = 1,2,3,4$ separately. Since $(A_{6,6} + I)^3 = (A_{6,6} + I)^4$, then $R_{6,6} = (A_{6,6} + I)^3$, which is shown below:

$$R_{6,6} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The first row of the reachability matrix indicates that attribute 1 is the direct or indirect prerequisite for all attributes. Row 2 shows that attribute 2 is the direct or indirect prerequisite of attributes 2, 3, 4, 5, and 6. Row 3 shows that attribute 3 is a direct prerequisite of itself but is neither a direct or indirect prerequisite of any of the other attributes. The rest of the rows can be interpreted in the same way. In the AHM, the reachability matrix is used to select a subset of items from the potential pool of items, which correspond to the dependencies of the attribute hierarchy.

In order to have maximum control over the attributes each item measures, Leighton et al. (2004) suggested that the attribute hierarchy should be identified prior to the development of test items. In other words, the attribute hierarchy should be used to guide the development of test items. When test items are not developed based on an attribute hierarchy, the hierarchy has to be extracted from existing items and it becomes problematic to ensure that all relevant knowledge and skills have been identified correctly. It is difficult in this situation to obtain a unique adjacency matrix in which the direct relationships among attributes are specified. In addition, the extraction of the attribute hierarchy from actual test items could also lead to the problem of the nonidentifiability of certain attributes, when items needed to reflect the relationships in the hierarchy are missing from the set of existing test items. Therefore, the construction of test items based on the attribute hierarchy in a domain of interest can improve the interpretability of student performance on test items.

Creating Potential Item Pool

The potential item pool is designed as the set of items that measure all the possible combinations of attributes when the attributes are assumed to be independent of each other. In this case, the adjacency matrix is a matrix of order $K \times K$ in which all the elements are 0, and the reachability matrix is a $K \times K$ identity matrix, where K is the number of attributes. The number of items in the potential item pool is $2^K - 1$. The potential item pool is represented by the incidence or Q matrix (Tatsuoka, 1983, 1984, 1996), which is an attribute-by-item matrix, which is of order $K \times (2^K - 1)$. In the Q matrix, each column represents one item, and the 1s in the column identify which attributes are required for successful performance on this item. The columns of the Q

matrix are obtained by converting the integers ranging from 1 to $2^K - 1$ to their binary form.The Q matrix for the six-attribute hierarchy (shown in Figure 1) is given by:

$$Q_{6,63} = \begin{bmatrix} 101010101010101010101010101010101010101010101010101010101010101 \\ 011001100110011001100110011001100110011001100110011001100110011 \\ 000111100001111000011110000111100001111000011110000111100001111 \\ 000000011111111000000001111111100000000111111110000000011111111 \\ 000000000000000111111111111111100000000000000001111111111111111 \\ 000000000000000000000000000000011111111111111111111111111111111 \end{bmatrix}.$$

For this six-attribute hierarchy, the number of items (columns) in the Q matrix is $2^6 - 1 = 63$, and therefore the Q matrix is of order 6×63. Column 1 of the Q matrix represents item 1, and it identifies that only attribute 1 is required in order for students to correctly respond to this item. Conversely, according to column 63 of the Q matrix, item 63 requires all six attributes for a successful response. The rest of the columns can be interpreted in the same manner.

Reducing the Potential Item Pool

As discussed in the previous section, the size of the potential item pool is equal to $2^K - 1$ when the attributes are assumed to be independent of each other. Hence, even for a small number of attributes, the potential item pool will be fairly large. However, 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, column 2 of the Q matrix is (010000), which indicates that only attribute 2 is required to correctly answer the item represented in this column. According to the reachability matrix of the attribute hierarchy, however, attribute 2 requires attribute 1. Therefore, a student must have mastered both attribute 1 and 2 in order for the student to correctly answer item 2. That is, item 2 must be represented by (110000), which is identical to the item represented by column 3 of the Q matrix. As a result, column 2 of the Q matrix can be removed. The removal of items in this manner ultimately produces a reduced Q matrix that reflects the dependency among attributes.

Alternatively, the reduced Q matrix can be derived using Boolean addition to remove items that do not match the constraints of the reachability matrix. For example, column 6 of the reachability matrix specifies that any item that probes attribute 6 must also measure attributes 1, 2, and 4. If the item does not measure these additional 3 attributes, the item will not match the attribute hierarchy and, consequently, will be removed. The reduced Q matrix of the attribute hierarchy shown in Figure 1 is as follows:

$$Q_{R_{6,11}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The reduced Q matrix shown above is of order 6×11. Thus, out of a potential pool of 63 items, if the attribute hierarchy is true, only 11 items are logically meaningful according to the attribute hierarchy shown in Figure 1. As explained by Leighton et al. (2004), the reduced Q matrix has a particularly important meaning for test development. It should be used as the cognitive specifications for test construction. For the attribute hierarchy shown in Figure 1, at least 11 items need to be developed based on the derived reduced Q matrix in order to achieve maximum diagnostic information. Multiple sets of items can be used to increase the number of items for ensuring the reliability of the test.

By describing the cognitive requirements of the domain of interest with the attribute hierarchy and specifying the items needed to measure the domain in the reduced Q matrix, the AHM makes a direct link between student cognition and the test design.

Generating Expected Response Patterns

Once the reachability matrix and the reduced Q matrix are identified, expected response patterns can be derived. 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." For example, a student who only possesses attribute 1 is expected to answer item 1 correctly and the rest of the items incorrectly. Conversely, if a student has mastered all attributes, the student is expected to correctly answer all the items in the reduced Q matrix, providing the hypothesized attribute hierarchy is true. As shown in the second column of Table 1, eleven expected response patterns are derived from the attribute hierarchy illustrated in Figure 1. The second row of Table 1 can be interpreted as follows: the attribute pattern (100000), which indicates a student has only mastered attribute 1, should produce the expected response pattern (10000000000) and obtain a total score of 1. Similarly, row 3 of Table 1 indicates that a student who has mastered attributes 1 and 2 is expected to correctly answer the first and second item correctly. It should be noted that two students with an equal total score do not necessarily possess the same attribute patterns. For example, a total score of 4 can be produced from attribute patterns (110110) or (110101). Therefore, a student's total score cannot be consistently associated with a single attribute pattern. In order to identify students' cognitive strengths and weaknesses, total scores are not sufficient. Students' attribute patterns must be estimated to indicate which attributes are absent and what remediation instructions are required to help students learn their unmastered attributes. Hence, the attribute patterns yielded by the AHM can provide more specific information regarding students' cognitive strengths and weaknesses than the single score derived from item response theory (IRT) or classical test theory (CTT).

Table 1

Expected Response Patterns for the Hierarchy Shown in Figure 1

Attribute Pattern	*Expect Response Pattern*	*Total Score*
100000	1 0 0 0 0 0 0 0 0 0 0	1
110000	1 1 0 0 0 0 0 0 0 0 0	2
111000	1 1 1 0 0 0 0 0 0 0 0	3
110100	1 1 0 1 0 0 0 0 0 0 0	3
111100	1 1 1 1 1 0 0 0 0 0 0	5
110110	1 1 0 1 0 1 0 0 0 0 0	4
111110	1 1 1 1 1 1 1 0 0 0 0	7
110101	1 1 0 1 0 0 0 1 0 0 0	4
111101	1 1 1 1 1 0 0 1 1 0 0	7
110111	1 1 0 1 0 1 0 1 0 1 0	6
111111	1 1 1 1 1 1 1 1 1 1 1	11

Stage 3: Classifying the Observed Response Patterns

In real testing situations, it is possible that a student, who has not mastered all the attributes required by an item, can still answer the item correctly by guessing or by having partial knowledge. It is also possible that a student, who has mastered all the attributes that an item is probing, might reach the wrong answer due to careless mistakes. Therefore, the observed student response vectors might consist of slips of the form 1 to 0 or 0 to 1. By classifying each observed response vector in the presence of slips into one of the expected response patterns, students' attribute mastery can be estimated.

Leighton et al. (2004) proposed two methods for the classification of observed response patterns in the AHM. In these two methods, the probability of a correct response to individual items is calculated for each expected response pattern using an IRT model. The three-parameter logistic IRT model is given by:

$$p(x_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + \frac{1 - c_j}{1 + e^{-1.7a_i(\theta_j - b_j)}},$$

where

a_j = 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 student i.

The two-parameter logistic IRT model is a special case of the three-parameter model in which the c_j parameter is set to 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 & Bock, 1990).

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 each of the expected response patterns to identify the slips from 1 to 0 and from 0 to 1. The likelihood of all slips from 1 to 0 and from 0 to 1 for student i is given by:

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

where

$P_{jk}(\theta_j)$ = the probability of a correct response to item k using the ability parameter for expected response pattern j,

$P_{jm}(\theta_j)$ = the probability of a correct response to item m using the ability parameter for expected response pattern j,

S_{i0} = the subset of items with slips from 0 to 1 for the observed response vector of student i, and

S_{i1} = the subset of items with slips from 1 to 0 for the observed response vector of student i.

The higher the value of $P_{ijExpected}(\theta_j)$ calculated by comparing the observed response vector to the expected response vector j, the more likely the observed response pattern originates from this expected response vector. Therefore, the observed response vector will be classified as originating from expected response vector j when the maximum value of $P_{ijExpected}(\theta_j)$ is achieved.

In Method B, the expected response patterns that are logically included in the observed student response vector are identified and a student is considered to possess all attributes logically included within his or her observed response vector. For those expected response patterns that are not logically included in the observed vector, the likelihood of slips only from 1 to 0 is calculated and compared to a cut-point assigned by researchers. The likelihood of slips from 1 to 0 is given by:

$$P_{ijExpected}(\theta_j) = \prod_{k \in S_{i1}} [1 - P_{jm}(\theta_j)].$$

If an expected response vector's likelihood value is greater than the cut-point, it is concluded that the student has mastered the attributes implied by this expected response vector.

Another method of analyzing the students' response patterns is to employ a neural network (Gierl, Cui, & Hunka, 2007) to approximate the functional relationship between students' item responses and their attribute mastery profiles. The power of the neural network approach lies in its ability to map any relationship between inputs and outputs (Dawson, 1998, 2004; Lippmann, 1987; Medler, 1998). An example of a neural network is presented in Figure 2.

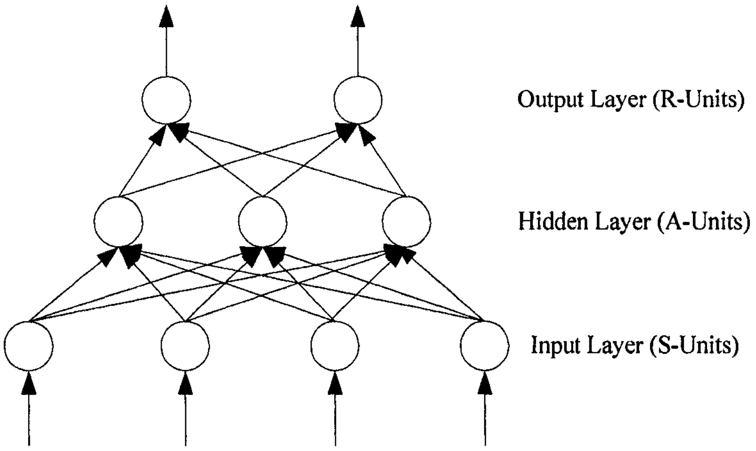

Figure 2. A Neural Network with Three Layers

This neural network contains three parallel layers – input, hidden, and output – where each unit in the input layer is connected to each unit in the hidden layer and each unit in the hidden layer, in turn, is connected to each unit in the output layer. The arrows denote these connections. The purpose of the network is to establish the functional relationship between input and output units, so the exemplars from the input layer are optimally associated with their responses in the output layer, as indicated by a goodness-of-fit measure which is typically an error term.

In the AHM, the input units of the neural network are students' responses to test items, while the output units are the probabilities that the student possesses individual attributes illustrated in the attribute hierarchy. The exemplars used to train the neural network are the expected response vectors derived from the attribute hierarchy while the target output is their associated attribute patterns assuming that the hierarchy is true. The relationship between the expected response vectors with their associated attribute vectors is established by presenting each expected response pattern to the network repeatedly until the error term of the neural network reaches an acceptable level. Once the relationship between input and output units are established successfully, a set of weight matrices are produced to transform any observed response vector to its associated attribute vector so the attribute probabilities can be computed. Let

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

and

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

then the attribute probability for attribute k, M_k^*, is given as

$$M_k^* = F(a_k),$$

where

q is the total number of hidden units,

v_{kj} 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 strength of using a neural network approach with the AHM is that this approach does not rely on IRT models or any assumptions about the distributional properties of the parameters. Rather, this approach can be used to estimate the probabilities that students have mastered each attribute by minimizing the error associated with the estimation. For a detailed description of using the neural network approach in the AHM, readers are referred to Gierl, Cui, and Hunka (2007).

Evaluating Some Other Issues in Educational Measurement with the AHM

Efforts have been made to evaluate other issues with the AHM. For example, Gierl, et al. (2007) described the concept of attribute reliability and developed a new procedure to assess it in the AHM framework. Additionally, Gierl, Zheng, and Cui (in press) used the AHM to identify and interpret differential group performance on tests. These new developments of the AHM are briefly described next.

Attribute Reliability

As described by Gierl et al. (2007), attribute reliability refers to the consistency of the decisions made in a diagnostic test about students' mastery of specific attributes. The reliability of an attribute is estimated by calculating the ratio of true score variance to observed score variance on the items that are probing each attribute. In the AHM, an item is often 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. In order to isolate the contribution of each attribute to an examinee's item-level performance, the item score is weighted by the subtraction of two conditional probabilities. The first probability is associated with attribute mastery (i.e., the probability that an examinee who has mastered the attribute can answer the item correctly) and the second probability is associated with attribute non-mastery (i.e., the probability that an examinee who has not mastered the attribute can answer the item correctly). The calculation of these probabilities was discussed in detail by Gierl et al. (2007), which will not be repeated here. The weighted scores for items that measure the attribute are used in the reliability calculation by adapting Cronbach's alpha for the AHM framework. The derived formula is given by

$$\alpha_i = \frac{k_i}{k_i - 1}\left[1 - \frac{\sum_{j \in S_i} W_{ij}^{\,2} \sigma^2{}_{X_j}}{\sigma^2{}_{\sum_{j \in S_i} W_{ij} X_j}}\right],$$

where

α_i is the reliability of attribute i,

S_i denote the subset of items that measures attribute i

k_i is the number of items that are probing attribute i in the Q_r (i.e., the number of elements in S_i),

$\sigma^2{}_{X_j}$ is the variance of the observed scores on item j,

$\sum_{j \in S_i} W_{ij}^{\,2} X_j$ is the weighted observed total score on the items that are measuring attribute i, and

$\sigma^2{}_{\sum_{j \in S_i} W_{ij} X_j}$ is the variance of the weighted observed total scores.

