

The Hierarchy Consistency Index:

A Person-fit Statistic for the Attribute Hierarchy Method

Ying Cui

Jacqueline P. Leighton

Mark J. Gierl

Steve M. Hunka

Centre for Research in Applied Measurement and Evaluation (CRAME)

University of Alberta

Abstract

The attribute hierarchy method (AHM) (Leighton, Gierl, & Hunka, 2004), which is based on the assumption that test items can be described by a set of hierarchically ordered attributes, is designed to estimate examinees' level of competency as well as their profiles for mastering a set of attributes. The AHM, by incorporating the assumption of attribute dependency, brings an important cognitive property into cognitive diagnostic testing. However, the validity of this new diagnostic model depends critically on the accuracy and adequacy of the attribute hierarchy. In this study, a person-fit statistic, called the hierarchy consistency index (HCI_i), is introduced to help assess the degree to which an observed examinee response pattern is consistent with the attribute hierarchy of the AHM. This type of model-data fit study can enhance the validity of diagnostic feedback produced by the AHM.

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

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

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

These new assessments make explicit the test developer's substantive assumptions regarding processes and knowledge structures a performer in a test domain would

use, how the processes and knowledge structures develop, and how more competent performers differ from less competent performers. (p. 578)

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

Cognitive diagnostic models (CDMs) have been developed to help construct diagnostic assessments and estimate students' attribute mastery patterns associated with different cognitive skills. Leighton, Gierl, and Hunka (2004; see also Gierl, Leighton, & Hunka, 2000) proposed a CDM called the Attribute Hierarchy Method (AHM). The AHM is based on the assumption that test items can be described by a hierarchically-ordered set of attributes. Attributes are defined as basic cognitive processes or skills required to solve test items correctly (Leighton et al., 2004). The attribute hierarchy can be used as a basis for the development of test items, which upon the administration to examinees, produces vectors of binary responses (1 or 0). An examinee's response vector is then used to determine the examinee's level of competency as well as the likelihood that the attributes are possessed by the examinee.

For the AHM to yield inferences about students, the attribute hierarchy in the domain of interest must be specified correctly. Hence, methods for assessing the accuracy and adequacy of the attribute hierarchy in describing the cognitive processes used by

students to solve test items must be developed. Generally, methods for evaluating the misfit of an item-score vector to a specific test model have been referred to as "person-fit" methods. Although numerous person-fit statistics have been proposed and investigated by researchers (Meijer & Sijtsma, 2001), as will be discussed shortly, most of these methods cannot be used to determine if the attribute hierarchy is truly representing the cognitive processes used by examinees to solve test items. Therefore, it is inappropriate to use these existing person-fit statistics with the AHM.

The purpose of the present study is to introduce a person-fit statistic called the hierarchy consistency index (HCI_i), which is designed explicitly to examine the degree to which an observed examinee response vector is consistent with the attribute hierarchy. The present paper is divided into five sections. The first section presents a brief overview of the AHM. The second section discusses the applicability of the existing person-fit statistics in the AHM framework and introduces the person-fit statistic, the HCI_i , designed to assess whether the examinee uses different cognitive skills (or in a different combination) from what the attribute hierarchy indicates when solving test items. The third section introduces a simulation approach for setting the critical value of the HCI_i . By comparing an observed HCI_i against the critical value of the HCI_i , researchers can determine whether the observed response pattern are consistent with the attribute hierarchy of the AHM. The fourth section applies the AHM and the HCI_i to a real data set obtained from a previously administered large-scale achievement test for illustrative purposes. The fifth and final section summarizes and discusses two lines of future research related to the HCI_i and its applicability to other cognitive diagnostic models.

An Overview of the Attribute Hierarchy Method

The AHM is a cognitive diagnostic model designed to estimate examinees' level of competency as well as profiles that reflect their mastery for a set of attributes. 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 cognitive processes or skills required to solve test items correctly (Leighton et al., 2004). The first step in using the AHM for cognitive diagnosis is to identify the attributes in the domain or for the task of interest. Correctly identifying the attributes influences the validity of the inferences about examinees made with the AHM. As mentioned by Leighton et al. (2004), methods from cognitive psychology, such as task and protocol analysis, could play an important role in the identification of attributes in a domain. Many studies have been conducted to identify the attributes required for successful performance on test items and tasks. For example, in a language testing study, Buck and Tatsuoka (1998) created the hypothetical attribute set for a 35-item listening comprehension test by using two main sources: an extensive literature review to seek the theoretical and empirical evidence for the attributes that affect performance on listening tests and the results from a series of verbal protocol studies conducted by Buck (1990, 1991, 1994) for examining the second language listening processes.

