JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

*Journal of Educational Measurement*
*Winter 2009, Vol. 46, No. 4, pp. 429–449*

# The Hierarchy Consistency Index: Evaluating Person Fit for Cognitive Diagnostic Assessment

**Ying Cui and Jacqueline P. Leighton**
*University of Alberta*

*In this article, we introduce a person-fit statistic called the hierarchy consistency index (HCI) to help detect misfitting item response vectors for tests developed and analyzed based on a cognitive model. The HCI ranges from −1.0 to 1.0, with values close to −1.0 indicating that students respond unexpectedly or differently from the responses expected under a given cognitive model. A simulation study was conducted to evaluate the power of the HCI in detecting different types of misfitting item response vectors. Simulation results revealed that the detection rate of the HCI was a function of type of misfit, item discriminating power, and test length. The best detection rates were achieved when the HCI was applied to tests that consisted of a large number of highly discriminating items. In addition, whether a misfitting item response vector can be correctly identified depends, to a large degree, on the number of misfits of the item response vector relative to the cognitive model. When misfitting response behavior only affects a small number of item responses, the resulting item response vector will not be substantially different from the expectations under the cognitive model and consequently may not be statistically identified as misfitting. As an item response vector deviates further from the model expectations, misfits are more easily identified and consequently higher detection rates of the HCI are expected.*

In educational assessment, measurement models (e.g., classical test theory [CTT], item response theory [IRT], or more recent cognitive diagnostic [CDM] models) play central roles in the development of test items and the interpretation of test results. The appropriateness of a measurement model used in test development links directly to the measurement accuracy of student performance. If a model fails to characterize the real aspects of student item response processes precisely, invalid inferences about the student performance will be made. As a result, it is critical to evaluate whether a student's item responses can be predicted or interpreted accurately by the model used in test development. Attempts to assess the fit of a student's item response vector to the measurement model have led researchers to studies of "person-fit" statistics. These studies have centered on evaluating how well a student's item response vector matches the expected vectors derived from a given model. The primary goal of these person-fit statistics is to identify students who are not being measured well by a test either because of invalid models (e.g., the assumptions of the model are not tenable), or perhaps because of aberrant response behavior such as cheating (i.e., copying answers from more able students), creative responding (i.e., answering easy items incorrectly for the reason of interpreting these items in a unique, creative manner), or random responding (i.e., randomly responding to multiple-choice items on the test).

 

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). According to Meijer and Sijtsma, these person-fit statistics can be grouped into two major categories: group-dependent statistics and IRT-based statistics. Group dependent person-fit statistics compare the observed item response vector to the expectations 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 relatively low proportion-correct scores are answered correctly but items with relatively high proportion-correct scores are answered incorrectly. Examples of group dependent person-fit statistics are Kane and Brennan's agreement, disagreement, and dependability indices, Harnisch and Linn's modified caution index, van der Flier's $U_3$, and Tatsuoka and Tatsuoka's norm conformity index $NCI_i$. Group dependent person-fit statistics can be used with tests that have been developed and analyzed with either a CTT or an IRT model, and where the unidimensionality of the test is assumed.

IRT-based person-fit statistics are specifically designed to evaluate the misfit of an observed response vector to an IRT model by calculating the IRT probabilities associated with a student's ability parameter and the item parameters. If according to the IRT model the probability of a correct response from a student is high, the hypothesis is posited that the student should answer that item correctly, and vice versa. A misfit is found when the hypothesis is not supported by the observed data. Examples of IRT-based person-fit statistics include Wright and Stone's (1979) $U$ statistic, Wright and Masters's (1982) $W$ statistic, Smith's (1985) $UB$ and $UW$ statistics, and Drasgow, Levine, and Williams's (1985) $l_z$ statistic. Although many statistics are available for examining the fit of a student's item response vector to a CTT or an IRT model, very few, if any, statistics are found in the literature that are designed specifically for testing the fit of student responses to CDMs. Existing person-fit statistics, which assume a unidimensional latent trait to interpret student performance, cannot be applied adequately in CDMs.

Over the past two decades, measurement specialists have shown growing interest in the development of CDMs with an attempt to incorporate cognitive models into test design and analysis so test results provide more diagnostic information about student thinking and learning. Designing test items according to a cognitive model has been recognized as an important way to improve the quality of test items and the validity of inferences drawn from tests (Borsboom, 2005; Embretson & Gorin, 2001; Leighton, 2004; Mislevy, 2006; Nichols, 1994). This is because a cognitive model provides an explicit understanding of the knowledge and skills normally used by students to solve standardized tasks in a test domain. By empirically analyzing student item responses to confirm the cognitive model used in test design, the validity argument about the targeted construct being measured by the test can be strengthened. Because cognitive models help clarify the psychology that underlies test

performance, compared to tests not developed and analyzed with a CDM, scores from tests using a CDM may be more interpretable and meaningful.

Although the advantages of CDMs in test design and analysis have been widely recognized, few, if any, person-fit statistics that have been found in the literature are explicitly designed to examine the fit of a student item response vector to CDMs. As the main feature of CDMs is to use a cognitive model to guide test design and analysis, the evaluation of person fit in CDMs should focus on assessing whether the set of knowledge and skills specified in the cognitive model can sufficiently account for student performance. The purpose of this study, then, is to (a) introduce a new person-fit statistic called the *hierarchy consistency index* (*HCI*) to test the fit of a student item response vector to a CDM used in test design and analysis, and (b) conduct a simulation study to assess the effectiveness or the power of the *HCI* in identifying misfitting item response vectors relative to CDMs.

