University of Alberta

The Hierarchy Misfit Index: Evaluating Person Fit for Cognitive Diagnostic Assessment

by

Qi Guo

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

Master of Education

in

Measurement, Evaluation and Cognition

Educational Psychology

Fall, 2012

Edmonton, Alberta

ABSTRACT

As cognitive diagnostic models (CDMs) become increasingly popular in modern educational measurement, it is important to develop a person fit index that examines the appropriateness of a CDM for each individual examinee. The purpose of this study is to propose a new person fit index, the hierarchy misfit index (HMI), for CDMs, and test the power and type 1 error of the HMI at detecting misfitting item response vectors using a simulation study. The results of the simulation study showed that the HMI had high powers and acceptable type 1 errors when a test consisted of highly discriminating items. But when a test consisted of low discriminating items, the HMI's type 1 errors were too high to be acceptable. A comparison was also made with a previously developed person fit index, the hierarchical consistency index, (HCI). The results showed that the HMI performed better in high item discrimination conditions.

ACKNOWLEDGEMENTS

I would like to express my thanks to several people that have helped me through this journey. First, I want to thank all my supervisors. My master supervisor, Dr. Jacqueline Leighton, patiently guided me through the entire thesis. I have learned so many valuable lessons from her, especially about writing. Not only she taught me how to write, but also she made me realize that how writing can help improve understanding. I also want to thank my current PhD supervisor, Dr. Ying Cui, who is so patient and encouraging. Throughout my master study, I have made a number of mistakes. Dr. Ying Cui is always patient and forgiving with me. At several times, I felt overwhelmed by the thesis. Thanks Dr. Ying Cui for giving me the encouragement.

Then, I want to express my thanks to my friends (other CRAME students). I want to thank Alexander Riedel, not only for the many interesting discussions we had about my thesis and all the proofreading he had done for me, but also for his impact on my attitude toward learning and education. When I first came in CRAME, I was not sure about my goal and my academic motivation was more extrinsic than intrinsic. It was after I met Alex, I began to see how pure and intrinsic learning can be and how selfless teaching can be. I also want to thank my other friends and colleagues, Hollis Lai, Wei Tang, and Amin Mousavi who provided many insightful feedbacks for my thesis.

In the end, I want to express my thanks to my family, my parents and my grandparents, for their unconditional love. I know every day, at the other side of

the ocean, my grandparents are whishing me well. I want to dedicate my work to my grandparents and wishing them happy and healthy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AHM: Attribute Hierarchy Method

CDM: Cognitive Diagnostic Model

CTT: Classical Test Theory

HCI: Hierarchy Consistency Index

HMI: Hierarchy Misfit Index

IRT: Item Response Theory

CHAPTER 1 INTRODUCTION

Inaccurate educational measurements could negatively impact students' learning. Therefore, it is crucial to design educational measurements that can reliably and accurately measure students' knowledge and skills, and are informative for instruction and learning. In order to generate such measurements, a variety of psychometric procedures must be considered in both the test development and the interpretation of results. Otherwise, large-scale assessment programs, especially those that result in high stakes decisions about students, would be open to legal action.

During test development and interpretation, the psychometric procedures that must be considered involve what are called "measurement models." In non-technical terms, a measurement model is a simplified explanation of how different factors, such as test difficulty and students' abilities, influence students' test scores. Mislevy (2006) provides a more formal definition of measurement model: a measurement model describes the mathematical relationship among a variety of measurement variables including students' observed responses (e.g., test scores), their underlying level of achievement or ability (e.g., knowledge and skills), the test or item characteristics (e.g., difficulty), and measurement error. Examples of measurement models include classical test theory (CTT; Haertel, 2006), item response theory (IRT) models (e.g., 1PL, 2PL & 3PL; for a review see Hambleton, Swaminathan, & Rogers, 1991), and the recently developed cognitive diagnostic models (CDM; for a review see Leighton & Gierl, 2007). These measurement models differ in their assessment goals, and as a result, they differ in the level of

detail of inferences they can support about students. For example, CTT and IRT models are designed with the purpose to compare and rank students. Therefore, their primary focus is to infer a student's location on an underlying ability continuum. CDMs, on the other hand, are designed with the purpose to inform teaching and learning. Therefore, their primary focus is to support inferences about the knowledge and skills students have acquired that permit them to answer test items correctly. Regardless of their different foci, the accuracy of the inferences made from a measurement model depends on how well the model's assumptions are reflected by the student's responses. For example, if a student is able to consistently and correctly answer difficult test items, most measurement models would predict that the student should also be able to consistently and correctly answer easy items. But if the student consistently fails to answer the easy test items defined by a measurement model, we know it may be inappropriate to make inferences about this student based on the measurement model. Therefore, a critical area of research in educational measurement is to evaluate whether a student's item response pattern is logically consistent with the measurement model's expectations (Cui, & Leighton, 2009).

Attempts to assess the consistency between a student's item-response pattern and a measurement model's expectations have led researchers to the studies of *person-fit statistics*. While various person-fit statistics are available for IRT models (for a review see Meijer & Sijtsma, 2001), very few person-fit statistics are specifically designed to examine the fit of a student's item response pattern to CDMs. One recently developed person-fit index for CDMs is the

hierarchy consistency index (HCI) (Cui, Leighton, Gierl, & Hunka, 2006). Cui and Leighton (2009) apply the HCI to detect different types of *misfitting item response patterns* (the definition of this term will be explained in details later). The results shows that while this index is powerful at detecting certain types of model misfits (e.g., misfits that result because students are guessing randomly), its power at detecting other types of misfitting item response patterns (e.g., misfits that result because the measurement model is incorrectly specified) needs to be improved (Cui & Leighton, 2009).

Purpose of the Study

Therefore, the objectives of the current study are to a) propose a new person-fit statistic called the hierarchy misfit index (HMI), which attempts to address some of the weaknesses of the HCI, b) conduct a simulation study (i.e., using a computer to generate hypothetical student item-response patterns that resemble real student item-response patterns), and c) assess the power and type 1 error of the proposed HMI in identifying misfitting item response vectors. There are two specific research questions:

1. Are the HMI's power and type 1 error acceptable across different simulation conditions?
2. Does the HMI's performance in terms of power and type 1 error offer an improvement compared to the HCI?

Organization of the Thesis

The thesis is divided into six chapters. The introduction is presented in Chapter 1 shown above. The literature review is presented in Chapter 2. It starts with an overview of CDMs. Then, an example of CDM, the attribute hierarchy method (AHM), is reviewed in detail. The AHM is chosen as an example of CDM because the simulation study is conducted under the AHM framework. The last part of Chapter 2 reviews the HCI, including its formula, interpretation, and power. The proposed HMI is explained in detail in Chapter 3. The research design and data simulation procedures are provided in Chapter 4. The results of the analysis are reported and discussed in Chapter 5. The answers to the research questions, limitations of the study, future research directions, and conclusions are presented in Chapter 6.

CHAPTER 2 LITERATURE REVIEW

Cognitive Diagnostic Models (CDMs): An Overview

While traditional measurement models such CTT and IRT are successful at ranking students on an underlying or latent ability continuum, they offer little information about students' cognitive strengths and weaknesses, and as a result, they are often not informative for instruction and learning (Nichols, 1994). For example, when a student fails to answer a test item correctly, the only inferences supported by CTT or IRT are that the student's ability is below the threshold that is required to answer the item correctly, or that the student makes a careless mistake by chance (e.g., the student was distracted while responding to the item). Such inferences are not informative for helping teachers improve their instruction and helping students focus on their learning. It is for this reason that CDMs were developed with the goal to uncover students' knowledge and skills. CDMs attempt to identify students' knowledge and skills in small parts, where each part represents a small piece of knowledge or skill a student needs to be successful in answering a test item. These small pieces of knowledge or skill are called attributes. Unlike other measurement models, CDMs are not designed to be used on an existing educational test, that is, one that has already been created. In fact, CDMs are designed to inform the development of test items, after the attributes of interest are identified. With CDMs, test items are developed to probe different combinations of the attributes. The end goal of CDMs is to identify which attributes a student has mastered and which attributes the student has not mastered (this will be referred to as an attribute pattern), and hopefully provide useful

diagnostic information to teachers about students' cognitive strengths and weaknesses.

In the past three decades, many CDMs have been proposed (e.g., Fischer, 1973; Embretson, 1984; Tatsuoka, 1983; mislevy, Steinberg, & Almond, 2003: Leighton et al., 2004). While they differ in the specific mathematical formulation, most of them share the common goal to support inferences about students' mastery of attributes. In this paper, Leighton et al.'s AHM (2004) will be introduced as an illustration of CDMs because the HCI was initially developed under the AHM framework.

The Attribute Hierarchy Method (AHM)

The AHM (Leighton, Gierl, & Hunka, 2004) is a recently developed CDM that is an extension of Tatsuoka's rule space model (Tatsuoka, 1983). Compared to other CDMs, the unique feature of the AHM is that it assumes that the attributes students use to answer test items are hierarchically related. This means that some attributes are prerequisites for other attributes. For example, in order to carry out the multiplication of 2-digit numbers, addition and the multiplication of single digit numbers need to be mastered first. Consequently, if a student is able to carry out the multiplication of 2-digit numbers, the AHM would predict that the student is also able to carry out addition and simple multiplication. The advantage of considering the hierarchical relationships among the attributes is that it reflects the structure found in many cognitive psychological theories (see Leighton & Gierl, 2011); for example, well known cognitive theories such as

Piaget's stage developmental theory describes cognitive skills as hierarchically related (Piaget, 1983). Even Bloom's taxonomy, which is not an empirically-verified cognitive theory but nonetheless widely applied, is hierarchically ordered (Bloom, 1956). The AHM does have limitations, however. Potential disadvantages of the AHM include the difficulty in specifying the hierarchy of attributes for educational tasks and the possibility that students may use different knowledge and skills from those specified in the attribute hierarchy to answer items (Leighton, Cui, & Cor, 2009).