In the AHM, attributes are considered to be hierarchically related and therefore can be ordered into a hierarchy based upon their logical and/or psychological properties. As explained by Leighton et al. (2004), the assumption of attribute dependency is consistent with the conclusion that “cognitive skills do not operate in isolation but belong to a network of interrelated competencies (Kuhn, 2001; Vosniadou & Brewer, 1992)” (p.

209). The ordering of the attributes into a hierarchy should be based on “empirical considerations (e.g., a series of well defined, ordered cognitive steps identified via protocol analysis) or theoretical considerations (e.g., a series of developmental sequences suggested by Piaget such as preoperational, concrete operational, and formal operational)” (Leighton et al., 2004, p. 209).

Once the attribute hierarchy is identified, a series of matrices (e.g., the adjacency, reachability, incidence, reduced incidence, and expected response matrices), initially introduced by Tatsuoka (1983, 1995, 1996), can be derived to facilitate the development of test items and the estimation of students’ profiles of their mastery and non-mastery of cognitive skills. A binary adjacency matrix (A) of order $K \times K$ specifies the direct relationship between each pair of attributes, where K is the number of attributes. To specify the direct and indirect relationship among attributes, a reachability matrix (R) of order $K \times K$ is used. The R matrix can be obtained using the equation $R = (A + I)^n$, where I is an identity matrix of order $K \times K$, and n is the integer between 1 and K that leads R to become invariant. That is, if $(A + I)$ times itself repeatedly using Boolean algebra until the product become invariant, then the obtained matrix is the R matrix.

The potential item pool, represented by an incidence matrix (Q) of order $K \times (2^K - 1)$, contains items that measure all the possible combinations of attributes when the attributes are assumed to be independent of each other. The columns of the Q matrix are obtained by converting the integers ranging from 1 to $2^K - 1$ to their binary form. In the Q matrix, each column represents one item, and the 1s in the column identify which attributes are required for successful performance on this item. When the attributes share dependencies, the size of the potential item pool can be significantly reduced by imposing

the constraints of the attribute hierarchy as embodied in the R matrix. The removal of items that do not match the constraints of the R matrix ultimately produces a reduced incidence matrix (Q_r). The Q_r matrix can be derived by determining which columns of the Q_r matrix are logically included in each column of the R matrix using Boolean addition. By describing the cognitive requirements of the domain of interest with the attribute hierarchy and specifying items needed to measure the domain in the Q_r matrix, the AHM attempts to make a direct link between student cognition and test design.

Once the R matrix and the Q_r matrix are identified, the expected response matrix (E) can be created. The E matrix is composed of the response patterns that can be clearly explained by the presence or absence of the attributes without any errors or “slips” given that the attribute hierarchy is true. The rows of the E matrix represent the expected examinees who possess cognitive attributes that are consistent with the hierarchy. *Expected* examinees do not make errors or slips that produce inconsistencies between the observed and expected response patterns.

In real testing situations, it is possible that an examinee, who has not mastered all the attributes required by an item, can still answer the item correctly by guessing or by having partial knowledge. It is also possible that an examinee who has mastered all the attributes that an item is probing might answer the item incorrectly due to careless mistakes. Therefore, the *observed* examinee response vectors might reflect slips of the form 1 to 0 or 0 to 1. By classifying each observed response vector into one of the expected response patterns, examinees’ attribute mastery can be estimated.

Leighton et al. (2004) proposed two methods for the classification of observed response patterns in the AHM. In these two methods, the probability of a correct response

to individual items is calculated for each expected response pattern using an IRT model.

The three-parameter logistic IRT model is given by:

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

where a_j is the item discrimination parameter for item j , b_j is the item difficulty parameter for item j , c_j is the pseudo-guessing parameter for item j , and θ_i is the ability parameter for examinee i . The two-parameter logistic IRT model is a special case of the three-parameter model in which the c_j parameter is set to 0. The one-parameter model also called Rasch model is another form of the logistic IRT model in which all the items are assumed to have equal discrimination power and no guessing. Item parameters can be estimated based on the expected response patterns using BILOG 3.11 (Mislevy & Bock, 1990).