## Cognitive Diagnostic Models: An Overview

The increasing desire to combine cognitive psychology with educational assessment provides a great incentive for developing and applying CDMs in test design and analysis. From a psychometric modeling perspective, CDMs evaluate student performance by establishing the relations of student response data to student mastery of different knowledge and skills within a particular test domain. In CDMs, the probability of a student's correct response to a test item is modeled as a function of item parameters in terms of the knowledge and skills measured by the item as well as student ability parameters characterized by the student's mastery of these knowledge and skills. The end goal of CDMs is to estimate students' mastery of knowledge and skills accurately and, hopefully, provide useful diagnostic information about student strengths and weaknesses for remediation purposes. In the past three decades, educational specialists have witnessed the emergence and development of many CDMs, such as Fischer's (1973) linear logistic trait model (LLTM), Embretson's (1984) multicomponent latent trait model (MLTM), Tatsuoka's (1983) rule space model, Hartz's (2002) reparameterized unified model (RUM), Mislevy, Steinberg, and Almond's (2003) Bayes net approach, and Leighton, Gierl, and Hunka's (2004) attribute hierarchy method (AHM). These CDMs differ primarily in the way cognitive models are formed and operationalized in the test domain and the way student proficiencies are formulated conceptually and estimated statistically. Given that the HCI was initially developed under the AHM framework, in the next section we provide a summary review of this method.

### *The Attribute Hierarchy Method (AHM)*

The AHM (Leighton et al., 2004), as an extension of Tatsuoka's rule space model, is a cognitive diagnostic tool designed to explicitly model the knowledge and skills underlying student problem solving on test items. The AHM is based on the assumption that test items can be described by a set of hierarchically ordered attributes. Attributes are defined as basic knowledge, skills, or cognitive processes. The assumption of attribute dependency is consistent with findings from cognitive research (e.g., Kuhn, 2001; Vosniadou & Brewer, 1992), where cognitive skills have been

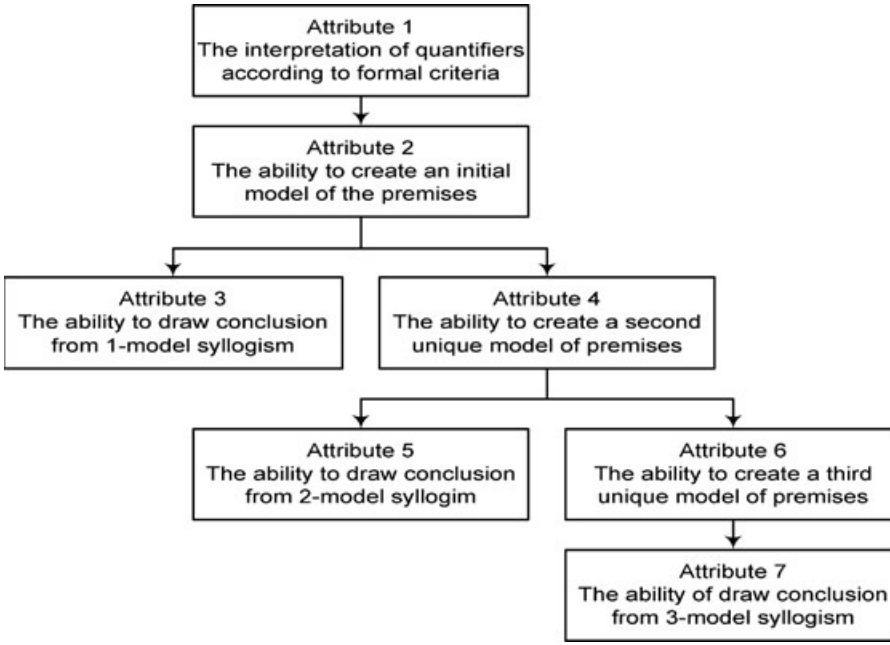


FIGURE 1. *A seven-attribute hierarchy of categorical syllogism performance (taken from Leighton et al., 2004).*

found not to operate in isolation but belong to a network of interrelated competencies. For example, as demonstrated by Leighton et al., in order to correctly answer an item that involves solving categorical syllogisms, certain attributes are required such as (a) understanding the meaning of quantifiers, (b) creating single or multiple representations of the quantified premises, and (c) drawing conclusions from representations. As shown in Figure 1, these attributes can be ordered into a hierarchy from basic to complex based on the expectation that complex attributes (e.g., drawing conclusions from representations) should not be possessed unless basic attributes (e.g., understanding the meaning of quantifiers) have been mastered. The AHM, by incorporating the assumption of attribute dependency, brings an important cognitive property into cognitive modeling methodologies.

The AHM is composed of three sequential stages. In the first stage, an attribute hierarchy is defined to describe the knowledge and skills that students are expected to use in the test domain. Once defined, the attribute hierarchy serves as a cognitive model that facilitates the prediction and explanation of student performance. 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 in a test domain. Once identified, the attributes need to be organized into a hierarchy. 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). This is a critical step because the validity of the attribute hierarchy underwrites the accuracy of the inferences to be made about student test-based performance. In the second stage, the attribute hierarchy is used as a basis for developing test items to ensure that each attribute has been measured with adequate reliability. This can be achieved by designing a sufficiently large number of items to measure each attribute. However, practical constraints such as time limit and student fatigue must be considered in determining the length of a test. In the third stage, statistical procedures are used to estimate each student's mastery of attributes in the hierarchy, thereby making specific inferences about students' strengths and weaknesses. For a full description of the technical details of the AHM, the reader is referred to Leighton et al. (2004).

## The Hierarchy Consistency Index

Given the assumption that student test performance can be described by a set of hierarchically related attributes, the AHM can be applied in test domains where students are expected to gain knowledge and skills gradually from simple to complex. In these domains, relatively simple knowledge and skills must be possessed in order for students to move to the next stage of learning in which more complex knowledge and skills are involved. For an educational test that is designed to measure a set of hierarchically ordered attributes, students are expected to answer correctly items that measure simple attributes if they have also produced correct answers to items requiring complex attributes. The logic of the new person-fit statistic HCI is to examine whether students' actual item response patterns match the expected response patterns based on the hierarchical relationship among attributes measured by test items.