This approach can provide information about attribute consistency in the measurement process and help determine whether more items are required in order to make consistent inferences about student's attribute-level performance.

Attribute Differential Functioning

Gierl et al. (in press) described a four-step procedure for estimating and interpreting group differences using the AHM. In the AHM, the hierarchy serves as a cognitive model that specifies the attributes students use in solving test items. As a result, the attribute hierarchy can guide the study of cognitive factors that produce differential performance by systematically evaluating which attributes elicit group differences. Attribute-level differential functioning, hereafter referred to ADF, can be evaluated on a *studied attribute* by comparing the probabilities that different groups possess this attribute. To ensure the ability of examinees from the focal and reference groups are comparable before the studied attribute is evaluated, examinees' score are aligned on the *matching attributes*. ADF occurs when examinees with the same matching attribute pattern but from different groups have unequal probabilities responding to items that measure the studied attribute.

An ADF analysis has four steps. In step 1, the attribute hierarchy is used to generate hypotheses about the nature of attribute-related group differences so the studied and matching attributes are identified. The ordering of the attributes provides a logical basis for generating ADF hypotheses because the hierarchy specifies the ordered dependencies among the attributes according to an underlying cognitive model of task performance. In step 2, the probability that examinees have mastered the studied attribute in both the focal and reference groups is estimated using the neural network. In step 3, the scores for examinees in the focal and reference groups are aligned using the matching attributes. In step 4, the magnitude and direction of group differences on the studied attribute are estimated and tested.

The ADF analysis can potentially bridge the gap between the substantive and statistical steps commonly applied in DIF detection so group differences can be more easily identified statistically and interpreted substantively.

Summary

By incorporating the assumption of attribute dependency, the AHM brings a fundamentally important cognitive feature into cognitive diagnostic models. The AHM can be used to estimate the specific patterns of attribute mastery underlying students' observed item responses. In order for the AHM to be used to make valid inferences about students, however, it is critical to correctly identify the attribute hierarchy in the domain of interest. It is unavoidable that students make slips (from 1 to 0 or from 0 to 1) in answering test items, which leads to the inconsistency between students' observed response pattern and the expectations of the given attribute hierarchy. Therefore, person-fit statistics are needed to explicitly evaluate the degree to which student response vectors are consistent with the attribute hierarchy thereby assessing the accuracy and adequacy of the attribute hierarchy in describing the knowledge and cognitive skills individual students use in solving test items. A good fit of the student response vector relative to the attribute hierarchy suggests that the student uses the knowledge and cognitive skills as specified in the attribute hierarchy to solve test items. As a result, the inferences to be made about the student's cognitive strengths and weaknesses with the AHM can be validated.

Chapter 3: The Hierarchy Consistency Index

This chapter is divided into three sections. In the first section, currently existing person-fit statistics are reviewed, followed by a discussion of why these statistics cannot be directly used in the AHM framework. In the second section, a person-fit statistic, called the hierarchy consistency index (HCI_i; Cui et al., 2006), is introduced. The HCI_i is designed explicitly to examine the degree to which an observed student response pattern is consistent with the attribute hierarchy. In the third section, a simulation approach that was used to identify critical values in order to statistically test the significance of the HCI_i is described.

A Review of Existing Person-fit Statistics

The validation of the underlying construct that is being measured by a test is one of the most important aspects in educational measurement. It is fundamentally important to investigate whether a student's item scores can be predicted or interpreted by the construct that is being measured. One way to accomplish this is to assess whether the pattern of a student's item responses fit one of the typical item-score patterns that are consistent with the test model used in the development and interpretation of test items. Attempts to evaluate the misfit of a student's item-score vector to the test model have led researchers to studies of "person-fit" statistics. Numerous person-fit statistics have been proposed and investigated, and each has its advantages and disadvantages (e.g., Donlon & Fischer, 1968; Harnisch & Linn, 1981; Kane & Brennan, 1980; Levine & Rubin, 1979; Meijer, 1994; Meijer & Sijtsma, 2001; Sijtsma, 1986; Sijtsma & Meijer, 1992; Tatsuoka & Tatsuoka, 1983; van Der Flier, 1982; Wright & Stone, 1979). These person-fit statistics are grouped into two major categories: group-dependent statistics and IRT-based statistics.

Group Dependent Person-fit Statistics

In calculating group dependent person-fit statistics, items are rearranged and numbered according to a decreasing proportion-correct score (increasing item difficulty) in classical test theory (CTT): $\pi_1 > \pi_2 > ... > \pi_J$, where J is the number of items in a test and π_j is the proportion-correct score on item *j*. Group dependent person-fit statistics compare the observed item response vector to the expectation under Guttman's (1944, 1950) deterministic model, in which the probability that a student correctly answers a relatively difficult item but fails to answer a relatively easy item is assumed to be zero. That is, if a student's number-correct score is r, the student is expected to have answered the first r easiest items correctly. A response vector is considered as misfitting when items with a relatively low proportion-correct score are answered correctly, and items with a relatively high proportion-correct score are answered incorrectly. For example, Harnisch and Linn (1981) discussed the modified caution index C_i^*:

$$C_i^* = \frac{\sum_{j=1}^{r} \pi_j - \sum_{j=1}^{J} x_{ij}\pi_j}{\sum_{j=1}^{r} \pi_j - \sum_{j=J-r+1}^{r} \pi_j},$$

where

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

π_j = the proportion-correct score on item j .

When a student has a number-correct score r and answers the r easiest items correctly and the rest of the items incorrectly,

$$C_i^* = \frac{\sum_{j=1}^{r} \pi_j - [\sum_{j=1}^{r} 1 \times \pi_j + \sum_{j=r+1}^{J} 0 \times \pi_j]}{\sum_{j=1}^{r} \pi_j - \sum_{j=J-r+1}^{r} \pi_j} = 0,$$

indicating the response vector of student i fits the model perfectly. Conversely, when the student answers the r most difficult items correctly and the rest of the items incorrectly,

$$C_i^* = \frac{\sum_{j=1}^{r} \pi_j - [\sum_{j=1}^{r} 0 \times \pi_j + \sum_{j=r+1}^{J} 1 \times \pi_j]}{\sum_{j=1}^{r} \pi_j - \sum_{j=J-r+1}^{r} \pi_j} = 1,$$

indicating a maximum misfit. Tatsuoka and Tatsuoka (1983) proposed a person-fit statistic called the norm conformity index NCI_i:

$$NCI_i = 1 - \frac{2\sum_{j=1}^{J-1} \sum_{h=j+1}^{J} x_{ij}(1 - x_{ih})}{r(J-r)},$$

where

J = the total number of items,

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

r = student i's number-correct score.

The NCI_i evaluates the misfit of an observed response vector to the test model by comparing the student' responses for each item pair with the Guttman pattern. There are many other group dependent person-fit statistics, such as Kane and Brennan's (1980) agreement, disagreement, and dependability indices, and van der Flier's (1982) U_3 statistic.

Group dependent person-fit statistics rely on item difficulty as determined by the proportion correct score of a group of students. In the AHM, however, item complexity is associated to a set of hierarchically ordered attributes. The evaluation of the misfit of observed item responses to the AHM should be focused on examining if the reduced Q matrix derived from the attribute hierarchy is truly representing the cognitive processes used by students to solve test items. Thus, it is inadequate to only use the item difficulty parameter to evaluate if a student's response vector fits the AHM model.

IRT-based Person-fit Statistics

IRT-based person-fit statistics can be used to evaluate the misfit of an observed response vector to the IRT probabilities calculated with an IRT model using the student's ability theta and item parameters. Broadly speaking, the IRT-based person-fit statistics consist of residual-based statistics, likelihood-based statistics, and caution-index-based statistics (Meijer & Sijtsma, 2001).

Residual-based statistics include Wright and Stone's (1979) U statistic, Wright and Masters's (1982) W statistic, and Smith's (1985) UB and UW statistics. These statistics are used to compare a student's response relative to the IRT probability of a correct response determined by the student's ability theta and item parameters. The difference between the observed response and the IRT probability represents the residual which could not be explained by the IRT model. An observed response vector is considered as misfitting when the mean squared residuals across items are relatively large.

Likelihood-based statistics are derived from the log-likelihood function to assess person fit (e.g., Drasgow, Levine, & McLaughlin, 1991; Drasgow, Levine, & Williams, 1985; Levine & Drasgow, 1982, 1983; Levine & Rubin, 1979; Molenaar & Hoijtink, 1990). The log-likelihood function, first used by Levine and Rubin (1979), is given by:

$$l_{0i} = \sum_{j=1}^{J} \{x_{ij} \ln P_j(\theta_i) + (1 - x_{ij}) \ln[1 - P_j(\theta_i)]\},$$

where

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

$P_j(\theta_i)$ = the IRT probability of a correct response to item j by student i.

A low value of l_{0i} suggests that the probability of obtaining this observed response vector is small when the hypothesized IRT model is true. In turn, this observed response vector will be determined as a misfit of the IRT model. In order for l_{0i} to be used to classify an observed response vector as misfitting, the distribution of l_{0i} under the null hypothesis of the fit between the response vector and the IRT model is needed. However, the null distribution of l_{0i} is unknown. In addition, as pointed out by Meijer and Sijtsma (2001), l_{0i} is not standardized, indicating that the classification of an observed response vector as model-fitting or misfitting is influenced by θ_i. In order to overcome these problems, Drasgow et al. (1985) developed a standardized statistic of l_{0i}, which is provided by:

$$l_{zi} = \frac{l_0 - E(l_{0_i})}{[Var(l_{0_i})]^{1/2}},$$

where $E(l_{0_i})$ and $Var(l_{0_i})$ are the expectation and variance of l_{0_i}, respectively:

$$E(l_{0_i}) = \sum_{j=1}^{J} \{P_j(\theta_i) \ln P_j(\theta_i) + [(1 - P_j(\theta_i)) \ln[1 - P_j(\theta_i)]\}$$

and

$$Var(l_{0_i}) = \sum_{j=1}^{J} \{P_j(\theta_i)[1 - P_j(\theta_i)][\ln \frac{P_j(\theta_i)}{1 - P_j(\theta_i)}]^2.$$

Drasgow et al. (1985) argued that l_{zi} is less influenced by the value of θ_i and the presence of non-normality of distribution when true θ_i values are used.

Several caution-index-based statistics that are of similar form to Sato's (1975) caution index C_i have been developed by Tatsuoka and Linn (1983). C_i is defined as the complement of the ratio of two covariances: the covariance between the observed response vector of student i and the item proportion-correct score vector, and the covariance between the theoretical Guttman score vector of student i and the item proportion-correct score vector. The caution index is given by

$$C_i = 1 - \frac{Cov(\mathbf{X_i},\mathbf{n})}{Cov(\mathbf{X_i}^*,\mathbf{n})},$$

where

$\mathbf{X_i}$ = the observed response vector of student i,

$\mathbf{n}$ = the item number-correct score vector across students, and

$\mathbf{X_i}^*$ = the theoretical Guttman score vector of student i.

To adapt this caution index in the IRT framework, Tatsuoka and Linn (1983) proposed several statistics, including $ECI1_i$, $ECI2_i$, $ECI3_i$, $ECI4_i$, $ECI5_i$, and $ECI6_i$. $ECI1_i$ can be calculated by adapting C_i using student i's IRT probability vector in place of this student's theoretical Guttman score vector. $ECI1_i$ can be written as

$$ECI1_i = 1 - \frac{Cov(\mathbf{X_i},\mathbf{n})}{Cov[\mathbf{p}(\boldsymbol{\theta}_i),\mathbf{n}]},$$

where

$\mathbf{X_i}$ = the observed item response vector of student i,

$\mathbf{n}$ = the item number-correct score vector across students, and

$\mathbf{p}(\boldsymbol{\theta}_i)$ = the IRT item probability vector of student i.

Since the rest of the caution-index-based statistics are developed in the similar manner to the $ECI1_i$, they are not reviewed in this chapter. For a detailed description of these statistics, readers are referred to Tatsuoka and Linn (1983).

In general, the IRT-based statistics compare the observed item responses with the calculated IRT probabilities using the estimate of the student's overall ability. However, estimations of students' attribute mastery patterns are often of more interest for cognitive diagnoses. As a result, the person-fit statistics should focus on evaluating the cognitive and statistical soundness of the inferences made about students' attribute mastery patterns made by the AHM. By only concentrating on the single estimate on the student's overall ability, the person-fit statistics developed for the IRT models are not adequate for the AHM.

Initial Person-fit Statistics for the AHM

The first classification method for the AHM, Method A, proposed by Leighton et al. (2004), could be used to evaluate person fit. Broadly speaking, this method can be considered as a likelihood-based procedure in the sense that the likelihood function of slips is used to assess person fit. In Method A, the likelihood of slips is calculated by comparing an observed response vector to each of the expected response vectors. The higher the likelihood value, the more likely it is that the observed response vector originated from the expected response vector. The observed response vector is judged as originating from the expected response pattern when the maximum likelihood is achieved. However, when the maximum likelihood value is very low, it can be concluded that the observed response vector is unlikely to have originated from any of the expected response vectors. Thus, the observed response vector will be judged as not fitting the AHM.

However, to a great degree, Method A relies on the accurate estimation of ability theta for each of the expected response patterns, which is a critical element in the calculation of the likelihood value. As a result, when a misfit is found, one can not tell whether it is caused by the misfit of the observed response pattern to the attribute hierarchy, the IRT model, or both.

The Hierarchy Consistency Index

This review of the literature on existing person-fit statistics revealed that existing person-fit statistics cannot be used to adequately evaluate the misfit of the observed response vectors to expected response vectors in the AHM. Hence, the current study was designed to develop a person-fit statistic, called the hierarchy consistency index (HCI_i), to help assess the degree to which an observed student response pattern is consistent with the AHM, ultimately enhancing the validity of diagnostic feedback produced by the AHM.

The proposed person-fit statistic HCI_i depends on item complexity as determined by the attribute hierarchy and its associated reduced Q matrix. In the AHM, the reduced Q matrix, which is derived from the attribute hierarchy, is used to describe the knowledge and cognitive skills required in order for students to solve each item correctly. Therefore, by comparing an observed student response vector to the expectations associated with the reduced Q matrix, the HCI_i can be used to assess whether the student uses different cognitive skills (or in a different combination) when solving test items from those indicated by the reduced Q matrix associated with the attribute hierarchy. To calculate the HCI_i, the reduced Q matrix needs to be specified. When the attribute hierarchy is used as a cognitive model for test development, the reduced Q matrix can be derived from the attribute hierarchy to guide the construction of test items. When test items are not developed based on the attribute hierarchy, the reduced Q matrix will have to be obtained by reviewing test items and identifying the attributes required by each item. However, as discussed earlier, the extraction of the attribute hierarchy from actual test items might be problematic if items needed to reflect the relationships in the hierarchy are missing from these test items. This can lead to the problem of the nonidentifiability of certain attributes.