Once item parameters and the theta value associated with each expected response pattern are estimated, the IRT probability of a correct response for each item can be calculated for each expected response pattern. In Method A, an observed response pattern is compared against each of the expected response patterns to identify the slips from 1 to 0 and from 0 to 1. The likelihood of all slips from 1 to 0 and from 0 to 1 for examinee i is given by:

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

where S_{i0} is the subset of items with slips from 0 to 1 for the observed response vector of examinee i , and S_{i1} is the subset of items with slips from 1 to 0. The higher the value of $P_{ijExpected}(\theta_j)$ calculated by comparing the observed response vector to one of the expected

response vectors, the more likely the observed response pattern originates from that expected response vector. Therefore, the observed response vector will be classified as originating from expected response vector j when the maximum value of $P_{ijExpected}(\theta_j)$ is achieved.

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

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

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

Another classification approach currently being investigated is to use neural networks to classify one examinee's observed response pattern into a single examinee attribute pattern. One benefit of using neural networks is that they do not rely on IRT models and assumptions about the distribution of examinees. Rather, they can be used to estimate the probabilities that examinees have mastered each attribute by minimizing the error associated with the estimation.

A Person-fit Statistic for the AHM

The AHM brings a fundamentally important cognitive feature into cognitive diagnostic models by incorporating the assumption of attribute dependency. In order for the AHM to be used to make valid inferences about students, however, it is important to correctly identify the attribute hierarchy in the domain of interest. For example, if the cognitive attributes summarized in the attribute hierarchy do not correspond to any real aspects of the cognitive processes measured within each student, then any diagnoses of the student based on the attribute hierarchy will be meaningless. In addition, the inclusion of superfluous attributes in the attribute hierarchy may lead to a high misclassification rate due to the unnecessarily high complexity of the model in terms of the large number of deceptive knowledge states around the true states (Tatsuoka & Ferguson, 1999). Furthermore, a model that fails to include some of the important attributes will not provide sufficient diagnostic information to permit test users to develop and implement interventions designed to maintain students' strengths and addresses students' weaknesses.

In short, it is important for the attribute hierarchy to be supported by psychological evidence that demonstrates students' problem-solving behavior has been measured. To date, such evidence is limited. Consequently, methods for assessing the accuracy and adequacy of the attribute hierarchy in describing the cognitive processes used by students to solve test items must be developed. One method for doing this is to employ a person-fit statistic. Numerous person-fit statistics based on CTT and IRT have been proposed and investigated (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).

The CTT-based statistics rely on item difficulty as determined by the proportion-correct score of a group of examinees. In the CTT-based statistics, items are ordered and numbered according to a decreasing proportion-correct score (increasing item difficulty). If an examinee's number-correct score is n , the examinee is expected to answer the first n easiest items correctly. Therefore, a response vector is considered as misfitting when items with relatively low proportion-correct scores are answered correctly, and items with relatively high proportion-correct scores are answered incorrectly.

In general, the IRT-based statistics compare the observed item responses with the IRT probabilities calculated using the estimate of the examinee's overall ability. A response vector is considered as misfitting when items with relatively high IRT correct response probabilities are answered incorrectly, and items with relatively low IRT correct response probabilities are answered correctly.

In the AHM, items are described by a set of hierarchically ordered attributes. In turn, the evaluation of the misfit of observed item responses to the AHM should be focused on examining if the Q_r matrix derived from the attribute hierarchy is truly representing the cognitive processes used by examinees to solve test items. Thus, it is inadequate to use the CTT-based statistics by only focusing on the item difficulty parameter or to use the IRT-based statistics by only focusing on the single estimate on the student's overall ability to evaluate if the examinee's response vector fits the model.

The Hierarchy Consistency Index

The proposed person-fit statistic, the HCI_i , depends on item complexity as determined by the attribute hierarchy and its associated Q_r matrix. In the AHM, the Q_r

matrix is used to describe the cognitive skills required in order for examinees to solve each item correctly. Therefore, by comparing an observed examinee response vector to the Q_r matrix, the HCI_i can be used to assess whether the examinee uses different cognitive skills (or in a different combination) from what the Q_r matrix indicate when solving test items. To calculate the HCI_i , therefore, the Q_r matrix needs to be specified. When the attribute hierarchy is used as a cognitive model for test development, the Q_r matrix can be derived from the attribute hierarchy to guide the construction of test items. When test items are not developed based on the attribute hierarchy, the Q_r matrix can be obtained by reviewing test items and identifying the attributes required by each item.