The formulation of the HCI takes into account the item complexity as determined by the attribute hierarchy and the reduced Q matrix. The reduced Q matrix is an attribute-by-item matrix that describes the set of items that can be created to measure each possible combination of attributes in the hierarchy. For example, the reduced Q matrix for the attribute hierarchy presented in Figure 1 is shown below.

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

The reduced Q matrix contains 15 columns, each representing one possible combination of attributes that is consistent with the prerequisite relationships among the attributes in the hierarchy. Based on the reduced Q matrix, items can be created to measure each combination of attributes. For example, column 1 of this matrix shows that an item can be created to measure attribute 1 only. For a student to produce

a correct response to this item, the mastery of attribute 1 is required. Column 2 of this matrix indicates an item that requires both attributes 1 and 2. It should be noted that although coded as requiring both attributes 1 and 2, the item does not necessarily need to probe attribute 1 directly. For example, consider an item that measures whether students can create a unique representation of logical quantifiers (i.e., attribute 2). This item might not ask students to interpret certain quantifiers (i.e., attribute 1) directly. However, if a student produces a correct answer to the item, one can infer that the student can properly interpret quantifiers as it is the prerequisite for attribute 2 (the ability to create a unique representation of quantifiers). In this sense, the item is coded as requiring both attributes 1 and 2. The remaining columns can be interpreted in the same manner.

Note that the reduced Q matrix includes all possible attribute combinations that are consistent with the attribute hierarchy. They can be divided into two categories: simple and complex combinations. A combination can be considered as simple when all the required attributes belong to one branch of the attribute hierarchy. For example, the attribute hierarchy in Figure 1 contains three branches: the first branch has attributes 1 to 3; the second has attributes 1, 2, 4, and 5; and the third has attributes 1, 2, 4, 6, and 7. In the reduced Q matrix associated with this hierarchy, columns 1 to 4 and columns 6, 8, and 12 show attribute combinations from only one branch of the hierarchy (columns 1–3 measure branch 1; columns 4 and 6 measure branch 2; and columns 8 and 12 measure branch 3) and therefore can be categorized as simple attribute combinations. An important feature of simple attribute combinations is that all the required attributes are from the same branch and therefore share prerequisite relationships. Items can be created to explicitly measure only one attribute—the most complex attribute in the combination. If a student answers the item correctly, the mastery of all the required attributes can be assumed. From an incorrect response, one can infer that the student has not mastered at least the most complex attribute required by the item. For simple attribute combinations, therefore, the interpretation of item results is relatively straightforward.

In comparison, complex attribute combinations include attributes from at least two branches. As attributes from different branches may not share prerequisite relationships (e.g., attribute 3 is not prerequisite to attribute 4, and vice versa), the associated items must be created to explicitly measure at least two attributes from the different branches. These items tend to be more difficult to interpret in comparison to items measuring simple attribute combinations. For example, column 5 of the reduced Q matrix is (1111000), representing an item that requires attributes 1 to 4. Because attributes 3 and 4 belong to two different branches, column 5 represents a complex attribute combination. Given that no prerequisite relationship exists between attributes 3 and 4, items that require the attribute combination (1111000) must be created in the manner that measure both attributes 3 and 4 explicitly. If a student fails to produce correct answers to these items, one cannot tell clearly whether it is because the student has not mastered attribute 3, 4, or both. Therefore, the interpretation of these items tends to be problematic. In addition, items that measure complex attribute combinations may require more time to develop considering the complexity involved in the process of incorporating multiple attributes that may not have a direct hierarchical ordering into each item.

To avoid the difficulty and ambiguity caused by complex attribute combinations in item development and interpretation, test items can be created according to simple attribute combinations only. In this sense, the reduced Q matrix can be condensed to contain only simple attribute combinations:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

For each column of this matrix, items can be created to focus on measuring the most complex attribute of the column. For example, the first column of matrix 2 represents an item that explicitly measures attribute 1, and the last column represents an item that directly measures attribute 7. In this way, each attribute is explicitly measured by one item. To increase the reliability of the test, multiple items can be designed to measure each attribute. Suppose that a test is designed in which each attribute is explicitly measured four times. The reduced Q matrix for this test is:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{3}$$

The first four columns represent items in the test that are created to explicitly measure attribute 1, the next four columns represent items that explicitly measure attribute 2, and so on.

The reduced Q matrix serves to provide the test blueprint within a content domain, as it shows explicitly the attributes measured or required by each item. The AHM is conjunctive in nature in the sense that the mastery of all the required attributes is necessary for successful performance on a given test item. Therefore, a student is considered to have mastered all of the required attributes for an item when the item is correctly answered. In turn, the student is expected to correctly answer all those items that require the subset of attributes measured by the correctly answered item. For example, if a student answers an item requiring attributes 1 and 2 correctly, the student is also expected to have answered correctly all items that measure attribute

1. The HCI 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

$S_{correct_i}$ is an index set that includes items correctly answered by student $i$,
$X_{i_j}$ is student $i$'s score (1 or 0) to item $j$, where item $j$ belongs to $S_{correct_i}$,
$S_j$ is an index set that includes items requiring 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 student $i$'s item responses and the expected responses associated with the reduced Q matrix. When student $i$ correctly answers item $j$, $X_{i_j} = 1$, the student is also expected to answer item $g$ that belongs to $S_j$ correctly, namely, $X_{i_g} = 1$ ($g \in S_j$). If the student fails to answer item $g$ correctly, $X_{i_g} = 0$, 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, $N_{c_i}$, contains the total number of comparisons for items that are correctly answered by student $i$. When the numerator of the HCI is set to equal the total number of misfits multiplied by 2, the HCI has the property of ranging from $-1$ to $+1$, which makes it easy to interpret. When no misfit is found, the numerator of the HCI will be 0 and the HCI 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 correctly answered item), the numerator of the HCI will be equal to $2N_{c_i}$ and the HCI will be $-1$. If the HCI value of a student response vector is close to $-1$, one can conclude that either the attribute hierarchy fails to provide a valid representation of the student's problem solving processes or aberrant response behavior (e.g., creative responding or random responding) is occurring. In either of these cases, the AHM should not be used to make inferences about student performance.