In the AHM, a student is considered to have mastered all of the required attributes for an item when the student answers the item correctly. Thus, the student is expected to correctly answer all those items that require the subset of attributes measured by the correct-answered item. Therefore, the HCI_i for student i is given by

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

where

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

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

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

X_{i_g} is student i's score (1 or 0) to item g, and

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

The term $\sum_{j \in S_{correct_i}} \sum_{g \in S_j} X_{i_j}(1 - X_{i_g})$ in the numerator of the HCI_i represents the number of misfits between student i's item response vector and the expected response vectors associated with the reduced Q matrix. When student i correctly answers item j, $X_{i_j} = 1$, then the student is expected to also correctly answer item g that belongs to S_j, namely, $X_{i_g} = 1$ $(g \in S_j)$. If the student fails to correctly answer item g, $X_{i_g} = 0$, then $X_{i_j}(1 - X_{i_g}) = 1$ and it is a misfit of the response vector i to the reduced Q matrix. Thus, $\sum_{j \in S_{correct_i}} \sum_{g \in S_j} X_{i_j}(1 - X_{i_g})$ is equal to the total number of misfits. The denominator of the HCI_i, N_{c_i}, contains the total number of comparisons for items that are correctly answered by student i. When the numerator of the HCI_i is set to equal the total number of misfits multiplied by 2, the HCI_i has the property of ranging from -1 to +1, which makes it easy to interpret. When a student's response vector fits the attribute hierarchy perfectly (i.e., the student's response vector matches one of the expected response patterns without any slips), the numerator of the HCI_i will be 0 and the HCI_i will have a value of 1. Conversely, when the response vector completely misfits the reduced Q matrix (i.e., the student correctly answers one item but fails to answer any item that requires the subset of attributes measured by the correct-answered item), the numerator of the HCI_i will be equal to $(2 \times N_{c_i})$ and the HCI_i will be -1. If the HCI_i value of a student response vector is close to -1, one can conclude that the student likely uses different knowledge and skills to solve test items as specified in the attribute hierarchy and its associated reduced Q matrix. As a result, the attribute hierarchy fails to provide a valid representation of the student cognition and consequently cannot be used to make inferences about the student performances. In addition, depending on the shape of the distribution of the HCI_i, the mean or the median of the HCI_i can be used as indicators of the overall model fit. A high mean or median would suggest an overall fit of students' item responses vectors relative to the attribute hierarchy.

To illustrate the calculation of the HCI_i, consider the attribute hierarchy presented earlier in Figure 1 and reproduced here for convenience.

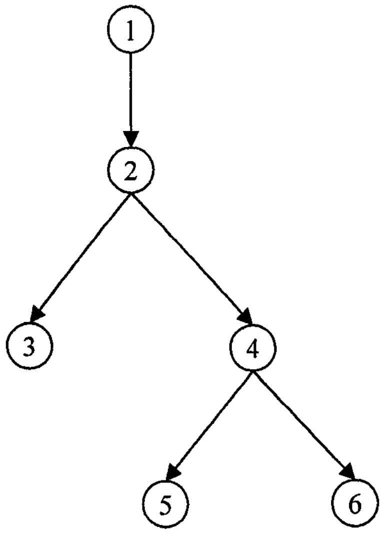

Figure 1. A Six-attribute Hierarchy

The reduced Q matrix associated with this attribute hierarchy is as follows:

$$Q_{R_{6,11}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Consider the observed response vector (11000100000) where items 1, 2, and 6 are correctly answered, namely $S_{correct_i}$ = $\{1, 2, 6\}$. According to the reduced Q matrix, item 6 measures attributes 1, 2, 4, and 5. Since student i correctly answers item 6, he or she is considered to have mastered the attributes required by this item. Therefore, student i is expected to also answer items 1, 2, and 4 correctly, each of which measures a subset of attributes required by item 6. That is, S_6 = $\{1, 2, 4\}$. Therefore, for item 6, there are three comparisons: item 6 vs. 1, 2, and 4. Since student i failed to answer item 4 correctly, $X_{i_6}(1-X_{i_4})=1$, a misfit between the student's response vector and the expected response vector derived from the reduced Q matrix is found. In the same manner, for items 1 and 2 that are also correctly answered by student i, $S_2=\{1\}$ and $S_1=\{\ \}$. Since S_2 contains item 1, which is correctly answered by student i, no misfit is found for item 2. S_1 is an empty set containing no elements, no comparison is made for item 1. Overall, the total number of misfits is 1, and the total number of comparisons is equal to 3 + 1 + 0 = 4. Hence, $HCI_i = 1 - \frac{2\times 1}{4} = 0.5$.

Table 2 displays some sample response vectors and their associated HCI_i values for the six-attribute hierarchy in Figure 1. The first row of Table 1 shows a student who correctly answers item 3 but fails to answer the rest of items correctly. In total, for this response vector, two pairs of item responses are compared where two misfits are identified. As a result, the corresponding HCI_i value is $1-\frac{2\times 2}{2}=-1$. It should be noted that different observed response vectors might have identical HCI_i values. For instance, both response vectors (11100110000) and (11000100000) produce an HCI_i value of 0.50. For response vector (11100110000), the total number of comparisons is 12, and three out of 12 comparisons are not consistent with the expectations associated with the reduced Q matrix. For response vector (11000100000), a total of four pairs of item responses are compared where one pair is not consistent with the expectations of the attribute hierarchy. This suggests that distinct response vectors might show the same degree of consistency with the attribute hierarchy.

Table 2

Sample Response Vectors and Their Associated HCI_i Values

Response Vectors	# of Correctly-answered Items	Total # of Comparisons	# of Misfits	HCI_i
00100000000	1	2	2	-1.00
00000100100	2	9	9	-1.00
00100010000	2	8	7	-0.75
00101010101	5	30	21	-0.40
00001111011	6	33	22	-0.33
01001111011	7	34	17	0.00
11100110000	5	12	3	0.50
11000100000	3	4	1	0.50
11111111011	10	30	1	0.95
11010101010	6	14	0	1.00
11111111111	11	42	0	1.00

It also should be noted that the distribution of the HCI_i under the null hypothesis that the student response vector fits the reduced Q matrix is unclear, and must be specified so that the critical value can be identified for significance testing. A simulation procedure was employed in the present study to approximate the null distribution of the HCI_i and in turn to set the critical value of the HCI_i to test the null hypothesis. Given that a higher HCI_i value suggests a better fit of a student response vector relative to the attribute hierarchy, if the observed HCI_i is smaller than the critical value, then the null hypothesis of the fit between a student response vector and the reduced Q matrix will be rejected at the significance level associated with the critical value.

Simulation Procedure for Identifying Critical values of the HCI_i

In hypothesis testing, the result of a statistic is evaluated by assessing the null hypothesis. To assess the null hypothesis, researchers first assume it is true and then test the reasonableness of this assumption by calculating the probability of obtaining the result due to chance. If the estimated probability is less than alpha, the null hypothesis will be rejected. The value of alpha is assigned by researchers based on theoretical and empirical considerations. The null hypothesis can also be evaluated by using the critical value to determine the critical region for rejection of the null hypothesis under the distribution curve of the tested statistic. The critical region is defined as the area under the null distribution curve that contains all the values of the statistic that allow rejection of the null hypothesis. By comparing the observed value of the statistic against the critical value, researchers will either reject or fail to reject the null hypothesis.

In the current study, the interest was to test the misfit of a student response vector to the attribute hierarchy by using the proposed statistic HCI_i. Hence, the HCI_i for a student response vector was evaluated by assessing the null hypothesis that the student response vector fits the attribute hierarchy well. Ideally, the probability of obtaining the observed HCI_i based on the distribution of the HCI_i when the null hypothesis is true is calculated. If the calculated probability turns out to be less than the alpha level, one can conclude the student response vector does not fit the attribute hierarchy well.

Unfortunately, the probability distribution is unknown for the proposed HCI_i under the null hypothesis that the student response vector fits the attribute hierarchy. To circumvent this problem, a simulation procedure was used to approximate the null distribution of the HCI_i using simulated data with known characteristics. This simulation procedure made possible the identification of the HCI_i value at the location in which the cumulative distribution function has a value that is equivalent to the critical value. Researchers can determine whether the student response vector fits the AHM by comparing the HCI_i of this response vector to the obtained critical value.

In order to produce this outcome, a set of student item response vectors was first simulated from the attribute hierarchy and the reduced Q matrix of the AHM. Since the purpose of this simulation was to approximate the distribution of the HCI_i under the null hypothesis, the simulated data set had a large sample size to decrease the errors due to random sampling. Each student response vector was generated by randomly adding slips from 1 to 0 and from 0 to 1 to one of the expected response patterns derived from the attribute hierarchy and the reduced Q matrix. The percentage of slips for each item was determined according to the prior knowledge about the nature of the item. In the next chapter, the procedures for randomly adding slips will be discussed in detail.

After simulating a set of student response vectors, the HCI_i value for each generated response vector was calculated and placed in ascending order. By doing this, the approximate distribution of the HCI_i under the null hypothesis was obtained. The HCI_i has the property of ranging from -1 to +1. A larger HCI_i value for a student response vector suggests a better fit of this response vector relative to the attribute hierarchy. Therefore, the critical region of the HCI_i is on the left side of its distribution. Using the alpha level of 0.05, the HCI_i value below which the 5% most extreme values fell was chosen as the critical value. In order for the null hypothesis to be rejected, the observed HCI_i must be smaller than the critical value. By rejecting the null hypothesis, one can conclude that the student likely uses different cognitive skills (or in a different combination) from those skills indicated by the attribute hierarchy and its associated reduced Q matrix.

In the next chapter, the simulated data sets were used to identify the critical values of the HCI_i for distinguishing good, moderate and poor fitting response vector under different simulation conditions. In addition, simulated data sets were also used to evaluate the effectiveness of the HCI_i in assessing the misfit of students response vectors relative to the attribute hierarchy, where the hypothesis is that higher HCI_i values should be obtained for the data sets with lower percentage of slips.

Chapter 4: Simulation Studies

The simulation studies were conducted for two purposes. The first purpose was to use simulated data of known characteristics to evaluate the effectiveness of the HCI_i in assessing the misfit of student response vectors to the attribute hierarchy and its associated reduced Q matrix in the AHM. The second purpose of simulations was to identify critical values of the HCI_i for distinguishing good, moderate, and poor fitting student response vectors under different simulation conditions by using the simulation approach for setting critical values of the HCI_i described in Chapter 3.

Method

Research Design

Student response data were simulated under a variety of conditions expected to affect the distribution and the effectiveness of the HCI_i. Four factors were manipulated: sample size, number of attributes, hierarchy structure, and percentage of slips. The levels of each factor were selected to reflect those that might be found in a real testing situation. First, sample size was set at 500, 1,000, and 1,500 to reflect small, moderate, and large sample sizes. Second, number of attributes was manipulated to range from five to seven with an increment of one attribute to examine whether this factor had an impact on the distribution and the effectiveness of the HCI_i. Third, the three types of hierarchy structure discussed by Leighton et al. (2004) – divergent, convergent, and linear structures - were considered. Different types of hierarchy structures describe distinct ordering of cognitive competencies required to solve problems successfully in a specific domain. Divergent structure represents hierarchies with divergent branches. This type of hierarchy is commonly present for a test domain with multiple divergent cognitive competencies. A hierarchy with convergent or linear structure ends at a single point, representing a test domain that involves a single end state of mastery. These three types of hierarchy structure were crossed with the three levels of the number of attributes producing a total of nine attribute hierarchies. These hierarchies are shown in Figures 3, 4, and 5, respectively. Fourth, the percentage of slips was set at 5%, 10%, and 20%. These levels of slips were selected to reflect a relatively good, moderate, and poor model-data fit, respectively. Thus, three levels of sample size, three levels of number of attributes, three types of hierarchy structure, and three levels of percentage of slips were considered in the current study so as to produce a total of 3x3x3x3=81 conditions. Each condition was replicated 100 times (Hawell, Stone, Hsu, & Kirisci, 1996) to obtain stable estimates of the HCI_i values and the critical values.

Data Generation

For each of the nine attribute hierarchies, the matrices of the AHM, including the adjacency matrix, the reachability matrix, the incidence matrix, the reduced Q matrix, and the expected response matrix, were derived. The obtained expected response matrix was used as a basis for the generation of student response data. A sample of 500, 1000, and 1500 expected item response vectors were separately generated based on each of the nine expected response matrices with the constraint that the total scores associated with the expected response patterns be normally distributed. Given that each generated sample

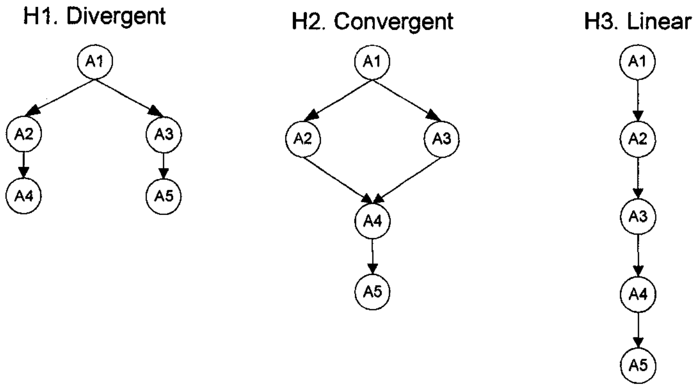

Figure 3. Three Five-Attribute Hierarchies Used for Simulation.

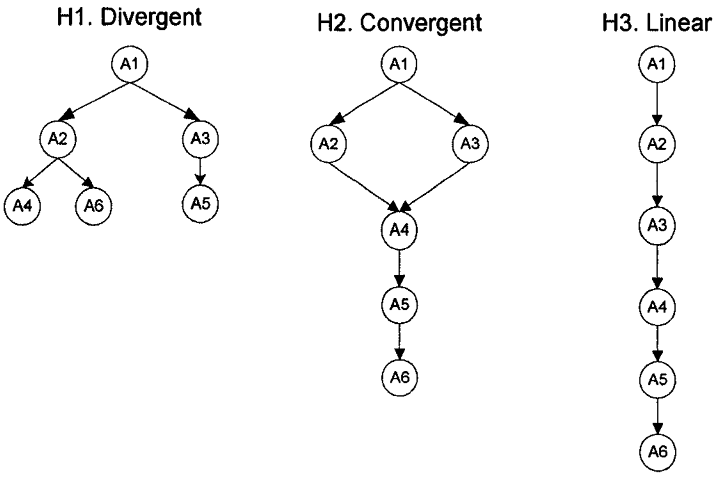

Figure 4. Three Six-Attribute Hierarchies Used for Simulations.