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

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

where

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

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

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

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

examinee i .

The term $\sum_{j \in S_{correct_i}} \sum_{g \in S_j} X_{i_j} (1 - X_{i_g})$ in the numerator of the HCI_i represents the number of misfits between examinee i 's item response vector and the Q_r matrix. If examinee i correctly answers item j , $X_{i_j} = 1$, then the examinee is expected to also correctly answer item g that belongs to S_j , namely, $X_{i_g} = 1$ ($g \in S_j$). If the examinee fails to correctly answer item g , $X_{i_g} = 0$, then $X_{i_j} (1 - X_{i_g}) = 1$ and it is a misfit of the response vector i to the Q_r matrix. Thus, $\sum_{j \in S_{correct_i}} \sum_{g \in S_j} X_{i_j} (1 - X_{i_g})$ is equal to the total number of misfits. The denominator of the HCI_i , N_{c_i} , contains the total number of comparisons for items that are correctly answered by examinee i . When the numerator of the HCI_i is set to equal the total number of misfits multiplied by 2, the HCI_i has the property of ranging from -1 to +1, which makes it easy to interpret.

To illustrate the calculation of the HCI_i , consider the attribute hierarchy shown in Figure 1. This hierarchy is also used by Leighton et al. (2004). The adjacency, reachability, and incidence matrices for this attribute hierarchy are presented in Leighton, et al. (2004) in matrices 1 to 3, respectively. The Q_r matrix associated with the attribute hierarchy shown in figure 1 is as follows:

$$Q_r = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Next, consider the observed response vector (1 1 0 0 0 1 0 0 0 0 0 0 0) where items 1, 2, and 7 are correctly answered, namely $S_{correct_i} = \{1, 2, 7\}$. According to column 7 of the Q_r matrix, item 7 is measuring attributes 1, 4, and 5. Since examinee i correctly answers item 7, he or she is considered to have mastered the attributes required by this item.

Therefore, examinee i is expected to also answer items 1 and 4 correctly, which are measuring the subset of attributes required by item 7. That is, $S_7 = \{1, 4\}$. Therefore, for item 7, there are two comparisons, item 5 vs. 1 and 4, respectively. Since examinee i fails to answer item 4 correctly, $X_{i_7}(1 - X_{i_4}) = 1$, a misfit between the examinee's response vector and the Q_r matrix is found. In the same manner, for items 1 and 2 that are also correctly answered by examinee i , $S_2 = \{1\}$ and $S_1 = \{ \}$, and no misfit is found. Overall, the total number of misfits is 1, and the total number of comparisons is equal to $2 + 1 + 0$

$= 3$. Hence, $HCI_i = 1 - \frac{2 \times 1}{3} = 0.33$.

When an examinee's response vector fits the Q_r matrix perfectly (i.e., the examinee's response vector matches one of the expected response patterns without any slips), the numerator of the HCI_i will be 0 and the HCI_i will have a value of 1.

Conversely, when the response vector completely misfits the Q_r matrix (i.e., the examinee correctly answers one item but fails to answer any item that requires the subset of attributes measured by the correct-answered item), the numerator of the HCI_i will be equal to $(2 \times N_{c_i})$, and the HCI_i will be -1. Therefore, if an examinee's HCI_i value tends to a value of -1, we can conclude that the examinee uses different cognitive skills (or in a different combination) from what the attribute hierarchy and the Q_r matrix indicate to

solve test items. In addition, the mean and standard deviation of the HCI_i can be used as indicators of the overall model-data fit. A high mean and low standard deviation suggest that the item response vectors fit the AHM model well.

Note, however, that the distribution of the HCI_i under the null hypothesis that the examinee response vector fits the Q_r matrix is unclear, and must be specified so that the critical value can be identified for significance testing. The current study proposed a simulation approach to approximate the null distribution of the HCI_i and, in turn, to set the critical value for the HCI_i to test the null hypothesis. If the observed HCI_i is less than the critical value, then the null hypothesis of the fit between an examinee response vector and the Q_r matrix should be rejected at the significance level associated with the critical value.