One might consider the HCI to be similar to Tatsuoka and Tatsuoka's (1983) norm conformity index (NCI) in the sense that both statistics evaluate the misfit of an observed response vector by comparing the student's responses for each item pair with the expectations of the test model. However, a critical distinction between these two statistics is that the NCI bases the comparisons on item difficulty as determined by the proportion correct score of a group of students, whereas the HCI depends on item complexity as determined by the prerequisite relationship among test items specified in the reduced Q matrix. To illustrate the calculation of the HCI, consider the attribute hierarchy presented in Figure 1 and reduced Q matrix 3. Suppose a student's item response vector is (11110000100000000000000000), in which items 1 to 4, and 9

are correctly answered, namely $S_{correct_i} = \{1, 2, 3, 4, 9\}$ According to the reduced Q matrix, item 9 measures attributes 1, 2, and 3. Because the answer to item 9 is correct this student is considered to have mastered the attributes required by the item. Therefore, this student is also expected to 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, each of which measures the same set or a subset of attributes required by item 9. That is, $S_9 = \{1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12\}$. Therefore, for item 9, there are 11 comparisons: item 9 versus items 1 to 8 and 10 to 12. Because this student failed to answer items 5 to 8 and 10 to 12 correctly, seven misfits are found. Similarly, for items 1 to 4 that are also correctly answered by the student, $S_1 = \{2, 3, 4\}$, $S_2 = \{1, 3, 4\}$, $S_3 = \{1, 2, 4\}$ and $S_4 = \{1, 2, 3\}$. For item 1, there are three comparisons: item 1 versus items 2, 3, and 4. Because items 2, 3, and 4 are all correctly answered by the student, no misfit is 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, the value of the HCI for the student item response vector is $1 - (2 \times 7)/23 = 0.39$.

A researcher or practitioner who applies the HCI to detect misfitting item response vectors may want to know when a pattern can be classified as normal[1] and when it can be classified as misfitting. To achieve this, ideally, a sampling distribution should be derived for the HCI. However, because in the AHM no specific probabilistic model is assumed to underlie item responses, the probability of a student item response vector cannot be derived. For this reason, it is difficult to derive the sampling distribution for the HCI. In person-fit research, cut scores for distinguishing normal or misfitting patterns for nonparametric statistics are often selected using simulated item response patterns that display normal response behaviors (e.g., Meijer, Muijtjens, & van der Vleuten, 1996; Meijer & Sijtsma, 2001). If item response patterns are simulated based on the measurement model used for the test, the sampling distribution of the nonparametric person-fit statistic can be approximated by calculating the value of the statistic for each person, and then ordering these calculated statistics from the lowest to the highest value. If we let $\alpha = 0.05$ be the probability of misclassifying a normal item response pattern as misfitting, then the 95th percentile value of the statistic can be used as the cut score for classifying an item response pattern. This approach is used later in the simulation study to select cut scores for the HCI.

## A Simulation Study

### *Method*

*Research design*. To assess the effectiveness or the power of the HCI in identifying the misfit of student item response vectors to the AHM, we conducted a simulation study. We generated student response data under a variety of conditions expected to affect the power of the HCI. The three factors manipulated were type of misfit, number of items, and item discrimination power. The levels of each factor reflect those that might be found in a real testing situation.

Three types of misfit were simulated. The first type of misfit was related to the misspecification of the attribute hierarchy. As discussed earlier, the attribute hierarchy serves as a cognitive model in the AHM. If not specified accurately, the attribute hierarchy will fail to provide a valid representation of the knowledge and skills used

by students in answering test items. Consequently, misfits of student item response vectors to the AHM would be expected. The other two types of misfit manipulated in the simulation dealt with two forms of aberrant student response behaviors: creative responding and random responding. These behaviors tend to produce item response vectors that are not in concordance with the theoretical predictions or expectations of the AHM. For example, creative response behavior often causes a high-ability student to produce incorrect responses to some easy items that measure basic attributes, which, in turn, leads to the misfit of the student item response vector relative to the expectations of the attribute hierarchy. Another type of aberrant response behavior often considered in person-fit research, not included in this study, is cheating. Many power studies using simulated data have defined cheating as a low-ability student copying correct answers on difficult items from a high-ability neighbor. However, as pointed out by Meijer and Sijtsma (1995), at least two assumptions underlying the simulation of cheating item responses are not practically reasonable: first, the student only copies answers on difficult items; and second, the neighbor always produces correct answers. Rather, they argued that whether the student who copies extensively from a neighbor produces a misfitting response pattern depends, to a greater degree, on the neighbor's item response pattern. Therefore, only person-fit statistics specifically designed to detect cheating (e.g., Frary, 1993) are effective in this situation. For this reason we did not consider cheating in this simulation study.

We manipulated the number of items that measure each attribute to range from two to six in increments of two. It is generally desirable to have a sufficient number of items measuring each attribute so that reliable inferences about student mastery or nonmastery can be made. By including this factor in the simulation design, it may be possible to determine whether the power of the HCI is affected by the number of items.