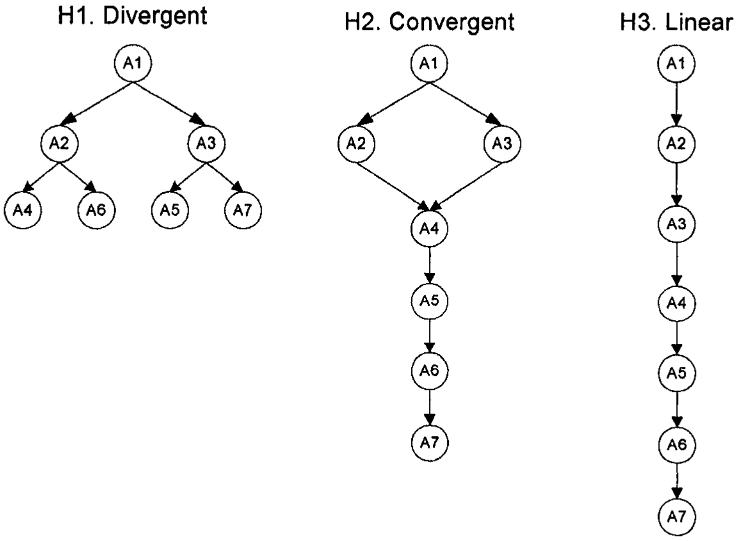

Figure 5. Three Seven-Attribute Hierarchies Used for Simulations.

only consisted of expected response patterns which were free from slips (from 1 to 0 and from 0 to 1), slips were randomly added to simulate real test-taking behaviors. In this simulation study, an assumption that all the items share an equal percentage of slips was made. This assumption may not be reasonable in the sense that the number of slips that students make in each item might vary with item characteristics, such as item difficulty and discriminating power, and student characteristics, such as ability level and gender. However, no studies were found in the literature that systematically investigated whether and how item and student characteristics affect the likelihood of slips made by students in answering test items. Therefore, in this study, uniform probabilities of 5%, 10%, and 20% were separately employed to create slips from 1 to 0 and from 0 to 1 for each generated expected response sample.

For example, to generate data based on the 5-attribute divergent hierarchy (H1 in Figure 3), first, the reduced Q matrix and the expected response matrix were derived from the hierarchy. The reduced Q matrix is shown as follows:

$$Q_{R_{5,11}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

The reduced Q matrix is of order 5 by 9 (i.e., attributes by items), suggesting that a minimum of 9 items are required in order to make inferences about students' mastery of the five attributes specified in the hierarchy. The first column of the reduced Q matrix is interpreted as showing that mastery of attribute 1 is required in order for students to answer item 1 correctly. The last column of the reduced Q matrix shows that item 11 requires the mastery of attributes 1 to 5 in order to reach the correct answer. The remaining columns can be interpreted in the same manner.

Using the reduced Q matrix, the expected response matrix was derived:

$$E_{10,9} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The expected response matrix is of order 10 by 9 (i.e., response vectors by items), suggesting that if the attribute hierarchy is truly representing student cognition, ten student response vectors, without any slips from 1 to 0 and from 0 to 1, are expected. In order to generate data based on expected response patterns, the frequency associated with each expected response pattern needs to be calculated for each sample size considered in the simulations (500, 1000, or 1500). These frequencies were calculated by making the assumption that the total scores associated with the expected response patterns are normally distributed. The obtained frequencies are shown in Table 3. According to the second row of Table 3, the frequencies associated with the response pattern (000000000) are 45, 90, and 134 when the sample size is set at 500, 1000, and 1500, respectively. It should be noted that the total scores associated with the expected response patterns are normally distributed, although the frequency distribution of expected response patterns shown in Table 3 appear to be positively skewed. This is because different expected response vectors may lead to a same total score. For example, expected response patterns (111111000) and (111100110) have the same total score of six. Based on the normality assumption, the frequency associated with the total score of six was calculated as 72 when sample size was 500. Correspondingly, two expected response patterns – (111111000) and (111100110) – were assigned to have an equal frequency of 36, respectively. By using these frequencies, a data matrix of expected response patterns was obtained with each sample size. For instance, the data matrix of 500 expected response patterns contained 45 row vectors of (000000000), 68 row vectors of (100000000), and so forth. This data matrix ended with 15 row vectors of (111111111).

Table 3

Frequencies for Each Expected Response Pattern Associated with Different Sample Sizes for the 5-Attribute Divergent Hierarchy

	Sample Size		
Expected response pattern	500	1000	1500
(0 0 0 0 0 0 0 0 0)	45	90	134
(1 0 0 0 0 0 0 0 0)	68	136	204
(1 1 0 0 0 0 0 0 0)	45	90	135
(1 0 1 0 0 0 0 0 0)	45	90	135
(1 1 1 1 0 0 0 0 0)	106	211	317
(1 1 0 0 1 0 0 0 0)	52	104	156
(1 1 1 1 1 1 0 0 0)	36	73	109
(1 0 1 0 0 0 1 0 0)	52	104	156
(1 1 1 1 0 0 1 1 0)	36	73	109
(1 1 1 1 1 1 1 1 1)	15	30	45

In order to simulate the real testing behavior, slips were added to the data matrix of expected response patterns. An assumption of equal percentage of slips across items was made in this study. The number of slips was determined by sample size and percentage of slips. For each item, the number of slips was equal to the number of responses (500, 1000, or 1500) multiplied by the percentage of slips (5%, 10%, or 20%). For example, in the 500 sample size condition with 5% slips, $500 \times 5\% = 25$ responses to each item were randomly selected from the 500 expected response vectors. And if the selected response was 1, indicating that the student was expected to answer the item correctly, a slip from 1 to 0 was created by altering the response from 1 to 0. On the other hand, if the selected response was 0, which indicated that the student was expected to answer the item incorrectly, a slip from 0 to 1 was created by altering the response from 0 to 1. Slips for each item were created separately in this way. In total, for the 500 sample size condition with 5% slips, $(25 \times n)$ slips were added to the data matrix of expected response patterns for each attribute hierarchy, where n is the number of items as indicated by the reduced Q matrix associated with the hierarchy. For example, nine items were used according to the reduced Q matrix derived from the 5-attribute divergent hierarchy (H1 in Figure 2). A total of $25 \times 9 = 225$ slips were created to simulate student responses to the nine items from the 5-attribute divergent hierarchy in the 500 sample size condition with 5% slips. This process was replicated 100 times, each with a different random seed. By doing this, 100 data sets were generated for each condition. Table 4 presents the number of slips added to the student responses for each item under different simulation conditions. The total number of slips added to each data set under different conditions is shown in Table 5.

Table 4

Number of Slips Added into Each Item under Different Simulation Conditions

	Percentage of Slips		
Sample Size	5%	10%	20%
500	25	50	100
1000	50	100	200
1500	75	150	300

Table 5

The Total Number of Slips Added under Different Simulation Conditions

Number of Attributes	Hierarchy Structure	Number of Items	Sample Size	Percentage of Slips		
				5%	10%	20%
5-attribute	Divergent	9	500	225	450	900
	Convergent	6	1000	300	600	1200
	Linear	5	1500	375	750	1500
6-attribute	Divergent	15	500	375	750	1500
	Convergent	7	1000	350	700	1400
	Linear	6	1500	450	900	1800
7-attribute	Divergent	25	500	625	1250	2500
	Convergent	8	1000	400	800	1600
	Linear	7	1500	525	1050	2100

By first creating a data matrix of expected response patterns and then randomly adding a certain percentage of slips, simulated data with known characteristics were generated. Using the same procedure for generating data for the 5-attribute divergent hierarchy, data for the remaining 8 hierarchies considered in the simulations (shown in Figures 3, 4, and 5) were generated. The expected response patterns and their corresponding frequencies for these 8 hierarchies are presented in Tables 6, 7, and 8.

Table 6

Expected Response Patterns and Their Associated Frequencies for the 6- and 7-Attribute Divergent Hierarchies

		Sample Size		
Divergent Hierarchy	Expected response pattern	500	1000	1500
6-attribute	(000000000000000)	31	62	93
	(100000000000000)	42	84	126
	(110000000000000)	27	54	81
	(101000000000000)	27	54	81
	(111100000000000)	72	144	215
	(110010000000000)	21	43	64
	(111111000000000)	25	49	74
	(110000100000000)	21	43	64
	(111100110000000)	25	49	74
	(110010101000000)	75	151	226
	(111111111100000)	36	73	109
	(101000000010000)	21	43	64
	(111100000011000)	25	49	74
	(111111000011100)	24	48	72
	(111100110011010)	24	48	72
	(111111111111111)	4	7	11
7-attribute	(0000000000000000000000000)	29	58	87
	(1000000000000000000000000)	36	72	107
	(1100000000000000000000000)	21	43	64
	(1010000000000000000000000)	21	43	64
	(1111000000000000000000000)	56	113	169
	(1100100000000000000000000)	12	25	37
	(1111110000000000000000000)	16	33	49
	(1100001000000000000000000)	12	25	37
	(1111001100000000000000000)	16	33	49
	(1100101010000000000000000)	31	62	93
	(1111111111000000000000000)	31	62	93
	(1010000000100000000000000)	12	25	37
	(1111000000110000000000000)	16	33	49
	(1111110000111000000000000)	17	33	50
	(1111001100110100000000000)	17	33	50
	(1111111111111110000000000)	7	15	22
	(1010000000000001000000000)	12	25	37
	(1111000000000001100000000)	16	33	49
	(1111110000000001110000000)	17	33	50
	(1111001100000001101000000)	17	33	50
	(1111111111000001111100000)	7	15	22
	(1010000000100001000010000)	31	62	93
	(1111000000110001100011000)	31	62	93
	(1111110000111001110011100)	7	15	22
	(1111001100110101101011010)	7	15	22
	(1111111111111111111111111)	1	1	2

Table 7

Expected Response Patterns and Their Associated Frequencies for the 5-, 6- and 7-Attribute Convergent Hierarchies

		Sample Size		
Convergent Hierarchy	Expected response pattern	500	1000	1500
5-attribute	(000000)	55	110	166
	(100000)	89	179	268
	(110000)	59	118	177
	(101000)	59	118	177
	(111100)	111	223	334
	(111110)	80	159	239
	(111111)	46	93	139
6-attribute	(0000000)	44	88	132
	(1000000)	69	139	208
	(1100000)	47	93	140
	(1010000)	47	93	140
	(1111000)	105	209	314
	(1111100)	88	175	263
	(1111110)	63	126	188
	(1111111)	38	77	115
7-attribute	(00000000)	36	72	108
	(10000000)	55	111	166
	(11000000)	37	75	112
	(10100000)	37	75	112
	(11110000)	93	187	280
	(11111000)	87	173	260
	(11111100)	71	141	212
	(11111110)	51	102	153
	(11111111)	32	65	97

Table 8

Expected Response Patterns and Their Associated Frequencies for the 5-, 6- and 7-Attribute Linear Hierarchies

		Sample Size		
Linear Hierarchy	Expected response pattern	500	1000	1500
5-attribute	(00000)	49	99	148
	(10000)	86	173	259
	(11000)	114	228	343
	(11100)	114	228	343
	(11110)	86	173	259
	(11111)	49	99	148
6-attribute	(000000)	40	79	119
	(100000)	67	134	201
	(110000)	92	184	276
	(111000)	102	205	307
	(111100)	92	184	276
	(111110)	67	134	201
	(111111)	40	79	119
7-attribute	(0000000)	33	66	99
	(1000000)	54	108	162
	(1100000)	75	150	225
	(1110000)	88	177	265
	(1111000)	88	177	265
	(1111100)	75	150	225
	(1111110)	54	108	162
	(1111111)	33	66	99

Evaluating the Effectiveness of the HCI_i

As discussed in Chapter 3, a higher HCI_i value suggests a better fit of the student response vector to the attribute hierarchy. Therefore, if the HCI_i works well in determining the degree to which a student response vector corresponds to the attribute hierarchy, the results should indicate that higher HCI_i values are obtained for response vectors with a lower percentage of slips. In other words, the highest HCI_i values are expected to be obtained for data sets with 5% slips, and the lowest HCI_i values for data sets with 20% slips. Of additional interest was to investigate whether the HCI_i is effective in examining the person fit of student response vectors across various forms of hierarchical structures.

For each generated data set, the HCI_i was applied to the simulated response vectors, and the median of the HCI_i values over the response vectors was calculated. The use of medians as the measure of central tendency was due to the markedly negatively skewed distribution of the HCI_i, which will be demonstrated in the results section. For data sets generated based on a same attribute hierarchy and at a same level of sample size (500, 1000, or 1500), the means of the median HCI_i values over the 100 data sets with 5%, 10%, and 20% slips were compared thereby providing a general criterion to evaluate the effectiveness of the HCI_i.

Identifying Critical Values for Interpreting the HCI_i

Simulated data sets were also used to identify critical values of the HCI_i for distinguishing good, moderate, and poor fitting response vectors relative to the attribute hierarchy. For each simulated data set, the HCI_i values were calculated and ordered according to decreasing misfit and the value below which the 5% most extreme misfitting values fell was taken as the critical value for the data set. As a result, for each condition, 100 critical values were calculated and the mean of the 100 critical values was used as the final critical value. In this study, the simulated data sets with 5% slips were used to set the critical values for distinguishing a good and a moderate person fit, and the simulated data sets with 10% slips were used to set the critical values for distinguishing a moderate and a poor person fit. For each attribute hierarchy, if the HCI_i value for a student response vector is greater than the critical value produced from the data sets with 5% slips, one can conclude that there is a good fit between the student response vector and the attribute hierarchy. If the HCI_i value for a student response vector is smaller than the critical value produced from the data sets with 5% slips but greater than the critical value produced from the data sets with 10% slips, one can conclude that there is a moderate fit between the student response vector and the attribute hierarchy. A smaller observed HCI_i value than the critical value produced from the data sets with 10% slips suggests a poor fit of the student response vector to the attribute hierarchy. As a result, for each attribute hierarchy, a guideline was produced for identifying good, moderate, and poor fitting response vectors to the attribute hierarchy.