The AHM is created in three sequential stages. In the first stage, an *attribute hierarchy* (a model of the prerequisite relationships among attributes) is developed from verbal reports of students, by consulting testing experts, or by reviewing the literature (Leighton et al, 2009). An example of an attribute hierarchy of interest is provided in Figure 1 next page.

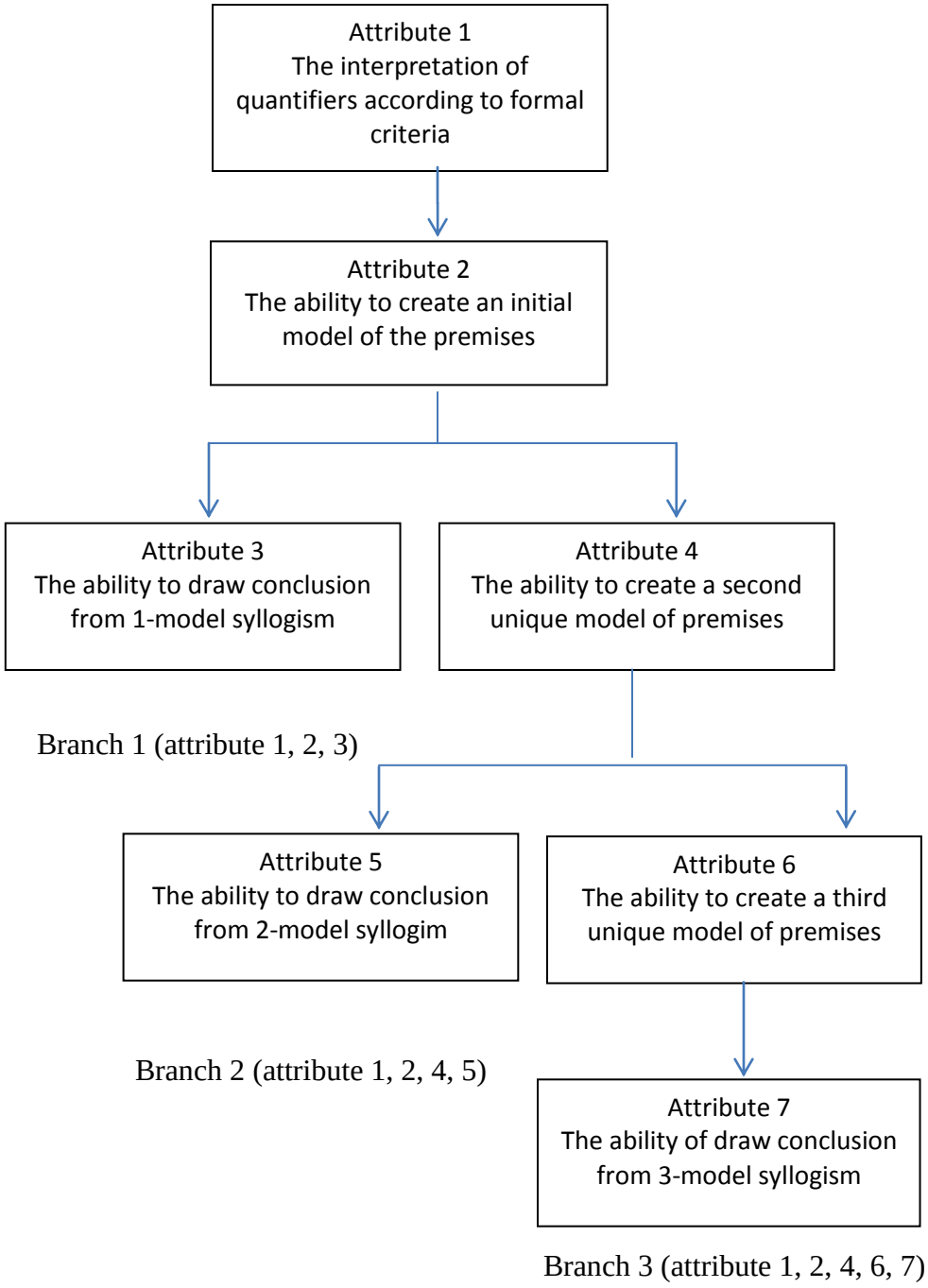

Figure 1. A seven-attribute hierarchy of categorical syllogism performance (taken from Leighton et al., 2004)

Figure 1 is an attribute hierarchy that attempts to model how people solve categorical syllogism problems. Categorical syllogism problems involve making a conclusion from two premise statements. An example of categorical syllogism

problem is as follow: given all A are B, and all B are C, what conclusion can we make about the relationship between A and C? The answer is all A are C. The attribute hierarchy in Figure 1 suggests that in order to solve categorical syllogism problems, people need to first be able to interpret the quantifiers (e.g., all, some) in the premise statements. Then, people need to be able to create mental representations of the premise statements. Some problems require only one mental representation (e.g., all A are B, and all B are C), other problems require two or three mental representations (e.g., some A are B, no B are C). The attribute hierarchy states that in order to solve problems that require multiple mental representations, people need to know how to solve problem that requires a single mental representation. The attribute hierarchy also suggests that after people create the mental representation(s), they need to be able to make a conclusion based on the mental representation(s). For a detailed review of the theory, see Leighton et al. (2004).

In the second stage of the AHM, the attribute hierarchy specified in the first stage is used to develop test items. In order to do this, test developers need to create a blueprint that outlines which attributes are required to correctly answer each item. Can test developers randomly decide the attribute combination required by each item? The answer is "no", if we assume there is a prerequisite relationship among the attributes. For example, is it possible to create an item that only requires the ability to do two-digit multiplication without requiring the abilities to do simple multiplication and addition? If we assume that in order to do two-digit multiplication, students need to know how to do single digit

multiplication and addition, then it is impossible to create an item that requires only the ability to do two-digit multiplication without requiring the abilities to do single digit multiplication and addition. This example shows that the attribute combination required to correctly answer each item is restricted by the prerequisite relationship among the attributes as specified by the attribute hierarchy. In other words, not all attribute combinations are logically coherent with the attribute hierarchy. If we put all the attribute combinations that are logically coherent with the attribute hierarchy together into a matrix, we get a blueprint of all the possible attribute combinations test items can measure. This blueprint is called the *reduced Q matrix* (Leighton, Gierl, & Hunka, 2004). The word "reduced" is used because attribute combinations that are not logically coherent with the attribute hierarchy are removed. Since the reduced Q matrix includes all possible attribute combinations consistent with the attribute hierarchy, it will be referred to as the *theoretical maximum reduced Q matrix* in this paper to distinguish it from other types of reduced Q matrices. The theoretical maximum reduced Q matrix for the attribute hierarchy shown in Figure 1 is shown in Table 1.

Table 1.

Theoretical Maximum Reduced Q Matrix for Figure 1

	Item														
Attribute	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12	i13	i14	i15
A1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
A2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
A3	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1
A4	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1
A5	0	0	0	0	0	1	1	0	0	1	1	0	0	1	1
A6	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
A7	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1

The theoretical maximum reduced Q matrix in Table 1 contains 7 rows and 15 columns of 1s and 0s. Each row represents an attribute, while each column represents a test item. Together, they show which attribute combination is required by each test item. For example, column 1 in Table 1 represents an item (item 1) that requires only attribute 1. Column 2 represents an item (item 2) that requires both attribute 1 and 2. It is important to note that while item 2 is coded to require both attribute 1 and 2, it does not need to directly probe attribute 1. That is, item 2 can be designed to probe attribute 2 directly and, indirectly, it also is expected to require attribute 1 since attribute 2 is dependent on attribute 1. For example, if a student is able to correctly answer an item that measures attribute 2, one can infer that the student can also correctly answer an item that measures attribute 1 as it is the prerequisite for attribute 2. For this reason, item 2 is coded as requiring both attribute 1 and 2. The rest of the columns can be interpreted in a similar fashion.

In practice, it is not necessary for a test to include all the attribute combinations in the theoretical maximum reduced Q matrix. This is because the development of items that measure complex attribute combinations can be difficult (e.g., it is difficult to develop an item that probes all the attributes at the same time). As a result, some attribute combinations in the theoretical maximum reduced Q matrix can be removed based on the needs of test developers (Cui & Leighton, 2009). However, it is important to note that not all attribute combinations in the theoretical maximum reduced Q matrix can be removed. Some attribute combinations in the theoretical reduced Q matrix are essential and must be retained. If any of the essential attribute combinations are removed, the diagnosis made from the AHM test will not be complete (Chiu, Douglas, Li, 2009). Therefore, it is important to understand which attribute combinations in the theoretical maximum reduced Q matrix are essential and which are not. Cui and Leighton (2009) named these essential attribute combinations as *simple attribute combinations*, and the non-essential attribute combinations as *complex attribute combinations*. A simple attribute combination only includes attributes belonging to a single branch of the attribute hierarchy (e.g., in Figure 1, attribute 1, 2, 3 belong to one branch, attribute 1, 2, 4, 5 belong to another branch, and attribute 1, 2, 4, 6, 7 belong to another branch), while a complex attribute combination includes attributes from at least two branches (e.g., attribute 5 and 6 are from different branches). For example, in Table 1, columns 1 to 4 and columns 6, 8, and 12 contain attribute combinations from only one branch and therefore can be classified as simple attribute combinations. The rest of the

columns contain attribute combinations from more than one branch and therefore can be classified as complex attribute combinations.