Item discrimination power was also considered. Students who have mastered the attributes required by an item with high discriminating power are expected to have a high probability of responding to the item correctly, whereas students who have not are expected to have a low probability. As reported by Jang (2006), designing items with high discriminating power is critical in determining the success of the diagnosis. Items that have low discriminating power not only compromise the accuracy of the estimate of student attribute mastery but can also lead to the paradoxical situation where students classified as nonmasters appear to perform nearly as well as those classified as masters in terms of their total test scores (Roussos, Henson, & Jang, 2005). Two levels of item discriminating power were included in the simulation. High item discriminating power was reflected by a relatively large difference between masters and nonmasters in terms of their probabilities of producing correct responses to test items, and low item discriminating power was indicated by a relatively small probability difference.

In total, three levels of types of misfit, three levels of number of items, and two levels of item discriminating power were considered in the current study so as to produce a total of $3 \times 3 \times 2 = 18$ conditions. Each condition was replicated 100 times. The main criterion for evaluating the power of the HCI was the detection rate of the HCI (i.e., the proportion of the simulated misfitting item response vectors correctly classified as misfitting by the HCI).

*Data generation*. The attribute hierarchy presented in Figure 1 was used as a basis for the generation of item response vectors. First, data matrices consisting of 2,000 *normal* student item responses to 14, 28, or 42 items (2, 4, or 6 items per attribute given a total of seven attributes in the hierarchy) with high or low discriminating power were generated. Note that three test lengths crossed with two levels of item discriminating power produced a total of six tests. Normal student item responses to each test were used to find the cut scores for classifying an item response vector as normal or misfitting. These cut scores were used for the misfitting samples described in a subsequent section. To generate a student's item responses, his or her attribute pattern must first be specified. Given seven attributes shown in Figure 1, there are $2^7 = 128$ possible attribute patterns. However, of the 128 attribute patterns, only 16 are logically meaningful considering the prerequisite relationships among attributes illustrated in the hierarchy shown in Figure 1. The attribute patterns are as follows:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{4}$$

The first row of matrix 4 indicates a student who did not possess any of the attributes, the second row indicates a student who possessed attribute 1 but did not possess attributes 2 through 7, and the last row indicates a student who possessed all seven attributes. For each attribute pattern, an equal number of students were generated, namely, $2000/16 = 125$ students per attribute pattern. A student's item response to each item was simulated based on three parameters: the student's attribute pattern, the set of attributes required by the item, and the item's discriminating power. For an item of high discriminating power, the probability of producing a correct response was set at .9 if the student had mastered all the attributes required by the item

as indicated by the student's attribute pattern or .1 if the student had not mastered at least one attribute required by the item. For an item of low discriminating power, the corresponding probabilities were set at .6 and .2, respectively. For example, consider a student who possesses only attributes 1 and 2. For an item that requires only attribute 1 for a correct response, the student had a 90% chance to answer the item correctly if the item has high discriminating power or a 60% chance if the item has low discriminating power. However, for a high discriminating item that requires attributes 1 to 3 for a correct response, the student who does not possess attribute 3 still has a 10% chance to produce a correct answer by guessing or by applying partial knowledge. Alternatively, if the item is low discriminating, the student has an even higher chance (20%) to respond correctly. For each test, the HCI values were calculated for each of the 2,000 simulated item response vectors and ordered from lowest to highest value. In this study, the probability of misclassifying a normal item response vector as misfitting (Type I error rate) was set at .10, higher than the conventional .05 level, to control the probability of failing to identify a vector (Type II error rate). This was warranted because the failure to identify a misfitting vector can lead to unintended consequences: it can falsely validate the inferences drawn from the test and result in inaccurate instructional decisions for the student. Using $\alpha = .10$, the 90th percentile values of the HCI were chosen as the cut scores. Because 100 replications were conducted for each test, the mean of the 100 cut scores was used as the final cut score for the misfitting samples described next.

Second, data matrices consisting of 2,000 *misfitting* item response vectors were simulated. As discussed earlier, three types of misfitting vectors were simulated in the study. The first type of misfitting vector was created by the misspecification of the attribute hierarchy. To generate this type of misfit we assumed that the hierarchy in Figure 1 did not provide an accurate representation of the attribute relationships for all students. Hypothetically, some students, who lacked the ability to create unique models of premises (attributes 2, 4, and 6), might still be able to successfully draw conclusions from a 1-, 2-, or 3- model syllogism (attributes 3, 5, and 7). Their attribute patterns can be represented by (1010000), (1010100), or (1010101). Note that these attribute patterns were not included in the attribute matrix shown in Matrix 4 because they were unexpected given the assumption that the attribute hierarchy in Figure 1 is true. For these "misspecified" students, the true relationship among attributes is presented in the hierarchy in Figure 2, where attributes 2, 4, and 6 are not the prerequisite attributes for attributes 3, 5, and 7, respectively. Therefore, the first type of misfitting item response vector was simulated by generating item responses from students with attribute patterns (1010000), (1010100), or (1010101).

The design also included two other types of misfitting response vectors produced by creative responding and random responding. Creative item responses were simulated by generating item responses from high-ability students with attribute patterns associated with Figure 1: (1101011), (1111011), (1101111), and (1111111), and inputting incorrect responses to items that solely measured attribute 1 (i.e., relatively easy items). Random responding students were simulated by randomly generating responses to each item with a probability of .25 for a correct response, which corresponds to the probability of producing a correct response by blindly guessing to a multiple-choice item with four options. For each type of aberrant response behavior,