A Simulation Approach for Significance Testing of the HCI_i

In hypothesis testing, the results of an experiment are evaluated by assessing the null hypothesis. This is done because the probability of chance events associated with the null hypothesis can be estimated, but there is no unique mathematics for the probability of the alternative hypothesis (Pagano, 1990). To assess the null hypothesis, we first assume it is true and then test the reasonableness of this assumption by calculating the probability of obtaining the results due to chance. If the estimated probability is less than a critical probability level called the alpha level, which is assigned by researchers based on the theoretical and empirical considerations, the null hypothesis will be rejected. Another way to evaluate the null hypothesis is to use the critical value. The critical value is the value that bounds the critical region for rejection of the null hypothesis. The critical

region is defined as the area under the distribution curve that contains all the values of the statistic that allows rejection of the null hypothesis (Pagano, 1990). The critical value is determined by the alpha level and the null distribution. By comparing the observed value of the statistic based on the sample against the critical value, researchers can either reject or fail to reject the null hypothesis.

In the current study, our interest is to test the misfit of the observed examinee response vector in the AHM by using the proposed statistic HCI_i . Hence, the HCI_i for an observed item response vector can be evaluated by assessing the null hypothesis that the observed item response vector fits the AHM well. Ideally, we can calculate the probability of obtaining the observed HCI_i based on the distribution of the HCI_i when the null hypothesis is true. If the calculated probability turns out to be less than the alpha level, then we conclude the observed item response vector does not fit the AHM well.

Unfortunately, the probability distribution of the proposed HCI_i under the null hypothesis that the item response vector fits the AHM is unknown. To circumvent this problem, a simulation is used to approximate the null distribution of the HCI_i . By using simulated data of known characteristics, we choose the HCI_i value where the cumulative distribution function has a value of alpha as the critical value. In turn, we can determine whether the observed item response vector fits the AHM by comparing the HCI_i of this observed response vector to the obtained critical value.

In order to produce this outcome, a set of observed item response vectors must first be simulated from the attribute hierarchy and the Q_r matrix of the AHM. Since the purpose of this simulation is to approximate the distribution of the HCI_i under the null

hypothesis, the simulated data set should have a large sample size to decrease the errors due to random sampling. In this study, the sample size is 5000. Each observed item response vector can be generated by randomly adding slips to one of the expected response patterns derived from the attribute hierarchy and the Q_r matrix.

At the second step, the HCI_i value for each generated response vector must be calculated and placed in an ascending order. By doing this, the approximate distribution of the HCI_i under the null hypothesis can be obtained. The HCI_i has the property of ranging from -1 to +1. A larger HCI_i value for an observed response vector suggests this response vector fits the AHM better. Therefore, the critical region of the HCI_i is on the left side of its distribution. If the alpha level is 0.05, then the HCI_i value below which that 0.05 of the most extreme values fall is chosen as the critical value. In order for the null hypothesis to be rejected, the observed HCI_i must be smaller than the critical value. By rejecting the null hypothesis, researchers can conclude that the examinee uses different cognitive skills (or in a different combination) from those indicated by the Q_r matrix. In the next section, this simulation approach for significance testing of the HCI_i was be applied to the real data set to examine whether it fits or misfits the AHM.

An Application of the AHM and HCI_i to the Real Data

Data from a previously administered large-scale standardized achievement test were used in the current study. This large-scale test contains 54 mathematic items. All of the items, including multiple-choice and constructed-response items, are scored dichotomously. Students were expected to solve these items by using key mathematical concepts in areas such as Number and Operations; Algebra; Geometry; and Statistics.

Method

Participants. A random sample of 5000 examinees was extracted from the original data set.

Since the items from this test were not developed from an attribute hierarchy and a Q_r matrix, items need to be reviewed to identify the cognitive attributes. Given the complexity of this task, only eight algebra items were selected and analyzed in the current study. First, the selected items were reviewed to identify each cognitive component associated with the best strategy required to solve each item. A cognitive component can be described as one mathematical process, step, and/or procedure required to reach the correct solution. Second, the identified cognitive components were grouped into cognitive categories. The categories contain components where remediation is possible. By doing this, the Q_r matrix can be constructed in which each item is described using its underlying cognitive categories/attributes. Third, the categories were ordered into a hierarchy based upon their logical and/or psychological properties.