Compared to complex attribute combinations, simple attribute combinations are easier to measure and interpret. This is because in a simple attribute combination, items can be designed to explicitly measure the most difficult attribute in the combination. For example, column 4 in Table 1 represents item 4, which is coded to measure attribute 1, 2 and 4. Since attribute 1, 2 and 4 belong to one branch, item 4 measures a simple combination. As a result, item 4 can be designed to directly measure the most difficult attribute in the combination, namely, attribute 4. If a student is able to answer the item correctly, the mastery of the most difficult attribute the item probes (i.e., attribute 4) and all its prerequisite attributes (i.e., attribute 1 and 2) can be assumed. Otherwise, if the student fails to answer the item correctly, one can infer that the student has not mastered the most difficult attribute probed by the item (i.e., attribute 4). For complex attributes combinations, item design and interpretation are more difficult because items need to explicitly probe the most difficult attributes for at least 2 branches, and when a student fails to answer the item correctly, it is difficult to tell whether the student has not mastered the attribute from the first branch or the second branch, or both. For example, column 5 in the reduced Q matrix (1) represents item 5, which is coded to measure attributes 1, 2, 3 and 4. Since attribute 1, 2, and 3 are from branch 1 and attribute 1, 2 and 4 are from branch 2 (see Figure 1), this is a complex combination. If a student fails to answer item 5 correctly, it is difficult to tell whether the student has not mastered

attribute 3 or 4 or both. For this reason, it may be desirable to simplify the theoretically maximum reduced Q to contain only simple attribute combinations. The resulting matrix is called *theoretical minimum reduced Q matrix*. Continuing with the example, Table 1 can be further simplified to form the theoretical minimum reduced Q matrix, which is shown in Table 2.

Table 2.

Theoretical Minimum Reduced Q Matrix

	Item						
Attribute	i1	i2	i3	i4	i5	i6	i7
A1	1	1	1	1	1	1	1
A2	0	1	1	1	1	1	1
A3	0	0	1	0	0	0	0
A4	0	0	0	1	1	1	1
A5	0	0	0	0	1	0	0
A6	0	0	0	0	0	1	1
A7	0	0	0	0	0	0	1

For each column of the matrix, items can be designed to probe the most difficult attribute in the column. For example, the fourth column in Table 2 represents an item that directly probes attribute 4. Similarly, the fifth column in Table 2 represents an item that directly probes attribute 5.

In order to ensure that each attribute has been measured with adequate reliability, a sufficiently large number of items need to be designed to measure each attribute. At the same time, it is also important to consider practical constraints such as the time limit of the test and student fatigue. For example, in practice, a test can be designed in which each simple attribute combination in Figure 1 is explicitly measured by four items. Such a reduced Q matrix will be

called the *test reduced Q matrix*, which, as the name suggests, is the reduced Q matrix used for the actual test, usually including multiple items measuring each simple attribute combination. This can be represented by the reduced Q matrix shown in Table 3, which is derived by replicating each column of Table 2 four times.

Table 3.

Test Reduced Q Matrix (based on theoretical minimum reduced Q matrix)

Attributes	1-4	5-8	9-12	13-16	17-20	21-24	24-28
A1	1111	1111	1111	1111	1111	1111	1111
A2	0000	1111	1111	1111	1111	1111	1111
A3	0000	0000	1111	0000	0000	0000	0000
A4	0000	0000	0000	1111	1111	1111	1111
A5	0000	0000	0000	0000	1111	0000	0000
A6	0000	0000	0000	0000	0000	1111	1111
A7	0000	0000	0000	0000	0000	0000	1111

In Table 3, there are total 28 items. A closer examination of Table 3 reveals that some of the items require the same attribute combination. For example, items 5, 6, 7, and 8 all require attributes 1 and 2; items 9, 10, 11, and 12 all require attributes 1, 2 and 3. For this reason, we say items 5, 6, 7, and 8 belong to an *item type*, and items 9, 10, 11, and 12 belong to another item type. An item type includes all the items that require the same attribute combination. In Table 3, there are total 7 item types, and each item type is measured by 4 items.

In the third stage of the AHM, the test developed in the second stage is administered to students. Students' test responses are scored as either correct or incorrect. The scoring produces an *observed item response vector* for each

student. An observed item response vector is a vector that shows which items a student correctly answered, and which items the student failed to correctly answer. For example, for a test developed based on the test reduced Q matrix shown in Table 3, a student's observed item response vector may look like this: (1111 1111 0000 1111 0000 0000 0000), which shows that the student is able to correctly answers items 1 to 8, and 13 to 16, but unable to correctly answer the rest of the items.

The next step in the AHM involves interpreting the observed item response vectors. More specifically, each observed item response vector is analyzed to determine which attributes the student has mastered. In order to understand the procedure, three new concepts need to be introduced first: attribute pattern, expected attribute pattern, and expected item response vector. The *attribute pattern* is a vector that summarizes which attributes a student has mastered. For example, a student may have an attribute pattern of (1101000), which means that the student has mastered the attribute 1, 2 and 4 but not the rest of the attributes. In this example, there are seven attributes (see Figure 1). In total, there are $2^7 = 128$ possible attribute patterns. However, not all of the 128 attribute patterns are logically consistent with the prerequisite relationships among the attributes described by the attribute hierarchy in Figure 1. For example, the attribute hierarchy in Figure 1 specifies that attribute 1 is the prerequisite for attribute 2. This means if a student has mastered attribute 2, he/she must also have mastered attribute 1. Consequently, attribute pattern (1100000) is consistent with the attribute hierarchy, but attribute pattern (0100000) is not. Attribute

patterns that are logically consistent with the attribute hierarchy of an AHM test are called the *expected attribute patterns*. The expected attribute patterns for the attribute hierarchy in Figure 1 are shown in Table 4.

Table 4.

Matrix of Expected Attribute Mastery Patterns

	Attributes						
Attribute Pattern (AP)	A1	A2	A3	A4	A5	A6	A7
AP1	1	0	0	0	0	0	0
AP2	1	1	0	0	0	0	0
AP3	1	1	1	0	0	0	0
AP4	1	1	0	1	0	0	0
AP5	1	1	1	1	0	0	0
AP6	1	1	0	1	1	0	0
AP7	1	1	1	1	1	0	0
AP8	1	1	0	1	0	1	0
AP9	1	1	1	1	0	1	0
AP10	1	1	0	1	1	1	0
AP11	1	1	1	1	1	1	0
AP12	1	1	0	1	0	1	1
AP13	1	1	1	1	0	1	1
AP14	1	1	0	1	1	1	1
AP15	1	1	1	1	1	1	1

The expected attribute patterns can be derived by transposing the theoretical maximum reduced Q matrix. For example, Table 4 is derived by transposing the theoretical maximum reduced Q matrix in Table 1.

It is important to note that given a test reduced Q matrix, each expected attribute pattern has a corresponding *expected item response vector*. For example, we can predict that under ideal condition (i.e., the student makes neither lucky guess nor careless mistakes), a student who has the expected attribute pattern (1100000) will have an expected item response vector of (1111 1111 0000 0000

0000 0000 0000). This is because if the student has only mastered attribute 1 and 2, the student should be able to correctly answer all items that requires only attributes 1 and 2 (i.e., items 1 to 8 according to the test reduced Q matrix in Table 3). The expected item response vectors associated with the expected attribute patterns in Table 4 are shown in Table 5.

Table 5.

Matrix of Expected Item Response Vectors

Expected item response vector (EV)	Items						
	1-4	5-8	9-12	13-16	17-20	21-24	24-28
EV1	1111	0000	0000	0000	0000	0000	0000
EV 2	1111	1111	0000	0000	0000	0000	0000
EV 3	1111	1111	1111	0000	0000	0000	0000
EV 4	1111	1111	0000	1111	0000	0000	0000
EV 5	1111	1111	1111	1111	0000	0000	0000
EV 6	1111	1111	0000	1111	1111	0000	0000
EV 7	1111	1111	1111	1111	1111	0000	0000
EV 8	1111	1111	0000	1111	0000	1111	0000
EV 9	1111	1111	1111	1111	0000	1111	0000
EV 10	1111	1111	0000	1111	1111	1111	0000
EV 11	1111	1111	1111	1111	1111	1111	0000
EV 12	1111	1111	0000	1111	0000	1111	1111
EV 13	1111	1111	1111	1111	0000	1111	1111
EV 14	1111	1111	0000	1111	1111	1111	1111
EV 15	1111	1111	1111	1111	1111	1111	1111

Now, we can go back to answer the question of how to analyze an observed item response pattern to determine which attributes a student has and has not mastered. Basically, a student's observed item response vector is matched

with the most similar expected item response vector. Then, the most similar expected item response vector's corresponding expected attribute pattern is assigned to be the student's *estimated attribute pattern*. For example, a student has an observed item response vector of (1111 1101 0000 0000 0000 0000 0000). The most similar expected item response vectors would be (1111 1111 0000 0000 0000 0000 0000), which has a corresponding expected attribute pattern of (1100000). Therefore, the student's estimated attribute pattern would be (1100000). This pattern matching procedure can be conducted using an artificial neural network (Gierl, Zheng, & Cui, 2008). However, the details of the artificial neural network procedure are not relevant to the current thesis and will not be further discussed.