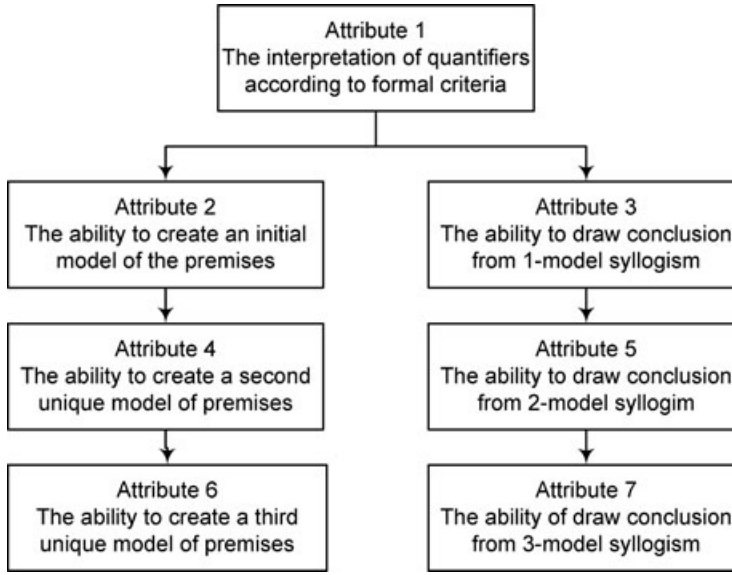


FIGURE 2. *A hypothetical seven-attribute hierarchy of categorical syllogism performance.*

100 data matrices, each consisting of 2,000 item response vectors, were generated for each of the six tests (i.e., 3 levels of test length $\times$ 2 levels of item discriminating power $=$ 6 tests). The HCI values for each simulated vector was calculated and compared to the corresponding cut score obtained in the normal samples to determine whether the vector was misfitting or normal. The detection rates of the HCI were obtained by counting the proportions of item response vectors in the misfitting samples that were successfully classified as misfitting by the HCI. They were used as indicators of the power of the HCI in detecting misfitting vectors. All methods described in this subsection were executed by a program written in *Mathematica* 6 code. The program is available from the first author upon request.

### *Results*

Table 1 shows the cut scores of the HCI identified for classifying a student item response vector as normal or misfitting. These cut scores were obtained using the normal samples generated from the attribute hierarchy in Figure 1. The differences among the cut scores or tests of high discriminating items may be considered negligible. In comparison, for tests of low discriminating items, the cut scores increased in magnitude as more items were included in the test. The cut scores were used to identify misfitting item response vectors (e.g., for the test of 14-high discriminating items an item response vector with an HCI value below .05 would be classified as misfitting and one with an HCI value above .05 would be classified as normal).

Table 2 shows the mean percentages of item response vectors that were correctly identified as misfitting by the HCI, with respect to test length, item discriminating

TABLE 1
*Cut Scores for the HCI as a Function of Test Length and Item Discriminating Power*

| | Test Length | | |
|---|---|---|---|
| | 14 Items | 28 Items | 42 Items |
| High discriminating | .05 | .06 | .07 |
| Low discriminating | −.60 | −.43 | −.36 |

TABLE 2
*Percentage of Misfitting Item Response Vectors Correctly Identified by the HCI*

| | | Test Length | | |
|---|---|---|---|---|
| | | 14 Items | 28 Items | 42 Items |
| High discriminating items | 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 | 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) |

power, and the type of misfit. For each condition, the *SD* of the percentages across the 100 data sets was calculated and is presented in parentheses. A three-way ANOVA could be used to determine the statistical significance of the observed difference. However, the large sample size ($N = 1{,}800$) and small within-group variances can lead to almost every effect being labeled as significant. That is, any slight association or variation, which may not have any practical importance, can be detected from statistical analysis. Therefore, the simulation results were only interpreted descriptively.

For tests associated with high discriminating items, the detection rates of the HCI were moderate for misfits caused by model misspecification (detection rates between 50.89% and 52.66%). Although the detection rates appeared to decrease as more items were included in the test, the differences by test length were almost negligible (with a maximum of 1.77). In comparison, the HCI showed high detection rates for misfitting item response vectors produced from both creative and random responses. The detection rates varied between 91.14% and 99.06% for creative vectors and between 93.01% and 99.82% for random vectors. Note that the detection rates consistently increased as test length increased from 14 to 42. Therefore, the number of items showed a positive impact on the detection rates of the HCI for creative and random item response vectors.

For tests associated with low discriminating items, the HCI showed relatively low detection rates for misfits caused by model misspecification. The values varied between 20.90% and 25.18%, and appeared to increase as test length increased. In comparison, the HCI showed high detection rates for creative item response vectors, values varying from 88.18% to 99.97%. For random vectors, the HCI produced

moderate to high detection rates. By increasing test length from 14 to 42 items, the detection rates increased considerably, from 53.01% to 82.38%.

The comparison for tests associated with high and low discriminating items revealed a significant impact on the detection rates of the HCI, particularly for identifying misfits caused by model misspecification and random responding. For example, given a 14-item test, 93.01% of random item response vectors were correctly identified as misfitting by the HCI when items with high discriminating power were used, whereas the detection rate dropped to 53.01% when items with low discriminating power were used.

### *Discussion*

*Item discriminating power*. The factor of item discriminating power showed a significant effect on both the cut scores and the detection rates of the HCI. The inspection of cut scores for tests of equal length (Table 1), but of different levels of discriminating power, revealed that as the discriminating power of test items decreased, the cut scores declined sharply (from 0.05 to $-0.60$, from 0.06 to $-0.43$, and from 0.07 to $-0.36$ for 14, 28 and 42-item tests, respectively). All the cut scores for tests of low discriminating power were negative, close to the lower bound of the HCI ($-1$). Considering that the cut scores were identified from the HCI values in the normal samples, one might ask why these item response vectors had such low HCI values. One plausible answer to this question is that low discriminating items might have led to inconsistencies in the student item response vectors in relation to the expectations of the AHM. For example, the mastery of the set of attributes required by an item does not always lead to a correct response. In fact, when the item does not discriminate well between masters and nonmasters, students who have mastered the required attributes have a fairly high chance of reaching incorrect answers (e.g., $1-60\% = 40\%$ in this study). Conversely, students who have not mastered the attributes required by an item may respond correctly. In fact, when items have low discriminating power, students who have not mastered the attributes have a relatively high chance of reaching correct answers (e.g., 20% in this study). When items expected to be answered correctly are indeed answered incorrectly, or items expected to be answered incorrectly are indeed answered correctly, inconsistencies are observed in the student's item responses. Hence, although normal vectors were used, low cut scores of the HCI were identified when items with low discriminating power were assumed in the generation of item response vectors.