Next, the HCI_i was applied to the observed item responses to the eight items in order to examine the degree to which each response vector is consistent with the identified attribute hierarchy and Q_r matrix. The simulation approach introduced in the third section of this paper was used to set the critical value for testing the null hypothesis that the observed item responses fit the AHM.

To set the critical value for the HCI_i , first, 5000 expected item response vectors derived from the identified attribute hierarchy and the reduced Q-matrix were generated with the constraint that the total scores associated with the expected response patterns be normally distributed. Second, a uniform probability of 0.05 was employed to randomly

add slips from 1 to 0 and from 0 to 1 on the expected response patterns derived from the attribute hierarchy and the Q_r matrix. For each item, 5 percent of correct responses were randomly selected and changed from 1 to 0. Conversely, 5 percent of incorrect responses were randomly selected and changed from 0 to 1. Therefore, for each item, $0.05 \times 5000 = 250$ slips were added on the expected responses to the item. A simulated data set can be generated by adding a total of $250 \times 8 = 2000$ slips on the 5000 expected response patterns. A hundred data sets were generated. Since each simulated response vector was generated by randomly adding a small number of slips on one of the expected response patterns, simulees were considered to use the cognitive skills indicated by the attribute hierarchy and the Q_r matrix to solve the test items. Third, for each of the 100 simulated data sets, the HCI_i values were calculated and ordered according to a decreasing misfit. The value below which 5% of the most extreme misfit values fall was taken as the critical value for the data set. As a result, 100 critical values were calculated and the mean of the 100 critical values was used as the final critical value. The mean of the HCI_i s for each data set was used as a criterion to evaluate the overall model-data fit. The 100 means were ordered and the fifth smallest mean was used as the critical value for the overall model-data fit.

By comparing the observed HCI_i to the critical value for testing the individual person-fit, the null hypothesis of the fit between the observed response vector and the AHM can be evaluated. A smaller HCI_i than the critical value suggests that the examinee uses different cognitive skills (or in a different combination) from what the identified attribute hierarchy and Q_r matrix indicate in solving test items. The mean of the HCI_i s

over the 5000 examinees were calculated and compared to the critical value for the overall model data fit to evaluate the fit of the AHM to the overall data set.

Results and Discussion. From the review of the eight algebraic items, eight attributes were identified for the math test by the first author of this paper. Given that the purpose of this application is purely to illustrate the procedures of the HCI_i using real data, the developed hierarchy was not corroborated by other reviewers. The eight attributes are as follows: 1) prerequisite skills, 2) linear functions, 3) quadratic forms, 4) simple substitutions, 5) complex substitutions, 6) simple exponential computations, 7) complex exponential computations, and 8) representations. The detailed description of these 8 attributes is presented in Table 1. Figure 2 shows the hierarchical relationships of the 8 attributes. The Q_r matrix identified by the review of test items is shown as follows:

$$Q_r = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

The Q_r matrix is of order 8 by 8 (i.e., attributes by items). The first column of the Q_r matrix is interpreted as showing that attributes 1 and 2 are required to solve item 1 correctly. The last column of the Q_r matrix shows that item 8 requires attributes 1, 2 and 8. According to the HCI_i value calculated for each observed response vectors, 4596 out of 5000 response vectors fit the AHM model perfectly ($HCI_i = 1$), and 64 vectors misfit the model maximally ($HCI_i = -1$). The mean and variance of the HCI_i s are .9156 and .3132, respectively.

By using the simulation approach introduced in the third section of this paper, the critical value for testing the null hypothesis that the observed response pattern fits the identified attribute hierarchy and Q_r matrix is set at -0.1120. By comparing the HCI_i values for the observed response vectors to this critical value, 4923 out of 5000 observed HCI_i values are greater than the critical value, indicating that these 4923 observed response vectors fit the identified attribute hierarchy and Q_r matrix. In addition, the critical value for testing the overall model-data fit is 0.8230, which is calculated from the simulated data sets. The mean of the observed HCI_i s is greater than the critical value for the overall data fit, indicating a good overall model data fit. Hence, we can conclude that the attribute hierarchy in Figure 2 and its associated Q_r matrix correctly represent the cognitive skills that examinees used in solving these algebra test items.