Meanwhile, the median of the HCI_i values for each data set was used as the critical value to evaluate the overall model-data fit. For each condition, 100 medians were calculated and ordered, and the fifth smallest median was used as the critical value for the overall model fit. For an observed data set, if the median of the HCI_i values is greater than the critical value for the overall model fit produced from the data sets with 5% slips, one can conclude that there is a good overall model fit. If the median of the HCI_i values for an observed data set is smaller than the critical value for the overall model fit produced from data sets with 5% slips but greater than the critical value produced from data sets with 10% slips, one can conclude that there is a moderate overall model fit. A smaller median of the HCI_i values than the critical value for the overall model fit produced from data sets with 10% slips suggests a poor overall model fit. Therefore, for each attribute hierarchy, a guideline was produced for identifying the good, moderate, and poor overall model-data fit. The results for the different attribute hierarchies were compared to investigate whether critical values of the HCI_i were influenced by sample size, number of attributes, and hierarchy structure.

Results

The results from the simulation studies are presented in four parts. Typical example frequency distributions of the HCI_i are first presented to give the reader a general idea of what the distributions of the HCI_i look like. The results used for evaluating the effectiveness of the HCI_i through the comparison of the median HCI_i values across different conditions are then presented, followed by the results associated with the critical values for evaluating the person fit of a student response vector to the attribute hierarchy. Finally, the results for the overall model-data fit are described.

The Frequency Distributions of the HCI_i

Figure 6 shows the HCI_i frequency distribution associated with one of the 100 data sets that were generated based on the expected response vectors of the 6-attribute divergent hierarchy (H1 in Figure 3) where 5% slips were added and the sample size was 500. The HCI_i appeared to be non-normally distributed and substantially negatively skewed, with a mean of 0.64 and a standard deviation of 0.53. The median and the mode of this frequency distribution were both 1.00. Of the 500 simulated response vectors, approximately 55% of vectors had an HCI_i value of 1, suggesting a perfect fit of student response vectors relative to the attribute hierarchy. Although data were generated based on the attribute hierarchy, almost 3% of the simulated response vectors had HCI_i values of -1. Suppose that a student, who has not mastered any attributes as specified in the attribute hierarchy, randomly guessed a difficult item correctly. Because the student correctly answered the difficult item but failed to answer any of its prerequisite items correctly, the HCI_i value for this student response vector was -1. In this case, one single slip to a difficult item led to a maximum misfit of the student response vector to the attribute hierarchy ($HCI_i = -1$), suggesting that the HCI_i is sensitive to the location of slips a student makes when the student has not mastered any attributes.

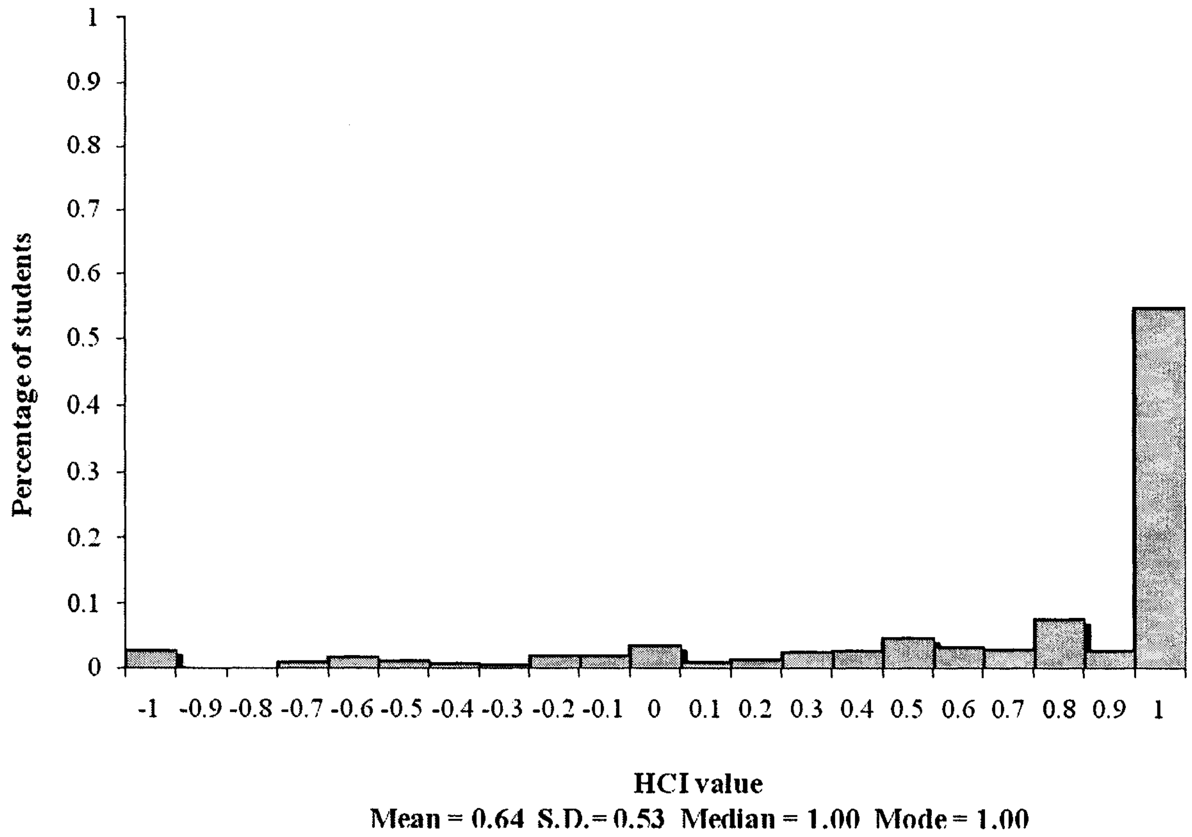

Figure 6. Tpyical Frequency Distribution of the HCI_i for the 6-attribute Divergent Hierarchy (5% Slips).

Figures 7 demonstrates the frequency distributions of a typical 5-attribute divergent hierarchy at the 10% slips condition. The HCI_i remained non-normally distributed but with a smaller negative skewness compared to the HCI_i distribution at the 5% slips condition. Of the 500 simulated response vectors, around 30% of vectors had an HCI_i value of 1 while 2% of vectors had an HCI_i value of -1. The frequency distributions of the HCI_i for the rest of the hierarchies considered in the simulations showed similar patterns as those for the 5-attribute divergent hierarchy in both shape and trend, and therefore are not presented here.

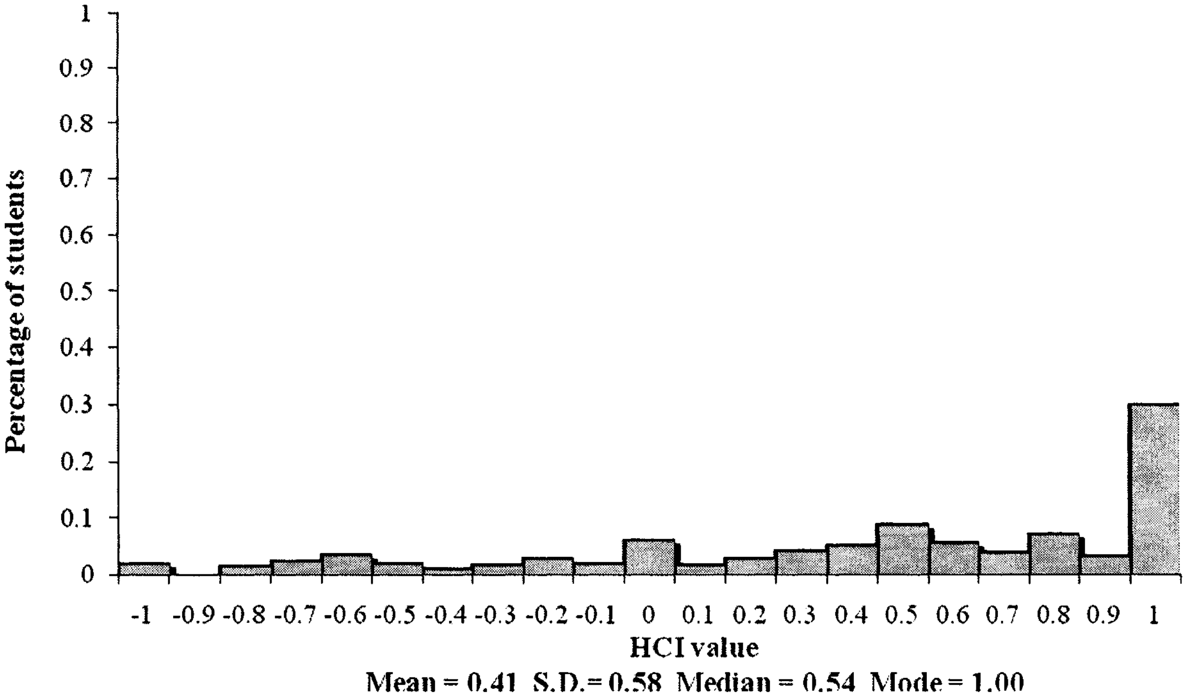

Figure 7. Tpyical Frequency Distribution of the HCI_i for the 6-attribute Divergent Hierarchy (10% Slips).

The Medians of the HCI_i Values

Given the substantively skewed distribution of the HCI_i, the median was chosen as the measure of central tendency for the purpose of evaluating the effectiveness of the HCI_i. The means and standard deviations of the median HCI_i values across the 100 replications under each condition are presented in Table 9. The standard deviations are presented in parenthesis.

Table 9

The Means and Standard Deviations of the Median HCI_i Values across the 100 Simulated Data Sets under Different Simulation Conditions

Hierarchy Structure	Number of attributes	Sample Size	Percentage of Errors		
			5%	10%	20%
Divergent	5 attributes	500	1.00 (0.00)	0.98 (0.05)	0.34 (0.03)
		1000	1.00 (0.00)	0.99 (0.03)	0.34 (0.02)
		1500	1.00 (0.00)	0.99 (0.02)	0.33 (0.01)
	6 attributes	500	1.00 (0.00)	0.56 (0.03)	0.20 (0.02)
		1000	1.00 (0.00)	0.56 (0.02)	0.20 (0.02)
		1500	1.00 (0.00)	0.56 (0.01)	0.19 (0.02)
	7 attributes	500	0.70 (0.03)	0.30 (0.02)	0.05 (0.02)
		1000	0.71 (0.02)	0.30 (0.02)	0.04 (0.01)
		1500	0.71 (0.02)	0.30 (0.02)	0.05 (0.01)
Convergent	5 attributes	500	1.00 (0.00)	1.00 (0.00)	0.55 (0.06)
		1000	1.00 (0.00)	1.00 (0.00)	0.55 (0.04)
		1500	1.00 (0.00)	1.00 (0.00)	0.54 (0.04)
	6 attributes	500	1.00 (0.00)	1.00 (0.00)	0.51 (0.01)
		1000	1.00 (0.00)	1.00 (0.00)	0.50 (0.01)
		1500	1.00 (0.00)	1.00 (0.00)	0.50 (0.01)
	7 attributes	500	1.00 (0.00)	1.00 (0.01)	0.50 (0.00)
		1000	1.00 (0.00)	1.00 (0.00)	0.50 (0.00)
		1500	1.00 (0.00)	1.00 (0.00)	0.50 (0.00)
Linear	5 attributes	500	1.00 (0.00)	1.00 (0.00)	0.91 (0.14)
		1000	1.00 (0.00)	1.00 (0.00)	0.95 (0.11)
		1500	1.00 (0.00)	1.00 (0.00)	0.96 (0.10)
	6 attributes	500	1.00 (0.00)	1.00 (0.00)	0.53 (0.06)
		1000	1.00 (0.00)	1.00 (0.00)	0.51 (0.03)
		1500	1.00 (0.00)	1.00 (0.00)	0.51 (0.03)
	7 attributes	500	1.00 (0.00)	1.00 (0.00)	0.49 (0.02)
		1000	1.00 (0.00)	1.00 (0.00)	0.50 (0.01)
		1500	1.00 (0.00)	1.00 (0.00)	0.50 (0.01)

When 5% slips were added to the expected response vectors, the mean of the medians of the HCI_i values were equal to 1.00 except for the 7-attribute divergent hierarchy. This suggested that at least 50% of simulated response vectors were consistent with the expected response vectors without any slips. For the 7-attribute divergent hierarchy, the mean medians of the HCI_i values were 0.70, 0.71, and 0.71 under the 500, 1000, and 1500 conditions, respectively. For the 10% slips condition, the mean medians of the HCI_i values remained at 1.00 for convergent and linear hierarchies. However, for the divergent hierarchies, the mean medians of the HCI_i values ranged from 0.30 to 0.99. These values appeared to decrease as the number of attributes increased after controlling for the factor of sample size. For example, when the sample size was fixed at 500, the mean medians of the HCI_i values were 0.98, 0.56, and 0.30 for the 5-, 6-, and 7-attribute divergent hierarchies, respectively. Sample size did not appear to affect the median values of the HCI_i within a hierarchy structure of a given number of attributes, with the maximum difference of only 0.01.

For the 20% slips condition, hierarchies of divergent structure produced the smallest mean medians of the HCI_i values, ranging from 0.04 to 0.34, while hierarchies of linear structure yielded the largest values, ranging from 0.49 to 0.96. The mean medians of the HCI_i values ranged from 0.50 to 0.55 for convergent hierarchies. As the number of attributes increased, the median of the HCI_i values decreased considerably for divergent hierarchies after controlling for sample size, suggesting that the number of attributes displayed a negative effect on the median HCI_i values for hierarchies with divergent structure. For example, the mean median HCI_i values were 0.34, 0.20, and 0.05, respectively, for data sets generated from the 5-, 6-, and 7-attribute divergent hierarchies. For convergent hierarchies, however, the mean medians of the HCI_i values only slightly decreased as the number of attributes increased after controlling for sample size. For linear structure, the medians of the HCI_i values dropped dramatically when the number of attributes increased from five to six, while the values slightly decreased when the number of attributes increased from six to seven. In addition, sample size did not show a significant impact on the median HCI_i values within a hierarchy structure of a given number of attributes. The maximum difference occurred when comparing the median values of the 500 sample size condition with respective values under the 1500 condition for the 5-attribute linear hierarchy.

To summarize, the magnitude of the median HCI_i values appeared to be stable for data sets of different sample sizes, with a maximum difference of 0.03, after controlling for hierarchy structure, number of attributes, and percentage of slips. The maximum difference was produced when data were simulated by adding 20% slips to the expected response vectors derived from the 5-attribute linear hierarchy. Given the negligible difference in the mean median HCI_i values across different sample sizes, one can conclude that sample size did not show an impact on the median HCI_i values.