The Hierarchy Consistency Index (HCI)

Definition and Computation. The accuracy of the inferences the AHM will support depends on how well the attribute hierarchy can help predict students' observed item response vectors. In order to assess the consistency between the attribute hierarchy and the observed item response vectors, the HCI was developed. The HCI is a person-fit statistic that examines how well a student's observed item response vector matches the expected item response vector of the student based on the hierarchical relationship among the attributes (Cui, Leighton, Gierl, & Hunka, 2006). The HCI formula is given by:

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

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, where item j belongs to $S_{correct_i}$,

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 where item g belongs to S_j, 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 represents the number of misfits between the student's observed item response vector and the expected item response vector associated with the test reduced Q matrix (an illustration of the calculation procedure will be provided in the next paragraph). The HCI was formulated to range between -1 and 1, where -1 represents a total misfit between a student's observed item response pattern and corresponding expected item response vector, and 1 represents a perfect fit between the student's observed item response vector and an expected item response vector. If the HCI value is close to -1, it indicates that the student's knowledge and skills for solving the test items are not well represented by the attribute hierarchy specified by the AHM (since it does not match very well with one of the expected item response vectors).

To illustrate the calculation procedure of the HCI, the example provided in Cui and Leighton's (2009) study is presented here. Consider the attribute hierarchy presented in Figure 1 and the test reduced Q matrix in Table 3. Suppose a student's observed item-response vector is (1111 0000 1000 0000 0000 0000 0000), in which items 1 to 4, and 9 are correctly answered, namely $S_{correct_i}$ ={1, 2, 3, 4, 9}. According to the test reduced Q matrix (Table 3), item 9 requires attributes 1, 2, and 3. Since the student correctly answered item 9, he or she is considered to have mastered the attributes required by this item, namely, attribute 1, 2, and 3. Therefore, this student is expected to also answer items 1 to 4 (measuring attribute 1), 5 to 8 (measuring attributes 1 and 2), and 10 to 12 (measuring attributes 1, 2, and 3) correctly, because each of these items measures the same set or a subset of attributes required by item 9. That is, S9={1,2,3,4,5,6,7,8,10,11,12,}. In other words, for item 9, there are 11 comparisons that can be made among items: item 9 vs. items 1 to 8 (8 comparisons) and 10 to 12 (3 comparisons). Since the student failed to answer items 5 to 8 and 10 to 12 correctly, seven misfits are found out of 11 total comparisons between the student's observed item responses and the expected responses. Similarly, potential misfits can be calculated for items 1 to 4, which are also correctly answered by the student, S1={2,3,4}, S2={1,3,4}, S3={1,2,4}, and S4={1,2,3}. For item 1, there are three comparisons: item 1 vs. items 2, 3, and 4. Since items 2, 3, and 4 are all correctly answered by student i, there are no misfit found for item 1. Likewise, no misfits are found for items 2, 3, and 4. Overall, the total number of misfits is 7, and the total number of comparisons is

equal to 11+3+3+3+3=23. Hence, using the formula presented in the previous page, the value of the HCI for the student's observed response vector is 1-2*7/23=0.39.

Interpretation of the HCI. Strictly speaking, any HCI value that is lower than 1 indicates some level of misfit between the student's observed item-response vector and the expected item response vector (a misfit in HCI refers an inconsistency between the number of items the student is observed to answer correctly with the number of items the student is expected to answer correctly). However, in practice, it is not realistic to classify all HCI values smaller than 1 as misfitting. As a result, a cut score is needed to identify the misfitting HCI values. If an HCI value is greater than the cut score, it will be classified as "normal"; if an HCI value is smaller than or equal to the cut score, it will be classified as "misfitting". According to Cui and Leighton (2009), the cut score can be obtained by the following procedure. First, a computer is used to generate 2000 observed item response vectors that are consistent with the attribute hierarchy. Observed item response vectors that are consistent with the attribute hierarchy will be referred to as the *normal item response vectors*. These normal item response vectors are then modified by randomly introducing *slips* (changes from 1 to 0, or 0 to 1) to all the possible expected item response vectors. The probability of the slips is determined by the item discrimination power, which will be discussed in detail later. Second, the HCI values for the 2000 normal item response vectors are calculated. Third, the HCI values are ordered from lowest to highest value. Fourth, if we let alpha = 0.10 be the probability of misclassifying a

normal item response vector as misfitting, then the 10^{th} percentile of the rank ordered HCI values can be used as the cut score. According to Cui and Leighton (2009), an alpha of 0.10 rather than 0.05 is used to increase the power of the HCI (i.e., the probability of correctly indentifying misfitting item response vectors).

The power of the HCI in Identifying Misfitting Item Response Patterns. The cut score enables the classification of the HCI values and the corresponding observed item response vectors into normal and misfitting. Like any other classification procedure in statistics, researchers are interested to know the power of the classification procedure. Power here refers to the probability of correctly classifying a *misfitting item response vector* (i.e., an observed item response vector that is inconsistent with the attribute hierarchy) as misfitting. In order to estimate the power of the HCI in identifying misfitting item response vectors, Cui and Leighton (2009) conducted a simulation study. Part of the simulation study involved generating misfitting item response vectors using a second attribute hierarchy shown in Figure 2. These generated observed item response vectors were considered as misfitting with respect to the attribute hierarchy in Figure 1 because they were generated from a different attribute hierarchy (Figure 2). Then, the HCI values for these misfitting item response vectors were computed (based on the attribute hierarchy in Figure 1). Cui and Leighton (2009) found that the HCI's power for identifying these misfitting item response vectors was low to moderate (0.21 to 0.53) across different conditions (e.g., different number of items, different item discrimination power). Cui and Leighton (2009) argued that the relative low powers of the HCI were mainly due

to the fact that the attribute hierarchy in Figure 1 was only partially different from the attribute hierarchy in Figure 2. That is, the two attribute hierarchies still shared many similarities (e.g., both attribute hierarchies assumed the same prerequisite relationships among attributes 1, 2, 4, and 6).

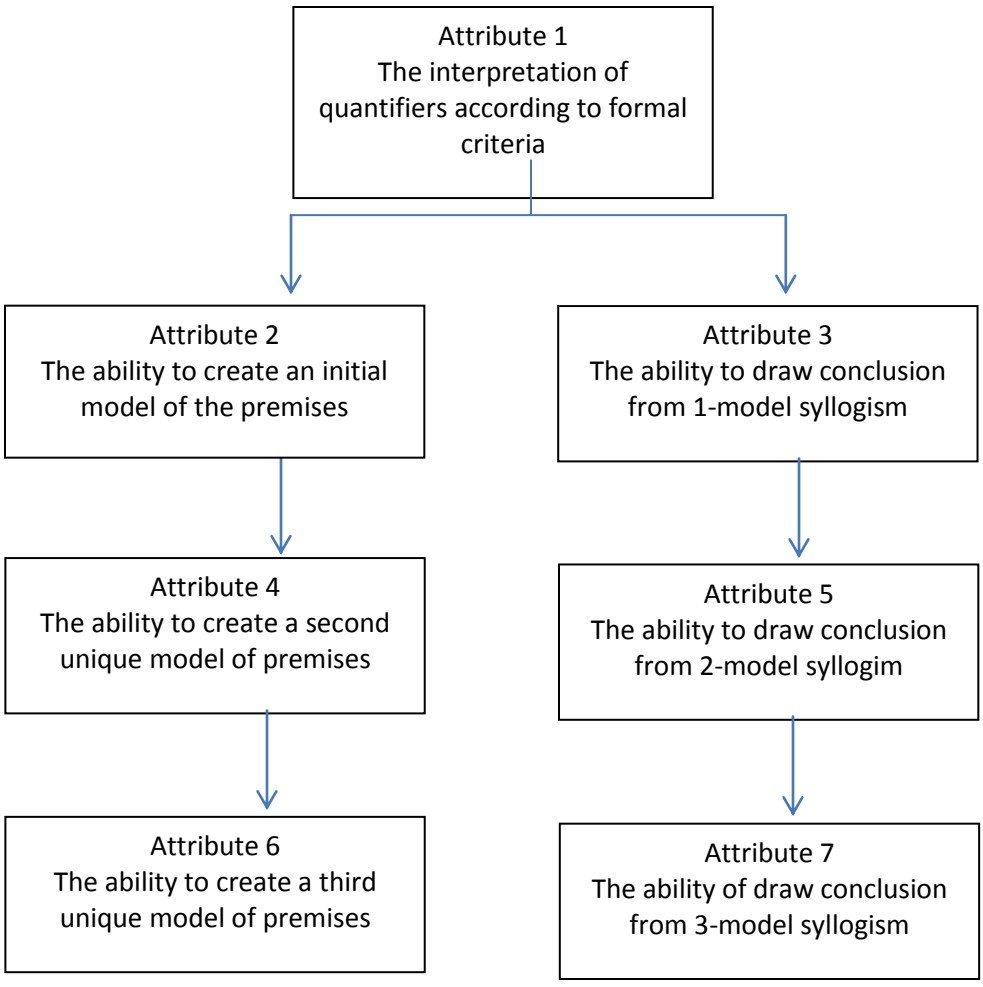

Figure 2. A hypothetical seven-attribute hierarchy of categorical syllogism performance

CHAPTER 3: THE HIERARCHY MISFIT INDEX (HMI)

In order to improve the power to detect misfitting item response vectors, a new person fit index, the HMI, is developed based on the HCI. Compared to the HCI, the most unique feature of the HMI is that it has two sub-indices that represent two types of misfits: *random misfit* and *systematic misfit*. The two sub-indices can be summed together to give an overall indication of person fit, the HMI. The definitions and computational formulas for the random misfit, systematic misfit, and the HMI will be presented in the following sections.