Conclusions and Discussion

In CDMs, it is fundamentally important to evaluate the misfit of the item responses to the cognitive model. The purpose of this paper is to introduce the person-fit statistic, the HCI_i , for the AHM. The HCI_i can be used to evaluate the degree to which an observed examinee response pattern is consistent with a cognitive model and to determine the overall model-data fit. Although it is developed in the AHM framework, the HCI_i should be helpful in other CDMs that are guided by cognitive models given that the index allows the researcher to evaluate the fit of the cognitive model relative to the examinee response data. Specially, the HCI_i should be useful for the Q-matrix based CDMs, such as 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, 2003; Doignon & Falmagne, 1999; Haertel, 1989; Junker & Sijstma, 2001; Macready & Dayton, 1977; C. Tatsuoka, 2002), and the noisy input deterministic and gate model (NIDA) (Junker & Sijstma, 2001). In these models, the HCI_i can be directly used to evaluate the fit of the observed response vectors to the Q matrix and consequently to determine whether examinees' cognitive processes differ from the cognitive processes hypothesized in the Q matrix.

Two lines of studies related to the HCI_i will be conducted in future research. The first line of future research will focus on assessing the performance of the HCI_i in evaluating the misfit of the observed response vectors to the Q_r matrix of the AHM. A variety of data sets based on different types of hierarchies discussed by Leighton et al. (2004) will be generated by randomly adding different percentage of errors. To assess the performance of the HCI_i , the primary hypothesis is that higher HCI_i values will be obtained for the data sets with lower percentage of random errors.

The second line of future research will involve the investigation of other approaches for setting the critical values for the HCI_i . In the current study, the simulation approach is proposed and used to set the critical values for the item responses for the eight math items. To apply this approach, a simulation study should be conducted when a different set of item is used, which is not efficient. A thorough study will be conducted to examine whether the critical value should be varied for different conditions. Several conditions will be considered, such as the number of items, the number of attributes, the structure of the hierarchy, and complexity of the hierarchy, to investigate whether a general criterion can be developed for evaluating the HCI_i .

References

- Buck, G. (1990). *The testing of second language listening comprehension*. Unpublished doctoral dissertation, University of Lancaster, England.
- Buck, G. (1991). The testing of listening comprehension: an introspective study. *Language Testing, 8* (1), 67-91.
- Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing, 11* (2), 145-170.
- Buck, G., & Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*, 119-157.
- DiBello, L., Stout, W., & Roussos, L. (1995). Unified Cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. Chipman, & R. Brennen (Eds.), *Cognitively diagnostic assessment* (pp. 361-389). Hillsdale, NJ: Earlbaum.
- 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.
- Gierl, M. J., Leighton, J. P., & Hunka, S. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice, 19*, 34-44.
- 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.

- Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement, 4*, 105-126.
- 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.
- Liu, J., Feigenbaum, M., & Walker, M. E. (2004). New SAT and PSAT/NMSQT Spring 2003 field trial design. Paper presented in the symposium, *Analysis of Spring 2003 new SAT and new PSAT/NMSQT field trial*, at the annual meeting of the National Council on Measurement in Education, San Diego.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.
- Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. F. Chipman, & P. L. Brennan (Eds.), *Cognitively Diagnostic Assessment*, Hillsdale, NJ: Erlbaum.
- Nichols, P. D. (1994). A Framework for Developing Cognitively Diagnostic Assessment. *Review of Educational Research, 64* (4), 575-603.
- Pagano, R. R. (1990). *Understanding statistics in the behavioral sciences*. St. Paul: West.
- Scriven, M. (1999). The nature of evaluation part I: relation to psychology. *Practical Assessment, Research & Evaluation, 6* (11).
- Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden, 7*, 131-145.
- Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological*

Measurement, 16, 149-157.

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

Tatsuoka, C., & Ferguson, T. (1999). *Sequential Classification on Partially Ordered Sets* (Tech. Rep. NO. 99-05). Department of Statistics, George Washington University, Washington DC.

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, & P. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 327-359), Hillsdale, NJ: Erlbaum.

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

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

van der Flier, H. (1982). Deviant response patterns and comparability of test scores.

Journal of Cross-cultural Psychology, 13, 267-298.

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

Mesa Press.

Figure 1

A six-attribute hierarchy.