Results also showed that the mean median HCI_i values under the 20% slips condition were consistently lower than the respective values under the 10% slips condition, which , in turn, were lower than, if not equal to, the values under the 5% slips condition after controlling for hierarchy structure, number of attributes, and sample size. For example, for the data sets simulated based on the 5-attribute divergent hierarchy and with the sample size of 500, the mean median HCI_i Values were 1.00, 0.98, and 0.34, respectively, when 5%, 10%, and 20% slips were randomly added into expected response vectors. Thus, the percentage of slips showed a negative effect on the median HCI_i

values after controlling for other factors, suggesting that higher HCI_i values tended to be produced by data sets with lower percentage of slips. Given that a higher HCI_i value suggests a better fit of a student response vector to the attribute hierarchy, to evaluate the effectiveness of the HCI_i the hypothesis is that the HCI_i values should decrease as the percentage of slips increases. Since the simulation results confirmed this hypothesis, one can conclude that the HCI_i works well in examining the degree to which a student response vector is consistent with the attribute hierarchy.

Critical Values for Testing the Person-Fit

The critical values, selected to statistically examine the person fit of a student response vector to the attribute hierarchy, are presented in Table 9. Two statistically-set critical values were selected under each simulation condition, one used for discriminating a good and a moderate person-fit (CV1), the other for discriminating a moderate and a poor person-fit (CV2). For example, according to the first row of Table 10, two critical values, -0.58 and -0.94, were identified for determining the person fit of a student response vector to the 5-attribute divergent hierarchy (H1 in Figure 3) when sample size was set at 500. If an observed HCI_i value for a student response vector is greater than -0.58, one can conclude there is a good fit between the student response vector and the attribute hierarchy. If the observed HCI_i value is smaller than -0.58 but greater than -0.94, one can conclude that the student response vector fits the attribute hierarchy moderately. If the HCI_i value for a student response vector is smaller than -0.94, the corresponding student response vector will be judged as not fitting the attribute hierarchy.

Table 10

Critical Values for Testing the Person-Fit

Hierarchy Structure	Number of attributes	Sample Size	Critical Values	
			CV1*	CV2*
Divergent	5 attributes	500	-0.58	-0.94
		1000	-0.60	-0.96
		1500	-0.60	-0.96
	6 attributes	500	-0.65	-0.78
		1000	-0.64	-0.78
		1500	-0.63	-0.78
	7 attributes	500	-0.79	-0.80
		1000	-0.79	-0.80
		1500	-0.79	-0.80
Convergent	5 attributes	500	-0.46	-0.96
		1000	-0.47	-0.97
		1500	-0.48	-0.99
	6 attributes	500	-0.44	-0.82
		1000	-0.44	-0.79
		1500	-0.45	-0.83
	7 attributes	500	-0.41	-0.72
		1000	-0.40	-0.71
		1500	-0.40	-0.71
Linear	5 attributes	500	-0.28	-0.83
		1000	-0.33	-0.82
		1500	-0.33	-0.85
	6 attributes	500	-0.28	-0.69
		1000	-0.32	-0.66
		1500	-0.31	-0.66
	7 attributes	500	-0.29	-0.65
		1000	-0.31	-0.64
		1500	-0.30	-0.64

Note: CV1 is the critical value identified for distinguishing a good and moderate fit
CV2 is the critical value identified for distinguishing a moderate and poor fit
The values were the mean of critical values across 100 replications under each simulation condition

All the identified critical values for person fit were negative, ranging from -0.28 to -0.79 for CV1 (critical values for distinguishing a good and a moderate person fit) and from -0.64 to -0.99 for CV2 (critical values for distinguishing a moderate and a poor person fit). These values were close to the lower bound of the HCI_i (-1), which indicates a maximum misfit of the student response vector relative to the attribute hierarchy. Using these low critical values, a response vector can be easily identified as fitting the attribute hierarchy. As a result, the critical values yielded by the statistical approach appeared to provide overly liberal criteria for testing person fit. Therefore, statistical analyses for identifying critical values for person fit were not pursued further.

Critical Values for Testing the Overall Model Fit

In order to evaluate the overall model data fit, two critical HCI_i values for examining the overall model data fit were selected under each simulation condition, one for discriminating a good and a moderate overall fit, and the other for discriminating a moderate and a poor overall fit. These critical values are presented in Table 11. The first row of Table 11 shows that two critical values, 1.00 and 0.86, were identified for examining the overall model data fit for the 5-attribute divergent hierarchy when sample size was set at 500. If the median of the HCI_i values for an observed data set is 1.00, suggesting at least 50% of response vectors produced an HCI_i value of 1.00, one can conclude that students' response vectors are, in general, statistically consistent with the expected response vectors associated with the given attribute hierarchy. If the median of the HCI_i values is smaller than 1.00 but greater than 0.86, one can conclude that there is a moderate overall model fit of students' response vectors relative to the attribute hierarchy. If the median of the HCI_i values is smaller than 0.86, a poor overall model fit is found between student response vectors and the attribute hierarchy.

Table 11

Critical Values for Testing the Overall Model Fit

Hierarchy Structure	Number of attributes	Sample Size	Critical Values	
			CV1	CV2
Divergent	5 attributes	500	1.00	0.86
		1000	1.00	0.88
		1500	1.00	0.96
	6 attributes	500	1.00	0.50
		1000	1.00	0.52
		1500	1.00	0.54
	7 attributes	500	0.65	0.26
		1000	0.65	0.27
		1500	0.67	0.28
Convergent	5 attributes	500	1.00	1.00
		1000	1.00	1.00
		1500	1.00	1.00
	6 attributes	500	1.00	1.00
		1000	1.00	1.00
		1500	1.00	1.00
	7 attributes	500	1.00	1.00
		1000	1.00	1.00
		1500	1.00	1.00
Linear	5 attributes	500	1.00	1.00
		1000	1.00	1.00
		1500	1.00	1.00
	6 attributes	500	1.00	1.00
		1000	1.00	1.00
		1500	1.00	1.00
	7 attributes	500	1.00	1.00
		1000	1.00	1.00
		1500	1.00	1.00

Note: CV1 is the critical value identified for distinguishing a good and a moderate fit
CV2 is the critical value identified for distinguishing a moderate and a poor fit
The critical values were the mean of median values across 100 replications under each simulation condition

Results from Table 11 indicate that the critical values for testing the overall model fit using the HCI_i appeared to vary with the type of hierarchy structures after controlling for sample size and number of attributes. For divergent structure, when sample size was set at 500, the critical values identified for discriminating between a good and a moderate overall model fit (CV1) were 1.00, 1.00, and 0.65 for the 5-, 6-, and 7-attribute hierarchies, respectively. Therefore, in order for the 5- and 6-attribute divergent hierarchies to be judged as having a good overall model data fit, at least 50% of student response vectors should have an HCI_i value of 1.00. However, for the 7-attribute divergent hierarchy, only a median HCI_i value greater than 0.65 is required to be classified as having a good overall model fit. Therefore, the number of attributes showed a negative impact on CV1s. In addition, sample size did not display a significant effect on the values of CV1s, with a maximum difference of only 0.02.

Furthermore, for divergent hierarchies, critical values identified for discriminating a moderate and a poor overall model fit (CV2) were found to be influenced by number of attributes and sample size. As the number of attribute increased, CV2s decreased considerably after controlling for sample size. For example, for the 5-attribute divergent hierarchy, when the sample size was 500, the median of the HCI_i values needed to be greater than 0.86 in order to be judged as having a moderate fit relative to the attribute hierarchy. However, for the 6- and 7-attribute divergent hierarchy, in order to reach a moderate fit, the median of the HCI_i values only needed to be greater than 0.50 and 0.26, respectively. These results showed that, for divergent hierarchies, it is relatively easier to obtain a moderate overall model fit for hierarchies with more attributes. In addition, sample size appeared to positively affect the critical HCI_i values for discriminating a moderate and poor fit. For example, for the 5-attribute divergent hierarchy, the values were 0.86, 0.88, and 0.96 for the sample sizes of 500, 1000, and 1500, respectively. The increments among the critical values were relatively larger for the 5-attribute hierarchy compared to those for the 6- and 7-attribute hierarchies.

On the other hand, for the convergent and linear hierarchies, critical values were consistently equal to 1.00 across different number of attributes and sample sizes. These results indicated that, in order for an observed data set to reach the statistical fit relative to a convergent or linear hierarchy, at least 50% of student response vectors must yield an HCI_i value of 1.00. These results showed that the magnitude of the critical HCI_i values for the overall model data fit varied for attribute hierarchies with different hierarchy structures.

Summary and Conclusions of Simulation Studies

The first purpose for conducting the simulation studies was to evaluate the effectiveness of the HCI_i index in examining the degree to which a student response vector is consistent with the attribute hierarchy. The second purpose of conducting the simulation studies was to identify the critical values of the HCI_i for testing both person fit and overall model fit under different simulation conditions.

The Effectiveness of the HCI_i

The HCI_i was shown to be effective in examining the degree to which a student response vector fits an attribute hierarchy under different conditions manipulated in the simulations. As discussed in chapter 3, a higher HCI_i value suggests a better fit of a student response vector relative to the attribute hierarchy. Therefore, the hypothesis for evaluating the effectiveness of the HCI_i is that higher HCI_i values should be able to increase as the percentage of slips decreases. Given that the simulation results showed that higher HCI_i values were consistently found for simulated response vectors with lower percentage of slips across different simulation conditions, one can conclude that the HCI_i works well in examining the degree to which a student response vector is consistent with the attribute hierarchy.

Additionally, the magnitude of the mean medians of the HCI_i values was found to vary across different simulation conditions. Although sample size was not found to affect the mean medians of the HCI_i values, the number of attributes and the hierarchy structure showed an impact on these values as shown in Table 10. As the number of attributes increased, the mean medians of the HCI_i values appeared to decrease. The hierarchy structure was also shown to affect the mean medians of the HCI_i values. Results showed that the highest mean medians of the HCI_i values were produced by data sets generated based on hierarchies of convergent and linear structures, while the lowest values were found for data sets generated from divergent hierarchies.

The impact of hierarchy structure and number of attribute can be explained by the logic of the HCI_i. The HCI_i is operationalized by assuming that a student who correctly answers item A should be able to correctly answer those prerequisite items that include a subset of the attributes measured by item A. Item response comparisons are made to examine whether a student answers one item correctly but fails to answer its associated prerequisite items. If so, misfits are found. As the number of attributes increases or a divergent hierarchy is used instead of a linear or convergent hierarchy, the number of items needed to measure the attributes increases, and items tend to share more complicated prerequisite relationships. As a result, when a student makes a slip on an item, more item response comparisons tend to be judged as misfitting for a hierarchy with a greater number of attributes, or a hierarchy of divergent structure. For example, for the 5-attribute linear hierarchy shown in Figure 3 (H3), the derived reduced Q matrix, of order (5, 5), is shown as follows:

$$Q_{R_{5,5}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

According to this reduced Q matrix, five items should be created to estimate students' attribute profiles for the 5-attribute linear hierarchy. On the other hand, the reduced Q matrix derived from the 5-attribute divergent hierarchy, of order (5, 9), is specified as follows:

$$Q_{R_{5,9}} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

This matrix suggests that nine items should be constructed to estimate students' mastery of attributes illustrated in the 5-attribute divergent hierarchy. Compared to the reduced Q matrix derived from the 5-attribute linear hierarchy ($Q_{R_{5,5}}$), four additional items (represented by columns 3, 5, 7, and 8 in $Q_{R_{5,9}}$) are required, which leads to more complicated prerequisite relationships among items. For instance, in the reduced Q matrix ($Q_{R_{5,5}}$), item 1 is the prerequisite of items 2, 3, 4, and 5 in the sense that students are not expected to solve these items correctly if they fail to answer item 1. However, according to the reduced Q matrix associated with the divergent hierarchy ($Q_{R_{5,9}}$), item 1 is the prerequisite of items 2 to 9. If a student who has mastered attribute 1 makes a random slip on item 1, four comparisons will be judged as misfitting for the linear hierarchy but eight comparisons for the divergent hierarchy. Therefore, for the divergent hierarchy, the HCI_i is more sensitive to the slips that students make in answering test items thereby producing relatively low HCI_i values. On the other hand, because fewer comparisons are needed for linear hierarchies, the HCI_i is less sensitive to the slips that students make in answering test items thereby producing relatively high HCI_i values. In a similar manner, as the number of attributes increases, more items are associated with the reduced Q matrix. As a result, hierarchies with a larger number of attributes are typically associated with lower HCI_i values.

Identifying Critical Values for Examining Person Fit

Results showed that critical values identified for testing person fit using the statistical approach were very liberal in the sense that all the critical values were negative, ranging from -0.28 to -0.79 for CV1 (critical values for distinguishing a good and a moderate person fit) and from -0.64 to -0.99 for CV2 (critical values for distinguishing a moderate and a poor person fit) (see Table 100). These values were close to the lower bound of the HCI_i (-1), suggesting that it is very easy for a response vector to be judged as fitting the attribute hierarchy thereby limiting the power of the HCI_i in identifying misfitting response vectors. A misfit of a student response vector relative to the attribute hierarchy suggests that inferences cannot be made about student performance based on the attribute hierarchy given that the student uses different knowledge and skills in solving test items from those specified in the hierarchy. As a result, failing to identify the misfit of a student response vector relative to the attribute hierarchy can falsely validate the inferences made about student's cognitive strengths and weaknesses and further lead to incorrect decisions about student performance. Therefore, the use of statistical approach for selecting critical values to examine person fit is not practical.

However, according to the author's practical experience with the HCI_i through the use of simulated and real data (Cui et al., 2004; Gierl et al., 2007; Leighton, Cui, & Cor, 2007), an HCI_i value of 0.80 and above would generally indicate a good fit between a student response vector and the expected response vector while a value greater than 0.6 would reflect a reasonable fit. Inferences should not be based on the attribute hierarchy about student performance if the student response vector produces an HCI_i value below 0.6, indicating that students likely used different sets of knowledge and skills from those specified in the attribute hierarchy when solving test items. These criteria are based on subjective judgment so they cannot be considered as infallible. However, they are more realistic and powerful than the initially proposed statistical criteria in detecting the misfit of student response vectors relative to the attribute hierarchy. Further research is needed to examine the reasonableness of these criteria for interpreting the HCI_i.

Identifying Critical Values for Examining Overall Model Fit

The critical values for evaluating the overall model fit were also identified across different simulation conditions. Unlike the critical values for person fit, the identified values for overall model fit did not appear to be overly liberal in identifying misfitting response data sets. The critical values for distinguishing a good and a moderate overall model-data fit (CV1) ranged from 0.65 to 1.00 while values for distinguishing a good and a moderate overall model-data fit (CV2) ranged from 0.22 to 1.00 (see Table 11).