Item discriminating power was also found to have a positive effect on the detection rates of the HCI after controlling for type of misfit and test length, which is in agreement with the findings from many person-fit studies (e.g., Meijer, Molenaar, & Sijtsma, 1994). As item discriminating power decreased, the power of the HCI decreased considerably in detecting misfits caused by model misspecification and random responding. This is understandable because, as discussed earlier, low item discriminating power can lead to inconsistencies in student item responses. When confounded with this type of inconsistency, misfits caused by model misspecification and random responding become difficult to identify. One cannot tell whether low HCI values are due to low item discriminating power, model misspecification,

or both. Therefore, the detection rates were relatively low when a test consisted of low discriminating items. However, for creative responding vectors, the effect of item discriminating power on the detection rate of the HCI was very small. The maximum difference in the detection rates between tests of high and low discriminating items was only 2.96%. This suggests that the HCI is relatively robust in detecting item response vectors reflecting creative responses even when low discriminating items make up the test.

*Types of misfit*. Three types of misfit were considered: model misspecification, creative responding, and random responding. The detection rates of the HCI were found to vary with different types of misfit. The simulation results indicated that the HCI produced moderate to high detection rates in identifying creative and random responding vectors but relatively low to moderate rates for misfits caused by model misspecification. It is worth considering why the detection rates varied with different types of misfit. To answer this question, the methods for generating different types of misfit must be considered. Misfits due to model misspecification were simulated first by generating item responses from the hypothetical students who lacked the ability to create unique models of premises (attributes 2, 4, and 6) but could still successfully draw conclusions from a 1-, 2-, or 3- model syllogism (attributes 3, 5, and 7). For these hypothetical students, the attribute hierarchy in Figure 1 failed to accurately specify the relationship among attributes. The hierarchy that reflected the true relationship among attributes for these hypothetical students is presented in Figure 2. The comparison of the two hierarchies revealed that although hierarchy 1 misspecified the prerequisite relationships between attributes 2 and 3, 4 and 5, and 6 and 7, it correctly identified the prerequisite relationships among attributes 1, 2, 4, and 6, and prerequisite relationships among attributes 1, 3, 5, and 7. In this sense, hierarchy 1 only *partially* misspecified the relationship among attributes. As a result, the generated misfitting item response vectors reflected a moderate level of misfit and were relatively difficult to identify, which led to low to moderate detection rates of the HCI (ranging from 20.90% to 25.18% for low discriminating tests and from 50.89% to 52.66% for high discriminating tests). In comparison, creative item response vectors were simulated by generating item responses from high-ability students with attribute patterns derived from the attribute hierarchy shown in Figure 1, and inputting incorrect responses to easy items that required only attribute 1. The generated item response vectors consisted of incorrect responses on the easiest items but correct responses on the most difficult items, which reflected comparatively strong misfit. Therefore, the HCI yielded high detection rates, ranging from 88.18% to 99.97% for low discriminating tests and from 91.14% to 99.06% for high discriminating tests. Random responding students were simulated by randomly generating responses to each item with a 25% chance of producing a correct response. In this way, the attribute prerequisite relationships did not play any role in generating item response vectors. As a result, strong misfits existed between the generated and the expected vectors, and moderate to high detection rates were yielded by the HCI (ranging from 53.01% to 82.38% for low discriminating tests and from 93.01% to 99.82% for high discriminating tests).

*Number of items*. Results from the simulation study showed that as the number of items increased, the HCI tended to provide higher detection rates, after controlling for type of misfit and item discriminating power. This is in agreement with the conclusions from many person-fit studies where test length was found to have a positive impact on the power of person-fit statistics in identifying misfitting item response vectors (e.g., Karabatsos, 2003; Meijer et al., 1996; Reise & Due, 1991). The influence of test length on the detection rates of the HCI can be addressed by the fact that the stochastic nature of response behaviors almost certainly introduces measurement errors in student item response vectors. For instance, students can answer an item successfully even when they do not master the set of attributes measured by the item. Student responses to a single test item cannot inform us whether the model-specified knowledge and skills are used in solving the item, or whether chance or some other aberrant behaviors are involved. However, as the test includes more items, it is much less likely that by chance the student can answer correctly all items that measure the same set of attributes by chance. Therefore, by lengthening the test, uncertainties due to the stochastic nature of response behavior tends to be weakened. In turn, systematic misfits (such as those caused by creative responding) become apparent and thereby relatively easy to identify.

## Conclusions

The logic of the HCI is that a student should not be able to answer an item correctly unless that student has solved its prerequisite items successfully. The evaluation of the misfit of an item response vector relative to the cognitive model is focused on assessing whether students' actual item response patterns match the expected response patterns based on the prerequisite relationships among test items. Although statistically sound, the simulation procedure used in this study may appear unrealistic to some practitioners. Alternative procedures for setting cut scores also could be considered. For example, in a study that investigated the usefulness of person-fit methodology for personality assessment, Meijer, Egberink, Emons, and Sijtsma (2008) found the cut score for the nonparametric person-fit statistic, $Z_{GE}$, based on its empirical distribution from the sample. The idea was that response vectors associated with extreme person-fit values or outliers of the sample distribution could be considered as misfitting. A rule of thumb for identifying outliers from a sample distribution is to find data points that deviate from the first and the third quartile by 1.5 times the difference between the two values (Tukey, 1977). The benefit of using descriptive techniques like this lies in their relative simplicity in using and understanding the cut scores when applied in practice. Further research is needed to investigate and compare the effectiveness of different techniques for identifying misfitting item response vectors.