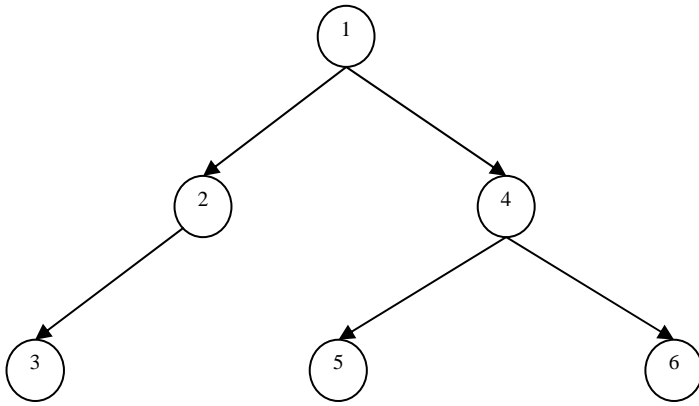


Figure 2

The attribute hierarchy for eight math items.

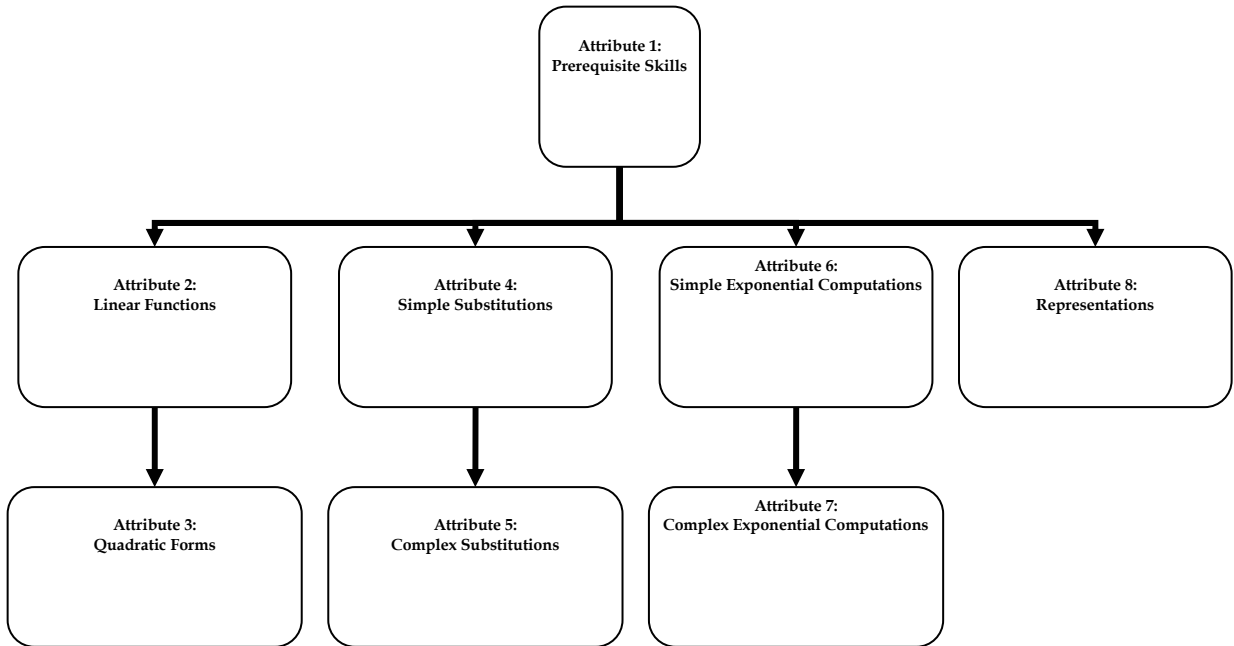


Table 1

The description of the 8 attributes

Attribute	Description
1	The understanding of the arithmetic operations implied by $+$, $-$, \times , $/$, $=$, absolute value, square, square root, exponent, $>$, $<$, \geq , \leq , and signed numbers; The ability to carry out basic computations, such as addition, subtraction, multiplication and division of whole numbers.
2	The ability to solve linear functions.
3	The ability to factor quadratic expressions and solve quadratic functions.
4	The ability to substitute the value of a variable for the letter.
5	The ability to substitute abstract expressions and rules.
6	The ability to carry out basic exponential computations, such as multiplication and division with two terms.
7	The ability to carry out more complicated exponential computations, such as multiplication and division with more than two terms.
8	The ability to translate words into mathematical expressions.