For divergent hierarchies, critical values for the overall model fit were found to be influenced by the number of attributes and sample sizes. The number of attributes appeared to show a negative impact on the critical values for the overall model fit, where the critical values decreased as the number of attributes increased. The critical values specifically for CV2s, tended to slightly increase as sample size increased. However, for convergent and linear hierarchies, critical values identified for the overall model fit were consistently equal to 1.00 across different number of attributes and sample sizes, indicating that at least 50% of simulated response vectors produced a HCI_i value of 1.00.

Based on the simulation results and the author's practical experience with the HCI_i, for divergent hierarchies, a median HCI_i value of 0.80 and above would normally indicate a good fit of the observed data set relative to the attribute hierarchy. A median HCI_i value of 0.60 and above would generally indicate a moderate fit of the observed data set to the attribute hierarchy. A model with a median HCI_i value smaller than 0.20 should not be used as a basis for making inferences about student performances. However, as the number of attributes becomes relatively large, the criteria for interpreting the HCI_i can be relaxed to some extent. For convergent and linear hierarchies, on the other hand, a median HCI_i value greater than 0.80 must be achieved in order for a set of observed response vectors to reach an adequate fit to the attribute hierarchy. Additional research must be conducted to investigate the reasonableness of the proposed criteria of using the HCI_i to examine the overall model fit.

Chapter 5: Summary and Conclusion

The growing demand for providing diagnostic information about students' cognitive strengths and weaknesses has led measurement specialists to investigate new ways of developing test items and interpreting students' performance. In order to make specific inferences about students' cognitive strengths and weaknesses, cognitive models in educational measurement are needed to make explicit the knowledge and cognitive skills required to solve test items correctly.

Leighton, Gierl, and Hunka (2004) proposed a cognitive diagnostic model, called the attribute hierarchy method (AHM), which is aimed at integrating cognitive models with a psychometric technique to model students' cognitive performances. The AHM makes explicit the assumption that test items can be described by a set of hierarchically ordered attributes. This method is composed of three sequential stages. In the first stage, an attribute hierarchy is defined to describe the knowledge structures and skill processes that students would use in the test domain. The attribute hierarchy serves as a cognitive model that helps construct test items and facilitates the explanation and prediction of student performance. This is a critical step because the validity of the attribute hierarchy links directly to the accuracy of the inferences to be made about students from the AHM. In the second stage, the attribute hierarchy is used as a basis for developing test items to ensure that each component of the attribute hierarchy has been adequately measured. In the third stage, statistical classification procedures are used to classify each student into one of the knowledge states, derived from the attribute hierarchy, thereby making specific inferences about students' cognitive strengths and weaknesses.

While the trustworthiness of the inferences to be made with the AHM critically depends on the validity of the attribute hierarchy used in the test domain, the question remains of how to validate the attribute hierarchy in the test domain. This study introduced a person-fit statistic, the HCI_i, to evaluate the degree to which a student response vector is consistent with the attribute hierarchy, which could serve as one source of evidence for validating the attribute hierarchy used with the AHM. Simulation studies were conducted to evaluate the effectiveness of the HCI_i in terms of examining the degree to which a student response vector fits the attribute hierarchy. In addition, simulated data were also used to identify the critical values of the HCI_i for testing both person fit and overall model fit under different simulation conditions.

This chapter is divided into five sections. In the first section, the proposed statistic, the HCI_i, is described, followed by a brief summary of the methods used in the simulation studies. In the second section, a summary and discussion of the simulation results are presented. In the third section, the limitations of the simulation studies are discussed. In the fourth section, the conclusions from the present study are provided. In the fifth and final section, the directions for future research are outlined.

A Summary of the HCI_i Index and Simulation Methods

The HCI_i Index

The HCI_i is a person-fit statistic designed explicitly to investigate the degree to which a student response vector is consistent with the attribute hierarchy. It can be calculated by:

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

where

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

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

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

X_{i_g} is student i's score (1 or 0) to item g, and

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

The HCI_i depends on item complexity as determined by the prerequisite relationship among test items specified in the reduced Q matrix. The logic of the HCI_i is that a student should not be able to answer an item correctly unless the student has solved its prerequisite items successfully. The HCI_i ranges from -1 to +1, where a higher HCI_i value suggests a better statistical fit of the student response vector to the attribute hierarchy.

Simulation Methods

Simulation studies were conducted for two purposes. The first purpose was to assess the effectiveness of the HCI_i in evaluating the degree to which an observed response vector fits the attribute hierarchy used in the AHM. To assess the effectiveness of the HCI_i, the hypothesis is that data sets with lower percentage of slips should be able to produce higher HCI_i values than data sets with higher percentages of slips. The second purpose was to identify the critical values of the HCI_i for examining both person fit and overall model fit. Different critical values were sought for identifying good, moderate, and poor fitting response vectors for nine different attribute hierarchies. Data were generated based on the nine different attribute hierarchies by randomly adding a 5, 10, and 20 percentage of slips to the expected response vectors associated with each hierarchy for three sample sizes – 500, 1,000, and 1,500.

Simulation Results and Discussion

The HCI_i was found to be non-normally distributed and substantially negatively skewed across simulation conditions. Simulation results indicated that the HCI_i performs well in determining the degree to which observed response vectors are consistent with the attribute hierarchy. Higher HCI_i values were obtained consistently for data sets with lower percentages of slips. Critical values were also identified for each simulated data set.

However, it was found that the identified critical values for examining person fit using statistical procedures were very liberal, meaning that a student response vector could be easily identified as fitting the attribute hierarchy by using these critical values. Therefore, it was concluded that statistical approach was not practically feasible. However, according to the author's practical experience with the HCI_i through the use of simulated and real data (Cui et al., 2006; Gierl et al, 2007; Leighton et al., 2007), it was recommended that an HCI_i value of 0.80 be used to distinguish a good and a moderate fit of a student response vector to the attribute hierarchy and an HCI_i of 0.60 be used to distinguish a moderate and a poor fit. Although these criteria are based on subjective judgment, they are more realistic and powerful than the initially proposed statistical criteria in detecting the misfit of student response vectors relative to the attribute hierarchy. Further research is needed to examine the reasonableness of these criteria of the HCI_i for testing person fit.

Unlike the critical values for person fit, the identified values for the overall model fit did not appear to be overly liberal in identifying misfitting response data sets. For divergent hierarchies, critical values for the overall model fit were found to be influence by the number of attributes and sample sizes. The number of attributes appeared to show a negative impact, while sample size tended to show a slightly positive effect on the critical values for the overall model fit. On the other hand, for convergent and linear hierarchies, critical values identified for the overall model fit were consistently equal to 1.00 across different number of attributes and sample sizes, indicating that at least 50% of simulated response vectors produced an HCI_i value of 1.00.

Based on the simulation results and the author's practical experience, for divergent hierarchies, it was recommended that a median HCI_i value of 0.80 be used to distinguish a good and a moderate fit of the observed data set relative to the attribute hierarchy and use a median HCI_i value of 0.60 be used to distinguish a moderate and a poor fit of the observed data set to the attribute hierarchy. A model with a median HCI_i value smaller than 0.20 was not recommended to be used as a basis for making inferences about student performances. However, as the number of attributes becomes relatively large, the criteria for interpreting the HCI_i can be relaxed to some extent. For convergent and linear hierarchies, on the other hand, it was recommended that a median HCI_i value greater than 0.80 must be achieved in order for a set of observed response vectors to reach an adequate fit to the attribute hierarchy. Additional research must be conducted to investigate the reasonableness of the proposed criteria for using the HCI_i to examine the overall model fit.

Limitations of the Simulation Studies

One limitation of the simulation studies in this thesis was that the number of items was not manipulated for a hierarchy and therefore its effect might be confounded with those caused by the type of hierarchy structure and number of attributes. In the AHM, the columns of the reduced Q matrix specify the items needed to be developed in order to achieve maximum diagnostic information. However, multiple sets of items can be used to increase the total number of items for ensuring the reliability of the test. In the present study, only one set of items specified in the reduced Q matrix was considered.

Given that different reduced Q matrices likely contain a different number of columns, the number of items might have an impact on the HCI_i and its associated critical values. For example, as the number of attributes increases, the number of items, as specified by the columns of the reduced Q matrix, also increases. Although results showed that the number of attributes had a negative impact on the HCI_i, it cannot be determined from the results of the present study whether this impact is due to the increase in the number of attributes, the increase in the number of required items, or both.

Likewise, the effect of the hierarchy structure is also confounded by the effect of the number of required items. With the same number of attributes, the reduced Q matrix derived from the divergent hierarchy contains more items than the matrices from the linear and convergent hierarchies. Again, for the divergent hierarchy, relatively low HCI_i values could be caused by the divergent structure, the increase in the number of items, or both.

In addition, data were generated by randomly adding a certain percentage of slips into the expected response vectors in the simulations. Hence, only random inconsistencies between observed response vectors and expected response vectors were considered in this study. However, other non-random sources of assessment errors as discussed by Meijer and Sijtsma (2001) – such as sleeping (e.g., inaccurately answering the first questions in a test because of problems getting started), plodding (working very slowly and methodically and, as a result, failing to finish later items in a test), and cheating (e.g., copying answers from another student) – were not considered. Additional studies are needed to investigate whether the HCI_i is effective in detecting the misfit of a student response vector to the attribute hierarchy that is caused by these non-random unusual testing behaviors.

Conclusions

This study introduced a person fit statistic, the HCI_i, which is designed to statistically evaluate the degree to which a student response vector is consistent with the attribute hierarchy. Tentative criteria were established for evaluating both person fit and overall model fit using the HCI_i. For person fit, an HCI_i value of 0.80 were used as the critical value for distinguishing a good and moderate fit and an HCI_i value of 0.60 for distinguishing a moderate and a poor fit. For overall model fit, two sets of criteria were established. For divergent hierarchies, it was recommended to use a median HCI_i value of 0.80 to distinguish a good and a moderate fit and use a median HCI_i value of 0.60 to distinguish a moderate and a poor fit of the observed data set to the attribute hierarchy. For convergent and linear hierarchies, it was recommended that a median HCI_i value greater than 0.80 must be achieved in order for a set of observed response vectors to reach an adequate fit to the attribute hierarchy. Further research is needed to examine the reasonableness of these criteria.

Although developed within the AHM framework, the HCI_i should be helpful in other cognitive diagnostic models that are guided by cognitive models given that the index allows the researcher to evaluate the fit of the cognitive model relative to the student response data. Specially, the HCI_i should be useful for the Q matrix based cognitive diagnostic models, such as the rule space model (Tatsuoka, 1983, 1984, 1990, 1995), the unified model (Dibello, et al., 1995), the deterministic input noisy and gate model (DINA) (de la Torre & Douglas, 2004; Doignon & Falmagne, 1999; Haertel, 1989; Junker & Sijstma, 2001; Macready & Dayton, 1977; C. Tatsuoka, 2002), and the noisy input deterministic and gate model (NIDA) (Junker & Sijstma, 2001). In these models, the HCI_i can be directly used to evaluate the fit of the observed response vectors to the expectations of the Q matrix and consequently to determine whether students' cognitive processes differ from the cognitive processes hypothesized in the Q matrix.

The HCI_i is straightforward to use and therefore it can be applied to a large sample of students so the results from the HCI_i can be generalizable to the target population. A low HCI_i value suggests that misfit is found between the student response vector and the attribute hierarchy. However, there are at least two possible interpretations for the misfit. First, the misfit of the student response vector to the attribute hierarchy could be due to the fact that the cognitive model, as specified in the attribute hierarchy, fails to accurately describe the prerequisite relationship among the attributes. As a result, observed student response vectors are not consistent with the expectations associated with the given attribute hierarchy. It is also possible that the prerequisite relationships among attributes are specified correctly in the attribute hierarchy but the reduced Q matrix fails to correctly specify the attributes that students use in solving each item. In other words, students use different combinations of attributes when solving test items than those described in the reduced Q matrix. In order to determine what actually causes the misfit of the student response vector to the attribute hierarchy, further substantive analyses are required.

In addition, it should be noted that the HCI_i focuses on the hierarchical structure used to configure the attributes but gives little attention to the specification of each individual attribute. For example, suppose that attribute A is the prerequisite of attribute B and item 1 measures attribute A, and item 2 measures attributes A and B. Given these two items, four types of student response vectors are possible, including (0, 0), (1, 0), (0, 1), and (1, 1). According to the logic of the HCI_i, if the prerequisite relationship of attributes A and B is specified correctly, students are not expected to answer item 2 correctly unless they answer item 1 successfully. As a result, the student response vector (0, 1) is not consistent with the prerequisite relationship between attributes A and B, and will be judged to be misfit. The HCI_i can successfully detect the misfit caused by the misspecification of the prerequisite relationship among attributes. However, the HCI_i is not able to identify the inaccuracy associated with the knowledge structure and processing skills specified by each attribute. The misfit of student response vectors relative to the attribute hierarchy would not be found when the attribute hierarchy specifies the prerequisite relationship among attributes successfully but fails to provide a precise interpretation for each individual attribute. By only focusing on the prerequisite relationship among attributes, the HCI_i results cannot be used to validate the interpretation of individual attributes in depth. As a result, more substantive evidence is required to conclude whether the attribute hierarchy truly represents the knowledge and skills students use as they answer the items. The use of think aloud procedures and protocol analysis (Ericsson & Simon, 1993; Leighton, 2004) and the use of experimental studies (e.g., Tatsuoka & Tatsuoka, 1997) are two procedures that could be used to validate substantively the attribute hierarchy. These two procedures provide relatively detailed pictures of how students actually solve items on tests, which helps validate and interpret the HCI_i results.

Directions for Future Research

At least four areas require additional research. The first area is related to the critical values of the HCI_i for examining both person fit and overall model fit. Although simulation studies were conducted to identify these critical values under different simulation conditions, the number of items was not manipulated for each attribute hierarchy. Consequently, this factor was intertwined with the other two factors manipulated in the present study – hierarchy structure and number of attributes – which makes it difficult to separate the effect associated with each of them. Therefore, in future research, simulation studies should be conducted to systematically investigate the effect of the number of items on the HCI_i and the critical values for examining person fit and overall model fit. Additionally, general criteria were recommended for interpreting the HCI_i results in the present study. These criteria were partially based on subjective judgments so further research is needed to investigate their reasonableness.

The second area that needs additional research is to investigate the usefulness of the HCI_i in determining whether different attribute hierarchies should be used to describe the knowledge and skills used by students from different groups. Currently, the AHM makes the assumption that one cognitive model, as specified by the attribute hierarchy, can be used to account for the test performance of students from different groups (e.g., ability, gender, or language groups) in terms of the mastery and nonmastery of attributes illustrated in the attribute hierarchy. However, this assumption may not be tenable. For example, a recent study conducted by Leighton, Cui, and Cor (2007) suggested that students of high ability appeared to differ from average- to low-ability students not only in terms of possessing more attributes illustrated in the attribute hierarchy but also in terms of using different strategies from those used by average- or low-ability students while solving test items.