Random Misfit

Not all misfits indicate that the attribute hierarchy is misspecified. In other words, even if an attribute hierarchy is accurately specified, there will still be some misfits that occur due to random measurement error (e.g., a high achieving student accidentally makes a careless mistake, or a low achieving student correctly answers an item by a lucky guess). Such misfits will be referred to as *random misfits*. In practice, in order to identify the amount of random misfit, an operational criterion is needed. Random misfit is operationally defined as the inconsistency among test responses to items that belong to the same item type (as a reminder, an item type includes all the items that require the same attribute combination). This inconsistency can be mathematically represented by the *random misfit index*. The computational formula for the random misfit index is:

$$\textit{Random misfit index} = \frac{\sum M_w}{k}, \qquad (2)$$

where,

M*w* represents the number of minority items (it will be demonstrated later) that are inconsistent with the rest of the items that measure the same attribute combination, and

k is the total number of items in the test.

To illustrate the calculation procedure for the random misfit index, consider a test that only measures five attributes and each attribute is measured by four items. For simplicity, assume the five attributes are linearly ordered (form one branch), where attribute 1 is the easiest attribute and attribute 5 is the most difficult attribute. Suppose a student's observed item response vector is {1111 0000 1100 1000 1110}. To calculate the random misfit index, inconsistencies within each item type are examined separately at first. M*w* represents the number of minority items that are inconsistent with the rest of the items that measure the same attribute combination. In the observed item response vector, the first four items measure attribute 1, and the student is able to answer them correctly. There is no inconsistency among the items, so M*w* equals to 0. The second four items measure attribute 2. The student fails to answer any of them correctly. Again, there is no inconsistency among the items, so M*w* equals to 0. For the third set of four items that measure attribute 3, the student is only able to answer two of them correctly. Since two items (item 9, 10, correctly answered) suggest the student has mastered attribute 3 and two items (item 11, 12, incorrectly answered) suggest the student has not mastered attribute 3, there are inconsistencies among the third

set of four items. M*w* equals to 2 in this case because 2 items are inconsistent with the rest of the items. For the fourth set of four items that measure attribute 4, the student is only able to answer one out of four items correctly. Since one item (item 13) is inconsistent with the rest of the items, M*w* equals to 1. For the last set of 4 items that measure attribute 5, the student is able to answer three out of four items correctly. Since one item (item 20) is inconsistent with the rest of the items, M*w* equals to 1. The random misfit index equals to (0+0+2+1+1)/k. Since the test includes 20 items (4 items per attibute*5attribute=20 items, k=20), the random misfit index equals to (0+0+2+1+1)/20=0.2. Given the definition of Mw, the M*w* cannot exceed half of the number of items that measure the same attribute combination. Therefore, the sum of M*w* cannot exceed half of the total number of items in the test. As a result, the random misfit index has a range between 0 to 0.5, where 0 represents no random misfit, and 0.5 represents severe random misfit (e.g., {1100 1100 1100 1100 1100}).

Systematic Misfit

In contrast to random misfits, systematic misfits occur because the attribute hierarchy is misspecified, that is the attribute hierarchy does not accurately represent the prerequisite relationship among students' knowledge and skills. When the attribute hierarchy is misspecified, students' observed item response vectors will systematically differ from the expected item response vectors. For this reason, the term systematic misfit is used.

Similar to random misfits, systematic misfits need an operational definition so that they can be clearly defined in practice. But in order to operationally define systematic misfits, two new concepts need to be introduced: the *observed item type pattern*, and the *expected item type pattern*. The systematic misfit is operationally defined as the inconsistency between the observed item type pattern and the expected item type pattern. In the following sections, the definitions and calculation procedures of the observed item type pattern and the expected item type pattern are introduced. Then, the computational formula of the systematic misfit index is introduced.

Observed Item Type Pattern. The *observed item type pattern* is a vector that indicates which *item types* (i.e., items that measure the same attribute combination are said to belong to the same item type) a student appears to be able to consistently and correctly answer. The observed item type pattern is different from the observed item response vector in that the observed item type pattern shows which **item types** a student can consistently and correctly answer, while the observed item response vector shows which **items** the student correctly answered. To illustrate this, the previous example is continued, consider the student whose observed item response vector is {1111 0000 1100 1000 1110}. The student's observed item type pattern could be summarized arbitrarily as follows: {1 0 0 0 1}, which indicates that the student can consistently and correctly answering item type 1 and 5 but not item type 2, 3 and 4 because item types 2, 3, and 4 have 50% or fewer items answered correctly.

The observed item type pattern is also different from the attribute pattern because an item type does not necessarily correspond to an attribute. There could be more item types than attributes. This is because an item type can be designed to measure combination of attributes from different branches of an attribute hierarchy (i.e., complex attribute combination). Another important and distinguishing property of the observed item type pattern is that it is estimated entirely based on the student's observed item response vector, and it is does not depend on the assumption that the attribute hierarchy is correctly specified.

In order to estimate which item types a student can consistently and correctly answer, a cut score is needed to determine how many items of an item type a student needs to correctly answer in order to be considered as a master of the item type. There are two approaches to determine the cut score: determining the cut score based on the likelihood of mastery vs. non-mastery, and determining the cut score based on an arbitrary standard.

In order to determine the cut score based on the likelihood of mastery vs. non-mastery, we need to know the *item discrimination power*. In the AHM framework, item discrimination power includes two probabilities: the probability of answering an item correctly given a student has mastered the attribute combination probed by the item, and the probability of answering an item correctly given a student has not mastered the attribute combination probed by the item (Cui & Leighton, 2009). For example, an item has an item discrimination power of 0.6/0.2. This means if a student has mastered the attribute combination required by the item, the probability of correctly answering the item by the

student is 0.6; if the student has not mastered the attribute combination required by the item, the probability of correctly answering the item by the student is 0.2. If the item discrimination powers of all the items that belong to an item type are known, the probability of producing an observed item response vector given mastery or non-mastery can be calculated, and whether the student is classified as a master or non-master can be determined by which probability is larger. For example, suppose there are four items that belong to the same item type, and they all have an item discrimination power of 0.9/0.1. The probability of a student producing the item response pattern (1111) given that the student has mastered the attributes combination required by the items is 0.6561 (0.9*0.9*0.9*0.9=0.6561). The probability of producing the same item response pattern given the student has not mastered the attribute combination is 0.0001 (0.1*0.1*0.1*0.1=0.0001). Since the probability of producing this item response pattern given mastery is much larger than the probability of producing it given non-mastery, it is more likely the student has mastered the item type. Therefore, the response pattern (1111) can be classified as mastery. If the number of correct responses in the item type is decreased (e.g., {1110}), the probability of producing such observed item responses given mastery will decrease and the probability of producing them given non-mastery will increase. At one point, the probability given mastery will equal to the probability given non-mastery. After that point, if the number of correct responses keeps decreasing, the probability given non-mastery will become larger than the probability given mastery. The point where the two

probabilities are equal can be used as the cut score to determine whether the student has mastered the item type.

In order to compute a cut score based on the probabilities given mastery or non-mastery, additional formulas are required. For the purpose of simplicity, the item discrimination is assumed to be the same across the items. The formula for calculating the probability of producing an observed item response vector given mastery is:

$$P(c|a) = a^c * (1-a)^{l-c} * \frac{l!}{c!*(l-c)!}, \quad (3)$$

where,

a is the probability of correctly answering an item given mastery

c is the number of correct responses to items that belong to an item type

l is the number of items that measure the item type.

The probability of an item response pattern given non-mastery is:

$$P(c|b) = b^c * (1-b)^{l-c} * \frac{l!}{c!*(l-c)!}, \quad (4)$$

where

b is the probability of correctly answering an item given non-mastery

c and l are the same as before.

The formula for the cut score is the solution to the equation: $P(c|a) - P(c|b) = 0$ (c is the unknown in this equation). This solution can be computed by many equation solving programs (e.g., Mathematica). After the cut score is computed, the estimated item mastery pattern can be computed based on the comparison of the actual number of correctly answered items that belong to an item type and the cut score. If the actual number of correctly answered items is greater than the cut score, the student will be considered as having achieved mastery of the item type.