Although the HCI is designed specifically to identify misfits at the individual student level, the mean of the HCI can serve as an indicator of overall model-data fit. When a large proportion of student item response vectors show low HCI values, it is likely that either the model does not provide a valid representation of student knowledge and skills, or test items fail to discriminate well between students who have and

have not mastered the attributes that the items are designed to measure. Although developed within the AHM framework, the HCI should be helpful with other CDMs that are guided by cognitive models given that the index allows the researcher to evaluate the fit of student item response vectors relative to the expectations of a cognitive model. Particularly, the HCI should be useful for other Q matrix-based conjunctive CDMs where a Q matrix is typically specified to indicate the attributes required by each item and the assumption is that the mastery of all the required attributes is necessary for successful performance on test items. Examples of Q matrix-based conjunctive CDMs are the rule space model (Tatsuoka, 1983, 1984, 1990, 1995), the unified model (DiBello, Stout, & Roussos, 1995), the deterministic input noisy and gate model (DINA; de la Torre & Douglas, 2004; Diognon & Falmagne, 1999; Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977; Tatsuoka, 2002), and the noisy input deterministic and gate model (NIDA; Junker & Sijtsma, 2001). In these models, the HCI can be directly used to compare student item responses to the expectations from the Q matrix so as to identify students whose item responses are the result of other variables aside from the mastery or nonmastery of the knowledge and skills hypothesized in the cognitive model. It should be noted that the HCI cannot be used with compensatory CDMs where the mastery of all the attributes measured by an item is not necessary for successful performance because of the assumption that high ability on one attribute can compensate for low ability on other attributes. Although the applications of skill diagnosis have focused more on domains where conjunctive CDMs are more appropriate (e.g., mathematics), the use of compensatory CDMs will likely increase as skill diagnosis is applied to a greater variety of content domains. Therefore, research is needed for methods that are specifically designed to identify misfits of item response vectors relative to compensatory CDMs.

The result that the best detection rates were achieved when the HCI was applied to tests that consisted of a relatively large number of high discriminating items suggests that the use of a sufficiently large number of such items to measure each attribute is critical for reliable skill diagnosis. It not only has the potential to improve the measurement reliability and accuracy of student ability estimates but also allows misfitting response behavior to be more identifiable. In addition, whether a misfitting item response vector can be correctly identified depends, to a large degree, on the number of misfits of the vector relative to the cognitive model. When misfitting response behavior only affects a small number of item responses, the resulting vector will not be substantially different from the expectations under the cognitive model and consequently may not be statistically identified as misfitting. As an item response vector deviates further from model expectations, misfits are more easily identified and consequently higher detection rates of the HCI are expected. A limitation of this simulation study is that the discriminating power was set constant across items, which is not expected to hold in actual testing applications. An area of future research is to investigate the power of the HCI in detecting misfitting vectors for tests that consist of items with various levels of discriminating power. In addition, this study only investigated the effects of type of misfit, number of items, and item discrimination on the detection rate of the HCI. Additional studies are needed to consider other factors (e.g., complexity of attribute hierarchy) that might affect the performance of the HCI.

To conclude, the use of person-fit statistics to identify misfitting item response vectors is only the first step in person-fit analysis as it does not provide clear indications of how misfits occur or what types of misfitting response behavior underlie test performance. To find the actual causes of misfits, additional information about students' response processes, such as students' verbal reports, eye tracking information, and reaction time (American Educational Research Association, National Council for Measurement in Education, & American Psychological Association, 1999) is needed. This type of information provides relatively detailed pictures of how students actually solve items on tests, which has the potential to help understand the reasons for misfits so that the results from person-fit statistics can be interpreted substantially and meaningfully.

## Note

[1]An item response pattern is classified as normal when it is not associated with aberrant response behaviors but is produced from model-specified knowledge and skills.

## References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, UK: Cambridge University Press.

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

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

Diognon, J. P., & Falmagne, J. C. (1999). *Knowledge spaces*. New York: 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., & 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., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*, 343–368.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 395–374.

Frary, R. B. (1993). Statistical detection of multiple-choice answer copying: Review and commentary. *Applied Measurement in Education*, *6*, 153–165.

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 Measurement*, *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, Champaign.

Jang, E. E. (2006). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Unpublished doctoral dissertation, University of Illinois, Champaign.

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.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*, 277–298.

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: 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*(2), 6–15.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205–237.

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

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

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., Egberink, I. J., Emons, W. H., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment*, *90*, 227–238.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, *18*, 111–120.

Meijer, R. R., Muijtjens, A. M. M., & van der Vleuten, C. P. M. (1996). Nonparametric person-fit research: Some theoretical issues and an empirical example. *Applied Measurement in Education*, *9*, 77–89.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent 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. (2006). Cognitive psychology and educational assessment. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 257–305). Westport, CT: Praeger and American Council on Education.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspective*, *1*, 3–67.

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

Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement*, *15*, 217–226.

Roussos, L. A., Henson, R. A., & Jang, E. E. (2005). *Simulation study evaluation of the fusion model system stepwise algorithm*. ETS Project Report. Princeton, NJ: Educational Testing Service.

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.

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

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

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

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

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.

## Authors

YING CUI is an Assistant Professor, Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB, Canada, T6G 2G5; yc@ualberta.ca. Her primary research interests include educational measurement and evaluation, cognitive diagnostic assessment, research design, and applied statistical methods.

JACQUELINE P. LEIGHTON is a Professor, Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB, Canada, T6G 2G5; jacqueline.leighton@ualberta.ca. Her primary research interests include educational assessment and evaluation, the development of high-level thinking tasks and items, construct validation, and statistics.