To investigate whether a cognitive model is equally accurate in interpreting the performance for students from different groups, the HCI_i could be employed. The HCI_i should be applied to each individual student response vector and the median HCI_i values for distinct groups can be calculated and compared to determine whether the cognitive model fits student response vectors from different groups evenly. A significant difference in the median HCI_i values among distinct groups suggests that students from distinct groups differ in terms of the strategies they use while solving test items. If so, the use of multiple cognitive models holds promise in providing more accurate representations of student knowledge structures and response processes and ultimately in improving the validity of the cognitive feedback produced with the AHM.

An issue raised by the use of the HCI_i for investigating group differences is how to determine whether differences in the median HCI_i values are sufficiently large to be able to conclude that a statistically significant difference exists. In addition, a challenge for future research is to investigate how to incorporate multiple cognitive models into test development and statistical classification techniques so that students' attribute profiles can be estimated accurately and efficiently.

The third area that requires further research is to investigate how to use the HCI_i to help determine the appropriateness of the grain size of the attribute hierarchy. According to Leighton and Gierl (2007), the grain size or the level of detail of a cognitive model is directly linked to the type of inferences made about student performance. In order to make inferences about students' cognitive strengths and weaknesses within a test domain, a cognitive model must be specified at a relatively small grain size. In the AHM, the attribute hierarchy serves as a cognitive model that specifies the knowledge and skills required in order for students to answer each item correctly. To make inferences about students' cognitive strengths and weaknesses, an attribute hierarchy with a relatively small grain size is required. However, there is a tradeoff between the grain size of a cognitive model and its ability to generalize the knowledge and cognitive skills over a group of students as well as a set of items. In order to address this issue, diversity of problem-solving processes must be considered.

Diversity of problem-solving processes is an empirical fact (Ericsson & Simon, 1993; Lohman, 2000). It is inevitable that students vary widely in the knowledge and cognitive skills they possess and use in solving test items. If an item elicits alternative solution paths, student diversity in solving problems is not surprising. Additionally, many researchers suggest that large intra-individual differences exist in strategy use. For example, Ericsson (1975) conducted a study that investigated subjects' sequences of moves in solving a sliding block puzzle. It was found that the similarity of move sequences among subjects starting from the same puzzle configuration was no greater than chance. Interestingly, the similarity of the solutions of each individual subjected to repetitions of the same problem was also no greater than chance. Accordingly, no single model could be expected to predict the exact sequences. However, when subjects' move sequences were analyzed at a more general or abstract level, most subjects followed the same orderly and predictable sequence. This study suggests that in order to uncover generalizable aspects of cognitive processes, cognitive models may have to be formulated in abstract terms, which may jeopardize the small grain size of models.

The AHM is mainly designed to help construct educational tests that are used to evaluate the performance of a group of students across a set of items. These tests commonly consist of multiple items to ensure that the inferences to be made about students' performance are highly reliable. As a result, the attribute hierarchy should be able to lend itself to being aggregated at different levels of grain sizes so it can be used for different assessment purposes (e.g., to produce fine-grained diagnostic information as well as coarser-grained summary information). Research must be conducted to investigate whether and how the grain size of the cognitive model influences the magnitude of the HCI_i. And further research is also needed to examine how the HCI_i

can help determine the appropriateness of the attribute hierarchy for capturing the commonality a group of students might have in terms of the knowledge and skills used in solving a set of items while at the same time making sufficiently detailed inferences about students' cognitive strengths and weaknesses.

The fourth area of future research is to investigate how to collect substantive evidence to complement the statistical results from the HCI_i for validating the attribute hierarchy. Two procedures could be used – the use of think aloud procedures and protocol analysis (Ericsson & Simon, 1993; Leighton, 2004) and the use of experimental studies (e.g., Tatsuoka & Tatsuoka, 1997). Although these procedures are often time consuming and costly, they can provide detailed information about student problem solving, which may help interpret and validate the HCI_i results and ultimately enhance the validity of the diagnostic feedback produced with the AHM.

Reference

Buck, G. (1990). *The testing of second language listening comprehension.* Unpublished doctoral dissertation, University of Lancaster, England.

Buck, G. (1991). The testing of listening comprehension: an introspective study. *Language Testing, 8 (1),* 67-91.

Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing, 11 (2),* 145-170.

Buck, G., & Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15,* 119-157.

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

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

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

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

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.

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

Donlon, T. F. & Fischer, F. e. (1968). An index of an individual's agreement with group determined item difficulties. *Educational and Psychological Measurement, 28,* 105-113.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171-191.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

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.

Ericsson, K. A. (1975). Problem-solving behavior with the Eight Puzzle IV: Process in terms of sequences of moves (No. 448). *Reports from the Department of Psychology.* Stockholm: University of Stockholm.

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,* 395-374.

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

Gierl, M. J. (in press). Using attributes to make cognitive inferences in skills diagnostic testing: an overview of the rule space model and attribute hierarchy method. *Journal of Educational Measurement.*

Gierl, M. J., Cui, Y., & Hunka, S. (2007). The attribute hierarchy method for cognitive assessment: Technical developments. Manuscript submitted for publication.

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., Zheng, Y, & Cui, Y. (in press). Using the attribute hierarchy method to identify and interpret differential group performance on tests. *Journal of Educational Measurement.*

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139-150.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Claussen (Eds.), *Measurement and prediction* (pp. 60-90). Princeton NJ: Princeton University Press.

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

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18,* 133-146.

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.

Hawell, M., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo Studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 12,* 55-73.

Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement, 4,* 105-126.

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). Hillsdale, 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. , Cui, Y., Cor, M. K. (2007) Testing expert-based and student-based cognitive models: an application of the attribute hierarchy method and hierarchical consistency index. Manuscript submitted for publication.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *27,* 3-16

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

Educational Measurement, 41(3), 205-237.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269-290.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. British Journal of Mathematical and Statistical Psychology, 35, 42-56.

Levine, M. V., & Drasgow, F. (1983). Appropriateness measurement: Validating studies and variable ability models. In D. J. Weiss (Eds.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 109-131). New York: Academic Press.

Lippman, R. P. (1987). An introduction to computing with neural nets. *IEEE ASSP Magazine, 4,* 4-22.

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.

Medler, D. A. (1998). A brief history of connectionism. *Neural Computing Surveys, 1,* 61-101.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18,* 311-314.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review and new developments. *Applied Measurement in Education, 8,* 261-272.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25,* 107-135.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439-483.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. F. Chipman, & P. L. Brennan (Eds.), *Cognitively Diagnostic Assessment*, Hillsdale, NJ: Erlbaum.

Mislevy, R. J., Almond, R. G ., Yan, D., & Steiberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K. B. Laskey & H. Prade (Eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 437-446). San Francisco, CA: Morgan Kaufmann.

Mislevy, R. J., Steinberg, L. & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55,* 75-106.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessment. *Review of Educational Research*, 64 (4), 575-603.

Sato, T. (1975). *The construction and interpretation of S-P tables.* Tokyo: Meiji Tosho.

Schooler, J. W., & Melcher, J. (1995). The ineffability of insight. In S. Smith, T. Ward, &

R. Finke (Eds.), *The creative cognition approach.* Cambridge, UK: Cambridge University Press.

Scriven, M. (1999). The nature of evaluation part I: relation to psychology. *Practical Assessment, Research & Evaluation*, 6 (11).

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden, 7*, 131-145.

Sijtsma, K, & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. Applied *Psychological Measurement, 16,* 149-157.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement, 45,* 433-444.

Snow R. E. & Mandinach, E. B. (1991). *Integrating assessment and instruction: A research and development agenda* (ETS Research Rep. No RR-91-8). Princeton, NJ: Educational Testing Service.

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. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95-110.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Fredrickson, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (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. Nichols, S. F. Chipman, & P. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 327-359), Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1996). Use of generalized person-fit indexes, Zetas for statistical pattern classification. *Applied Measurement in Education, 9,* 65-75.

Tatsuoka, K., Birenbaum, M., Lewis, C., & Sheehan, K. (1993). *Proficiency scaling based on conditional probability functions for attributes* (Tech. Rep. NO. RR-93-50). Princeton, NJ: Educational Testing Service.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7,* 81-96.

Tatsuoka, M.K., & Tatsuoka, K. K. (1997). Computerized cognitive diagnostic adaptive testing effect on remedial instruction as empirical validation. *Journal of Educational Measurement, 34*, 3-20.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 20,* 221-230.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-cultural Psychology, 13,* 267-298.

Varadi,F. & Tatsuoka, K. K. (1992). *BUGLIB Modified Version* [Computer program]. Trenton, NJ: Tanar Software.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: a study of conceptual change in childhood. *Cognitive Psychology, 24,* 535-585.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement.* Chicago: Mesa Press.

Whitely, S. E. (1980). Multicomponent latent trait model for ability tests. *Psychometrika, 62*, 495-542.

Appendix A

Mathematica Functions for Calculating the Hierarchy Consistency Index

rndoff[n_,d_] rounds a real number *n* to a number with *d* decimal digits.

```
Rndoff[n_,d_]:= N[10 -d Round[n 10d]]
```

Index[reQ_,n_] creates a vector of item numbers in the *reQ* that require the subset of attributes measured by item *n*, where *reQ* is the reduced Q matrix.

```
Index[reQ_,n_]:=Module[{reQ1,indexF,index},
   reQ1=Map[Plus[Transpose[reQ][[n]],#]&,Transpose[reQ]];
   indexF=Range[Dimensions[reQ][[2]]];
   index=Union[Flatten[Map[Position[#,-
     1]&,Transpose[reQ1]]]];
   index=Complement[indexF,index,{n}];
Return[index];
];
```

HCI[obs_,reQ_,d_] calculates the HCI_i for examinee *i* with the response vector *obs*, with *d* decimal places. The algorithm briefly is as follows:

1) Create a vector of item number that examinee *i* answered correctly
2) Create a vector of item number that requires the subset of attributes measured by each item that is answered correctly by examinee *i*
3) Calculate the total number of comparisons
4) Calculate the HCI_i and return the value with d decimal digits

Calls: rndoff[n_,d_]; Index[reQ_,n_]

```
HCI[obs_,reQ_,d_]:=Module[{n,NC,HCI},
   n=Flatten[Position[obs,1]];
   index=Index[reQ,#]&/@n;
   NC=Length[Flatten[Index[reQ,#]&
               /@Flatten[Position[obs,1]]]];
    If[NC=<0,NC=1];
    HCI=1-2*Total[1-obs[[#]]&/@Flatten[index]]/NC;
   Return[{rndoff[HCI,d]}];
   ];
```

Appendix B

Mathematica Functions for Simulating Data

np[mn_,std_,i_] calculates the probability of the occurrence of the total score *i* under the normal curve with a mean of *mn* and a standard deviation of *std.*

$$\texttt{np[mn\_, std\_, i\_] := Integrate}\left[\frac{\texttt{Exp}\left[-\frac{(\texttt{x}-\texttt{mn})^2}{2\,(\texttt{std})^2}\right]}{(2\,\pi)^{0.5} * \texttt{std}}, \{\texttt{x}, \texttt{i}-0.5, \texttt{i}+0.5\}\right]$$

frequency[erp_,nexaminee_] calculates the frequency of occurrence of the expected response vector *erp* given the total number of examinees is *nexaminee*. The algorithm briefly is as follows:

1) Calculate the total score for each expected response vector.
2) Calculate the mean and standard deviation of the total scores.
3) Consider a normal curve with x-axis metric of total-score units. The real limit of each total score is used to construct an interval between which the area under the normal curve is the proportion of the occurrence of the total score.
4) The proportion of each total score is divided by the sum of all the proportions since the sum is not equal to 1.
5) Because several expected response vectors may lead to a same total score, the proportion of the expected response vector is equal to the proportion of its total score divided by the number of expected response vectors of the same total score.
6) Multiply the proportion of each expected response vector by the total sample size desired to get the frequency of the expected response vector.

Calls: np[mn_,std_,i_]

```
frequency[erp_,nexaminee_]:=
    Module[{totalscore,u,sd,np1,ntotal,np2,np3,f},
    totalscore=Total[Transpose[erp]];
    u=Mean[Total[Transpose[erp]]];
    sd=(Variance[Total[Transpose[erp]]])^0.5;
    np1=np[u,sd,#]&/@totalscore;
    ntotal=Count[totalscore,#]&/@totalscore;
    np2=np1/ntotal;
    np3=np2/Total[np2];
    f=Round[np3*nexaminee];
Return[f];
];
```

RandomRelist[x_List] gives a list with the same members as the input list, *x*, but in a random reordering.

```
RandomRelist[x_List]:=Block[{n=x,p},
    Do[p=Random[Integer,{1,i}];
       n[[{p,i}]]=n[[{i,p}]],
       {i,Length[x]}
    ]; (*end do*)
n];
```

Note: **RandomRelist[x_List]** could be replaced by the function, **RandomPermutation[x].** In order to use the latter function, the package **<<DiscreteMath`Combinatorica`** must be first read into Mathematica.

slipsgen[nex_,erm_,h_,asp_] creates a matrix of nex expected response vectors from expected response matrix and then randomly generates slips of form from 1 to 0 and of form from 0 to 1 at a probablity of *asp* level. *h* is employed to control the random seed used in the simulation so that results can be replicated. In this study, h was assigned from 1 to 100 separately for the 100 data sets under each simulation condition. The algorithm briefly is as follows:

1) Calculate the frequency of occurence of each expected response vector from the expected response matrix, given the total number of examinees.
2) Create the data matrix of *nex* expected response vectors
3) Calculate the number of random slips (*nslips*) by multiplying the total number of examinees with the probability of slips.
4) For each item, *nslips* examinee responses are randomly selected, and altered to 1 if a correct response is selected or altered to 0 if an incorrect response is selected.

Calls: frequency[erp_,nexaminee_] ; RandomRelist[x_List];

```
slipsgen[nex_,erm_,h_,asp_]:=
    Module[{j,nslips,f,serm,a,g,k,index},
    f=frequency[erm,nex];
    serm=Flatten
     [Table[erm[[#]],{f[[#]]}]&/@Range[Length[erm]],1];
    nslips=Round[asp*Length[serm]];
    j=1;
    While[j=<Length[Transpose[erm]],
      a=Transpose[serm][[j]];
      SeedRandom[h*100+j];
```

```
          index=Take
             [RandomRelist[Range[Length[serm]]],nslips];
          k=1;
          While[k=<Length[index],
             g=index[[k]];
             a[[g]]=1-a[[g]];
             k++
          ];(*end while*)
          serm=Transpose[ReplacePart[Transpose[serm],a,j]];
          j++
         ];(*end while*)
    Return[serm];
    ];
```