To illustrate these procedures, the hypothetical 5 attributes test example will be continued. Suppose the item discrimination power for each item is 0.6/0.2. Thus, "a" equals to 0.6, and "b" equals to 0.2. Since the test has 4 items measuring each attribute, "*l*" equals to 4. The cut score will be the solution to the following equation:

$$\text{Cut score} = \text{Solution}\,[P(c|a) - P(c|b) = 0] \qquad (5)$$

$$= Solution[0.6^c * (1-0.6)^{4-c} * \frac{4!}{c!*(4-c)!} - 0.2^c * (1-0.2)^{4-c} * \frac{4!}{c!*(4-c)!}]$$

In this case, the solution to the above equation is 1.55, meaning that if a student correctly answers two or more items that belong to the same item type, the student will be classified as a master of the item type (since 2 >1.55); and if the student fails to answer at least two items, the student will be classified as a non-master of the item type. Suppose a student's item response vector is {1111 0000 1100 1000 1110}. By applying the cut score of 1.55, the student's estimated item mastery profile will be {1 0 1 0 1}. For example, the student is able to correctly answer

four items that belong to item type 1. Since 4 is greater than 1.55, the student is considered as a master of item type 1. Similarly, since the student correctly answers 0 items that belong to item type 2, and 0 is smaller than 1.55, the student will be considered as a non-master of item type 2. It is important to note, this method assumes that the item discrimination powers are known, and that each item has approximately the same item discrimination power. The disadvantage of this method is that in practice, there may be some situations where these assumptions cannot be satisfied. When the assumptions of the probability method cannot be satisfied, the cut score can be determined arbitrarily. For example, the test developer can decide that if a student is able to correctly answer more than half of the items that belong to an item type, the student can be considered as master of the item type.

Expected Item Type Pattern. In contrast to the observed item type pattern, which is independent from the attribute hierarchy, the *expected item type pattern* is an item type pattern that is logically consistent with the attribute hierarchy. The expected item type pattern is different from the expected item response vector and the expected attribute pattern. In order to illustrate how the expected item type pattern is derived, the previous example is continued. The student has an observed item type pattern of {1 0 1 0 1}. Since the attribute hierarchy is linear, it means that the in order to master an attribute, its previous attributes must be mastered (e.g., in order to master attribute 4, attributes 1 to 3 must be mastered). The observed item type pattern suggests that the student is able to master item type 5 which measures attribute 5. From this, the attribute

hierarchy will predict that the student should also master attributes 1 to 4, and consequently be able to correctly answer item types 1 to 4. Therefore, the expected item type pattern for this student will be {1 1 1 1 1}. However, since the observed item type pattern is different from the expected item type pattern at item type 2 and 4, we can say there are two misfits between the observed item type pattern and the expected item type pattern.

Computational Formula for Systematic Misfit Index. The systematic misfit index formula is shown below:

$$\textit{systematic misfit index} = \sum_{j \in S_{mastered_i}} \sum_{g \in S_j} (1 - m_{i_g}), \quad (6)$$

where

$S_{mastered_i}$ includes item types that are considered to be mastered by student i according to the observed item type pattern,

j represents the item type that the student has mastered, therefore, $j \in S_{mastered_i}$,

S_j includes all the item types that require the subset of attributes measured by item type j, in other words, S_j includes item types that measure the prerequisite attributes of item type j.

m_{i_g} is a dichotomous value (1 or 0) that represents whether student i has mastered item type g (according to the observed item type pattern), where item g belongs to S_j.

To illustrate the calculation of the systematic misfit index, the 5-attributes test example is continued. The student's observed item response vector {1111 0000 1100 1000 1110} is converted to the observed item type pattern, {1, 0, 1, 0, 1}, as shown previously. The observed item type pattern shows that item types 1, 3, and 5 are mastered by the student. In terms of the formula, S_{mastered_i}={1, 3, 5}. Since j represents an item type that the student has mastered, $j \in S_{\text{mastered}_i}$ (i.e., j can take the value of 1, 3, or 5). We will first examine the most difficult item type the student has master, namely item type 5 (j=5). According to the hypothetical linear attribute hierarchy, item type 5 measures attributes 1, 2, 3, 4 and 5. Since the student has mastered item type 5, he or she is considered to have mastered all the attributes required by this item type, namely, attribute 1, 2, 3, 4 and 5. Therefore, this student is expected to also have mastered item type 1 to 4, which measure attribute 1, 2, 3, and 4 respectively. In terms of the formula, $S_{j=5}$ ={1, 2, 3, 4}. In other words, for item type 5, there are 4 comparisons that can be made among the item types: item type 5 vs. item types 1 to 4. Since the student has not mastered item types 2 and 4, two misfits are found. To put these in terms of the formula, g represents an item type that measures the subset of attributes of item type j. Therefore, g belongs to S_j. For j = 5, g can take the value of 1, 2, 3, or 4. The term, m_{i_g}, represents whether the student has mastered item type g. The student has mastered item type 1 and 3, but not 2 and 4, $m_{i_{g=1}} = 1, m_{i_{g=2}} = 0, m_{i_{g=3}} = 1, m_{i_{g=4}} = 0$. Therefore, the sum of $(1 - m_{i_g})$ equals to $(1-1)+(1-0)+(1-1)+(1-0) = 2$, meaning there are two misfits. Similarly, potential systematic misfits can be calculated for item types 3 and 1,

which also have been mastered by the student. For item type 3, there is one misfit, item type 2, which is expected to be mastered, but it is not. For item type 1, there is zero comparison because item type 1 measures attribute 1, which does not have any prerequisite attribute. Overall, the total number of misfits is 3 (2+1+0=3). Hence, the value of the systematic misfit index for the student's item response vector is 3.

Since the systematic misfit index represents the total number of misfits between the observed item type pattern and the expected item type pattern, it will always be an integer greater than or equal to 0. When the systematic misfit index equals to zero, it indicates there is no systematic misfit in the observed item response vector. When the systematic misfit index is greater than 0, it indicates the presence of systematic misfits in the observed item response vector. A special case is observed when an observed item type pattern is null (all 0), the systematic misfit index will be set to 999. This is because a null observed item type pattern provides little information about person fit. That is, a student who has a null observed item type pattern either randomly guesses every item, or skips the test. Neither behavior is meaningful for the evaluation of person fit.

Combination of Random and Systematic Misfit Index

Since the random misfit index is always a decimal (i.e., smaller than or equal to 0.5), and systematic misfit index is always an integer (i.e., greater than or equal to 0), they can be summed together to give an overall index of misfit, the HMI.

$$\text{HMI} = \text{random misfit index} + \text{systematic misfit index} \quad (7)$$

The decimal part of the HMI is the random misfit index and the integer part of the HMI is the systematic misfit index.

Using the HMI to Detect Misfitting Item Response Pattern

Both random and systematic misfits can be problematic in practice. For random misfit, if the number of random misfit is high, it indicates the items within an item type are not consistent with each other. This problem can be detected by a high random misfit index that is close to 0.5. The cut score can be determined arbitrarily according to the needs of the test developers. For systematic misfit, any systematic misfit index that is greater than 0 will be considered as serious violation of the attribute hierarchy. Combining random and systematic misfit index together, any HMI value that is greater than the cut score of the random misfit index will be considered as misfitting. To illustrate this property, consider the following HMI values: 1.30, 2.10, 0.26, and 0.10, and assume the cut score of the random misfit index is set to 0.25. If systematic misfits exist, then the integer part of the HMI indices will be greater than or equal to 1. For example, 1.30 and 2.10 have integers that are greater than or equal to 1 and thus they indicate systematic misfits. If the integer part of the HMI index is greater than or equal to 1, the HMI index as a whole will also be greater than or equal to 1, and consequently it will be greater than the cut score of the random misfit index. In this example, 1.30 and 2.10 are both greater than the cut score of the random misfit index, 0.25. This shows that if systematic misfits are present, the HMI

value will always be greater than the cut score of the random misfits. Severe random misfits are indicated by any HMI value that is greater than the cut score of the random misfits (i.e., 0.25). To summarize, if systematic misfits are present, the HMI value will be greater than the cut score of random misfit; if severe random misfits present, the HMI value will also be greater than the cut score of random misfit. Therefore, if the HMI value is greater than the cut score of the random misfit, the observed item response vector will be classified as misfitting.

CHAPTER 4 METHODS

Research Design

In order to test how well the HMI can detect misfitting item response vectors, a simulation study was conducted to examine the HMI's power and type 1 error. Power is the probability of correctly detecting misfitting item response vectors. Type 1 error is the probability of incorrectly classifying normal item response vectors as misfitting. The simulation study was conducted under the framework of Cui and Leighton's (2009) simulation study, with the intention to compare the HMI with the HCI. The data were generated based on the attribute hierarchy in Figure 1. Three factors were mainipulated: types of misfits, number of items, and item discrimination.

Types of Misfits. Three types of misfits were manipulated: model misspecification, creative responding, and random responding to every item. Normal responding (no misfit) was also investigated as a control. Misfits due to model misspecification refer to the misfits that occur when the attribute hierarchy does not accurately reflect student's knowledge and skills. Misfits due to creative responding occur when high achieving students fail to answer easy questions due to misinterpretation or boredom. Misfits due to random responding occur when students randomly guess in response to every item. The three types of misfits mentioned above can all be classified as systematic misfit because all of them involve systematic violations of the attribute hierarchy. It is important to examine the HMI's power at detecting the three different types of misfits because it has

been shown that the HCI has high power at detecting misfits due to random responding and creative responding; but relatively low power at detecting misfits due to model misspecification (Cui & Leighton, 2009). In addition to the above three types of misfits, normal item response vectors were also examined. Unlike the above three types of misfits, normal item response vectors do not contain any systematic misfit. This is because normal item response vectors are generated by randomly introducing low probability slips to expected item response vectors. Any unexpected responses in normal item response vectors are the result of random errors (i.e., random misfits). As a result, the percentage of normal item response vectors that are classified as misfitting by the HMI is an indicator of the HMI's type 1 error.

Number of Items. Number of items (or number of items per attribute) refers to how many items are used to measure each attribute. Three levels of number of items were manipulated: 3, 4, and 6 items per attribute. These three levels of number of items were chosen because they reflect realistic testing limitation in practice (Cui & Leighton, 2009). It is important to examine how the number of items per attribute influences the HMI's power because previous research has shown that the HCI's power increased as the number of items per attribute increased (Cui & Leighton, 2009).

Item Discrimination. As described earlier, in the AHM, item discrimination is represented by two probabilities: the probability of correctly answering an item given that a student has mastered the attributes the item requires; and the probability of correctly answering an item given that a student

has not mastered the attributes the item requires. Two levels of the item discrimination were included: high and low. High item discrimination was defined by two probabilities: 0.9 and 0.1, meaning that students who have mastered the attributes required by an item have a probability of 0.9 of answering the item correctly; and students who have not mastered the attributes the item requires have a probability of 0.1 of answering the item correctly. Low item discrimination was also defined by two probabilities: 0.6 and 0.2, and they can be interpreted in a similar fashion. According to Cui and Leighton (2009), HCI's power tends to be high when the items discriminations are high. Thus, it is important to examine how item discrimination influences the power of the HMI.

In total, four types of misfit (including the control), three levels of number of items, and two levels of item discrimination were examined in the simulation study producing a total of 4x3x2=24 conditions. Each condition was replicated 100 times. An overall picture of the study design is illustrated in Table 6. The power of the HMI was evaluated by the proportion of the three simulated misfitting item-response vectors (i.e., model misspecification, creative responding and random responding) that were correctly classified as misfitting by the HMI; and the type 1 error of the HMI was evaluated by the proportion of the simulated normal item-response vectors (control vectors) that were incorrectly classified as misfitting by the HMI.

Table 6.

Overall scheme of the stimulation study

			Number of items per attribute		
			3 items	4 items	6 items
Item discrimination power	High item discrimination	Type of misfits	2000 normal responses (control)	2000 normal responses	2000 normal responses
			2000 model misspecifications	2000 model misspecifications	2000 model misspecifications
			2000 creative responding	2000 creative responding	2000 creative responding
			2000 random responding	2000 random responding	2000 random responding
	Low item discrimination	Type of misfits	2000 normal responses (control)	2000 normal responses	2000 normal responses
			2000 model misspecifications	2000 model misspecifications	2000 model misspecifications
			2000 creative responding	2000 creative responding	2000 creative responding
			2000 random responding	2000 random responding	2000 random responding

Data Generation

As shown in Table 4, the 3 levels of number of items and 2 levels of item discrimination create a total of 6 forms of the tests (i.e., each cell in Table 4 represents a unique test). For each form of test, 2000 item response vectors are generated for each type of misfit (i.e., model misspecification, creative responding, random responding, and normal responding). Since the number of items and the item discrimination powers are relatively easy to manipulate, the following sections will be centered on the data generation procedure for the 4 types of misfits. The procedure to manipulate the number of item and the item discrimination power will be explained within the data generation procedures for the 4 types of misfits.

Generating Normal Item Response Vectors. Two thousands normal item response vectors were generated for each of the 6 test forms. The normal item response vectors were generated by introducing random slips (1 to 0, or 0 to 1) to expected item response vectors. The probabilities of the slips are determined by the item discrimination. Expected item response vectors and item discrimination power will be explained in the next two paragraphs.

Expected Item Response Vector. As mentioned previously in the AHM review section, the expected item response vectors are derived from the expected attribute patterns and the test reduced Q matrix. An expected attribute pattern shows which attributes a student has mastered (assuming the attribute hierarchy is correctly specified). The test reduced Q matrix shows which attributes are

required by each item. Combining the two pieces of information, we can predict the student's test response under ideal situation (no lucky guesses, and no careless mistakes). For example, assuming a test has Table 3 as its test reduced Q matrix, and a student has an expected attribute pattern of (1100000), the student's expected item response vector would be (1111 1111 0000 0000 0000 0000 0000). Since the student has only mastered attribute 1 and 2, the student will only be able to correctly answer items that require only attributes 1 and 2, namely, item 1 to 8.

Item Discrimination Power. The expected item response vectors do not have any aberration from the attribute hierarchy. This is unrealistic because students do guess and make careless mistakes. Therefore, random errors need to be added to the expected item response vectors. This can be done by randomly introducing slips (change 1 to 0, or 0 to 1) based on item discrimination. As mentioned previously, in the AHM framework, item discrimination values includes two probabilities: the probability of correctly answering an item given a student has mastered the required attributes of the item; and the probability of correctly answering an item given that a student has not mastered the required attributes of an item. The probability of the slips can be calculated from the item discrimination value. Specifically, the probability of making the 1-to-0 slip equals 1 minus the probability of correctly answering an item given that a student has mastered the required attributes of the item. The probability of making the 0 to 1 slip equals the probability of correctly answering an item given a student has not mastered the required attributes of the item. For example, for a test form that is made up of items with discriminations of 0.6/0.2, the probability of making the

1-to-0 slip would be set to 0.4 (i.e., 1 - 0.6 = 0.4), and the probability of making the 0-to-1 slip would be set to 0.2.

Using this method, around 133 normal responding vectors (i.e., 2000 total normal responding vectors/15 attribute patterns = 133.3) were generated for each attribute pattern in Table 5 for each of the 6 test forms.

Generating Model Misspecification Vectors. Two thousand model misspecification vectors were generated for each of the 6 test forms. The data were generated using the same procedures described in the previous section except that the attribute hierarchy in Figure 1 was replaced by the attribute hierarchy in Figure 2. This means that the attribute hierarchy in Figure 1 was no longer an accurate model for these item response vectors, which were generated based on the attribute hierarchy in Figure 2.

Generating Creative Response vector. Two thousand creative response vectors were generated for each of the 6 test forms. Creative response vectors were generated by first selecting four high ability expected attribute patterns from Table 4 (i.e., {1101011}, {1111011}, {1101111}, and {1111111}), and then changing the four expected attribute patterns to: {0101011}, {0111011}, {0101111}, and {0111111} to represent high achieving students who misinterpreted items that probed the easiest attribute (i.e., the first attribute). These modified attribute patterns were combined with the test reduced Q matrix to generate corresponding expected item response vectors. Then, random slips were introduced based on item discrimination as before.

Generating Random Response Vectors. Two thousands random response vectors were generated for each of the 6 combinations. Random response data were generated by randomly generating responses to each item (0 or 1) with a probability of 0.25 for a correct response, which corresponds with the probability of correctly guessing in response to a multiple choice item with four options.

Power and Type 1 Error

The HMI value for each simulated item-response vector was calculated. Since items in a test form were established to always have the same item discrimination power, the cut score for an observed item type pattern was determined by formula (5). The cut score for the HMI was arbitrarily determined to be 0.5, since in this particular study, only systematic misfits were of interest. That is if an HMI value was greater than 0.5, then it was classified as misfitting; otherwise, it was classified as normal. The powers of the HMI were determined by counting the proportions of item-response vectors in the misfitting conditions that were correctly classified as misfitting by the HMI. The type 1 errors of the HMI were determined by the proportion of the item-response vectors in normal conditions that were incorrectly classified as misfitting by the HMI.

The data generation and HMI calculation were executed by programs written in Mathematica code (Wolfram Mathematica 6). The data generation program can be requested from Cui, and HMI calculation program can be requested from Guo.

CHAPTER 5 RESULTS

The results of the HMI are presented in Table 7. The HMI's power and type 1 error will be reported across different conditions.

Table 7.

The Power and Type 1 Error of the HMI at Detecting Misfitting Item Response Vectors

				Number of items per attribute		
				3 items	4 items	6 items
Item discrimination power	High discriminating items	Type of misfits	Model misspecification	0.9834	0.9822	0.9945
				(0.00)	(0.00)	(0.00)
			Creative responding	1.0000	1.0000	1.0000
				(0.00)	(0.00)	(0.00)
			Random responding	0.9610	0.9864	0.9830
				(0.00)	(0.00)	(0.00)
			Type 1 error	0.1026	0.1350	0.0460
				(0.01)	(0.01)	(0.00)
	Low discriminating items	Type of misfits	Model misspecification	0.9055	0.9305	0.9353
				(0.01)	(0.01)	(0.01)
			Creative responding	1.0000	1.0000	1.0000
				(0.00)	(0.00)	(0.00)
			Random responding	0.9271	0.9202	0.9267
				(0.01)	(0.01)	(0.01)
			Type 1 error	0.6900	0.5085	0.4669
				(0.01)	(0.01)	(0.01)

Power

Types of Misfit. The HMI had high power at detecting different types of misfits. The HMI was most powerful at detecting creative responding (100% detection across 6 conditions). The HMI's powers at detecting model misspecification and random responding were high, ranging from 91% to 99% across conditions.

Number of Items. The HMI's power at detecting model misspecification and random responding tended to increase with the number of items that measure each attribute. Specifically, in 3-items per attribute conditions, the HMI's powers at detecting model misspecification and random responding ranged from 91% to 98%. In 4-items per attribute conditions, the HMI's powers ranged from 93% to 98%. In 6-items per attribute conditions, the HMI's powers ranged from 98 to 99%. Overall, as the number of items per attribute increased, the power of the HMI also increased.

Item Discrimination. In high item discrimination conditions, the HMI's powers at detecting model misspecification and random responding were significantly higher than in low item discrimination conditions. Specifically, the HMI's power at detecting model misspecification in high item discrimination conditions ranged from 98% to 99%, but in low item discrimination conditions, the power ranged from 91% to 94%. The HMI's power at detecting random responding in high item discrimination conditions ranged from 96% to 98%, but in low item discrimination conditions, the power ranged from 92% to 93%. Since

the HMI had 100% detection rate for creative responding at each condition, item discrimination did not appear to influence the HMI's power at detecting creative responding.

Type 1 Error

The Type 1 errors of the HMI were related to both the number of items and item discrimination. In high item discrimination conditions, the type 1 errors of the HMI were in the acceptable range (from 5 to 13%). In these high item discrimination conditions, it seemed the type 1 error of the HMI decreased as the number of item increased, even though the type 1 error was slightly higher in the 4 item per attribute condition than the 3 item per attribute condition. In low item discrimination conditions, the type 1 errors of the HMI were high and fell in the unacceptable range (from 47% to 69%). As the number of item per attribute increased, the type 1 error became smaller in low item discrimination conditions.

CHAPTER 6 DISCUSSION

The first part of this chapter will focus on answering the two research questions outlined in chapter one. Then, limitations and future research directions will be discussed.

Are the HMI's Powers and Type 1 Error Acceptable across Different Conditions?

In high item discrimination conditions, the HMI's power at detecting different misfitting item response vectors was high, ranging from 96.1% to 100%, and the HMI's type 1 errors were acceptable, ranging from 4.6% to 13.5%. As the number of items measuring per attribute increased, the power increased and type 1 error decreased. There is one exception that the HMI's type 1 error was lower in the 3-items high condition than in the 4-items condition (both refer to high item discrimination conditions). This is likely related to the property of the cut score for observed item type patterns. Specifically, in the 3-items condition, a student can have 4 possible scores on an item type: 0, 1, 2, 3 (i.e., how many items of an item type a student correctly answered). The cut score is a number between 1 and 2, therefore, it divides all the possible scores of an item type into two classes (i.e., 0 and 1 belong to non-mastery, 2 and 3 belong to mastery). But in the 4 items condition, a student can have 5 possible scores on an item type: 0, 1, 2, 3, 4. The cut score is right on 2. When a student correctly answers 2 out of 4 items, the likelihoods of mastery and non-mastery are equal. An arbitrary decision must be made, and in this study a score of 2 out 4 is classified as non-

mastery. This arbitrary classification decision makes the observed item type pattern less accurate. As a result, the 3-item condition's type 1 error is actually smaller than the 4-item condition. As the number of items measuring an attribute increases, this problem has less impact. For example, in the 6-item condition, even though the cut score problem still exists, the large number of items is enough to overcome the effect and reduces the type 1 error (4.6%).

For the low item discrimination conditions, the HMI's powers are still relatively high, ranging from 90.6% to 100%, but the HMI's type 1 errors become unacceptably high, ranging from 46.7 to 69.0%. While increasing the number of items measuring an attribute from 3 to 6 decreases the type 1 errors, it is not enough to make the type 1 error fall into the acceptable range. The high type 1 errors are likely due to that the low item discrimination power (0.6/0.2) increases the probability to produce severe random misfits which resemble systematic misfits. This result shows that in order for the HMI to function properly, it is crucial to have high item discrimination. The cut score problem mentioned in the previous paragraph is not a concern for the low item discrimination condition. This is because the probability based cut score for estimated item mastery patterns are: 1.16, 1.55 and 2.32, which do not require any arbitrary decision to make the classification.

Is there any Improvement from the HCI?

The HCI's powers at identifying misfitting item response vectors are shown in Table 8. Before comparing the two indices, it is important to note the

small differences between the two study designs. That is in Cui and Leighton's (2009) study, the numbers of items measuring an attributes were 2, 4, and 6, but in the current study, it was 3, 4, 6. The current study used 3 rather than 2-item condition because the author wanted to examine the possible cut score problem previously mentioned. Also, in Cui and Leighton's (2009) study, the HCI's type 1 error was always 10%. The rest of the design is identical.

Table 8.

Percentage of Misfitting Item-Response Vectors Correctly Identified by the HCI

				Number of items per attribute		
				K=3	k=4	k=6
Item discrimination power	High discriminating items	Type of misfits	Model misspecification	52.66 (0.01)	51.74 (0.01)	50.89 (0.01)
			Creative responding	91.14 (0.01)	97.11 (0.00)	99.06 (0.00)
			Random responding	93.01 (0.01)	99.04 (0.00)	99.82 (0.00)
	Low discriminating items	Type of misfits	Model misspecification	20.90 (0.01)	23.35 (0.01)	25.18 (0.01)
			Creative responding	88.18 (0.01)	99.43 (0.00)	99.97 (0.00)
			Random responding	53.01 (0.01)	71.68 (0.01)	82.38 (0.01)

In high item discrimination conditions, the HMI seems to function better. As shown in Table 8, the HCI's powers at detecting model misfits were relatively

low, ranging from 50.9% to 52.7%. The HMI had better powers (98.2% to 99.5%), yet maintaining comparable or better type 1 errors (10.1%, 13.5%, and 4.6%).

In low item discrimination conditions, both the HMI and the HCI were not functional. The differences were that the HCI tended to have low type 1 errors and low power, while the HMI tended to have high type 1 errors and high power. Both studies point out the importance of high item discrimination.

Beside power and type 1 errors, the HMI has the advantage of being easier to interpret. That is, the HMI has two indices that distinguish random and systematic misfits, while the HCI only has one value which makes the distinction difficult. Another advantage of the HMI is that it does not require simulation studies to determine the cut score for detecting misfitting item response vectors. In contrast, in order to use the HCI to detect misfitting item response vectors, a simulation study needs to be done first to determine the cut score. This is slightly more difficult to use. The HCI does have a cut score system that does not depend on simulation. However, this cut score system is arbitrarily determined and its power in identifying misfitting item response vectors has not yet been tested.

Limitations and Future Research Direction

Some parts of the simulation study were not realistic. For example, the study made all items have the same item discrimination. This is unlikely to occur in practice. But it is possible to develop items that have similar item discrimination values. Also, the simulation of creative responding and random

responding data may not be realistic. For creative responding, the simulation made high achieving students systematically fail at the easiest item type. Specifically, if there are 4 items measuring attribute 1, creative responding students will most likely fail 3 or 4 items. This may not be realistic unless they choose to skip all the easy items. For random responding, the simulation made the students randomly guess for every item. This is usually not the case because most students will be able to master some items, and randomly guess other items they do not know. The current study kept these limitations in order to be comparable with Cui and Leighton's (2009) study.

The author of the study believes that the HMI can still be improved. Currently, the systematic misfit index value cannot be interpreted in a numeric way. A large systematic index does not always mean more severe systematic misfits. Therefore, a future study direction would be to modify the systematic misfit index such that its numeric value can better represent the severity of the systematic misfit.

Conclusion

As cognitive diagnostic assessment becomes increasingly more popular in modern educational assessment, it is important to have a person fit index that indicates whether or not CDA is appropriate for individuals. The HMI provides a new way to assess person fit for CDA. Building on the HCI, the HMI improves the power at detecting misfitting item response vectors and also includes two sub-indices that provides easier and clearer interpretation of the nature of the misfit.

REFERENCES

Bloom B. S. (1956). Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain. New York: David McKay Co Inc.

Chiu, C., Douglas, J., and Li, X. (2009). Cluster analysis for cognitive diagnosis: theory and applications. *Psychometrika, 74(4),* 633-665. doi:10.1007/s11336-009-9125-0

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.

Cui, Y., and Leighton, J.P. (2009). The hierarchy consistency index: evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46(4),* 429-449. doi: 10.1111/j.1745-3984.2009.00091.x

Embretson, S.E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175-186. doi:10.1007/BF02294171

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Pschologica, 37,* 395-374. doi:10.1016/0001-6918(73)90003-6

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

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. doi:10.1111/j.1745-3984.2004.tb01163.x

Leighton, J. P., & Gierl, M. J. (Eds.) (2007). *Cognitive diagnostic assessment for education: Theory and practices.* Cambridge University Press.

Leighton, J.P., & Gierl, M.J. (2011). *The learning sciences in educational assessment: the role of cognitive models.* New York, NY, US: Cambridge University Press.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25,* 107-135. doi:10.1177/01466210122031957

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. (2006). Cognitive psychology and educational assessment. In R. Brennan (Ed.), *Educational Measurement (4th Ed.).* Phoenix, AZ: Greenwood.

Nichols, P.D. (1994). A framework for developing cognitively diagnostic assessment. *Review of Educational Research, 64(4),* 575-603. doi:10.3102/00346543064004575

Piaget, J. (1983). "Piaget's Theory". In P. Mussen (Ed.) *Handbook of child psychology*. Wiley.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95-110. doi:10.1007/BF02294208