International Journal of Learning, Teaching and Educational Research
Vol. 16, No. 1, pp. 1-21, January 2017

Item Consistency Index: An Item-Fit Index for Cognitive Diagnostic Assessment

Hollis Lai,[1] Mark J. Gierl,[2] Ying Cui,[2] Oksana Babenko[3]

[1] School of Dentistry, Faculty of Medicine & Dentistry
[2]Centre for Research in Applied Measurement and Evaluation
[3]Department of Family Medicine, Faculty of Medicine & Dentistry
University of Alberta, Canada

Abstract. An item-fit index is a measure of how accurately a set of item responses can be predicted using the test design model. In a diagnostic assessment where items are used to evaluate student mastery on a set of cognitive skills, this index helps determine the alignment between the item responses and skills that each item is designed to measure. In this study, we introduce the Item Consistency Index (ICI), a modification of an existing person-model fit index, for diagnostic assessments. The ICI can be used to evaluate item-model fit on assessments designed with a Q-matrix. Results from both a simulation and real data study are presented. In the simulation study, the ICI identified poor-fitting items under three manipulated conditions: sample size, test length, and proportion of poor-fitting items. In the real-data study, the ICI detected three poor-fitting items for an operational diagnostic assessment in Grade 3 mathematics. Practical implications and future research directions for the ICI are also discussed.

Keywords: Item Consistency Index; cognitive diagnostic assessment; test development

Introduction

In educational testing, items are developed to elicit a correct response when examinees demonstrate adequate knowledge or understanding on the required tasks and skills within a specified domain. The methods of specifying knowledge, the conceptualization of content domains, and the design of how an item elicits responses are currently undergoing significant change with the evolution of our test designs. But one outcome that remains the same is that an item must assess the tasks and skills as intended, and the quality of each item must be judged to be high if it is to be included on the test. In most test designs, item discrimination power is a statistical criterion that is synonymous with describing item quality.

Item discrimination helps describe how well an item can differentiate examinees at different performance levels. Depending on the test design and how the scale of examinee performance is realized, different measures of item discrimination may be used. Additional information about item discrimination can also be garnered from measures of item-model fit. An item-model fit index describes the overall difference between real responses on a given item with a corresponding set of expected responses predicted by the test design. Item-model fit indices can be summarized, in general, as a ratio between the expected and actual correct responses on each item to compare the proportion of correct responses across examinees of different abilities with an expected correct proportion from the test design model. Different criterions that represent the examinee overall performance such as total score, estimated ability, or pseudo-scores have been used to group the responses of examinees' with similar ability to produce variations of item-model fit (Bock, 1972; Yen, 1981; Rost & von Davier, 1994; Orlando & Thissen, 2003). Application of item-model fit indices include the identification of poor performing items, cheating, or test administration anomalies, along with addressing issues related to dimensionality, item construction, calibration, and model selection (Reise, 1990).

Cognitive Diagnostic Assessment and Model Fit

Demand for more assessment feedback to better guide instruction and learning has led to the development of more complex test designs. Cognitive diagnostic assessment (CDA) is an example of a test design that yields enhanced assessment feedback by providing test takers with specific information about their problem-solving mastery on a given domain (Gierl, Leighton, & Hunka, 2007). The cornerstone of a CDA is the use of a cognitive model to guide test development. The use of a cognitive model allows CDA to provide enhanced feedback because cognitive information can be extracted from the examinees' item responses which, in turn, provide more detailed and instructionally relevant results to test takers. Compared to traditional tests where an item response is linked to a single outcome scale, the cognitive inferences made in CDA allow each item to measure multiple skills related to student learning. Due to the complexity of interpreting and modeling different aspects of cognitive skills, many approaches to modeling and scoring examinee responses are available. Sinharay, Puhan, and Haberman (2009) summarized three common features among different methods of CDA:

(1) tests assess student mastery based on a cognitive model of skills; (2) items probe student mastery on a pattern of skills expressed in a Q-matrix; and (3) items probing the same pattern of skill mastery should elicit a similar pattern of student responses.

An essential part of CDA development relies on the definition of a Q-matrix. The Q-matrix is an item-by-attribute matrix used to describe the skills probed by each item. For example, if a CDA is designed to determine examinee mastery on four skills, and an item was designed to elicit a correct response if the examinee has mastered the first and the fourth skill, then the row corresponded to that item in the Q-matrix would be expressed as {1,0,0,1}. The Q-matrix and the student response patterns are used to calibrate the model parameters and provide students with diagnostic results related to their cognitive problem-solving strengths and weaknesses.

To ensure that CDA results provide the most accurate information to examinees about their cognitive skills, the quality of CDA items must be scrutinized. The evaluations of the claim that items are to probe a specified set of skills have varied by the scope of how item-skill relations are represented. Model-data fit has traditionally been used to evaluate how items are aligned with construct of the skills based upon item responses. Few studies have investigated the relations of item-skill alignment. Wang, Shu, Shagn, and Xu (2015) have developed a measure which allows the evaluation of skill-to-item fit based on the DINA model that assumes a probabilistically scaled skill representation. To evaluate item-model fit in CDA, items need to be evaluated beyond the relationship of the correct responses on a particular item and single outcome scores. Because each item is designed to provide student mastery information on multiple skills, an item-model fit index is needed to ensure item responses are aligned with the intended cognitive skills.

Evaluating Model-Fit for CDA

The rationale evaluating model fit in CDA can be considered in two approaches, evaluating the fit with the expected psychometric properties of the test items or evaluating the fit of responses with the blueprint of skills. Existing developments tend to focus on the former approach. For example, Jang (2005) compared total raw score distributions between observed and predicted responses using the mean absolute difference (MAD). Jang's approach to evaluating model-fit is akin to IRT model fit approaches, where the level of fit is determined by total score differences between the expected and examinee results. But with each correct response of a CDA item linked to mastery on a vector of skills, evaluating item-model fit for CDA need to consider the fit of an item with the pre-requisite skills rather than a single test-level outcome.

Sinharay and Almond (2007) also developed an approach for evaluating item fit for CDA by assuming that examinees categorized with the same skill pattern should also have the same diagnostic outcome. With their approach, the proportion correct response for examinees with the same skill pattern is compared with the expected proportion predicted by the cognitive model. Differences between the expected and observed correct proportions are then summed across all skill patterns and weighted proportionally by sample size. That is, model-fit for item j was defined as:

$$X_j^2 = \sum_k \frac{N_k(O_{kj} - E_{kj})^2}{E_{kj}(N_k - E_{kj})},$$

where N_k is the number of examinees with skill pattern k, O_{kj} is the number of examinees with skill pattern k that responded correctly to item j, and E_{kj} is the product of the expected proportion of correct response for pattern k multiplied by N_k. Although this approach can be applied to account for fit among multiple sets of skills, results rely on an expected correct response rate of a given item for each skill pattern. As the expected correct response for a given set of skill pattern is not readily available, application of this method for determining model fit may be problematic. Moreover, a poor sample representation of a skill pattern or psychometrically indistinguishable skill patterns will also misestimate item-model fit. One way to avoid the influence of misclassification on an item-model fit measure for CDA is to comparatively evaluate items that measure the same skills. That is, items measuring the same skills are expected to elicit similar response patterns with one other.

Hierarchy Consistency Index (HCI)

One statistic developed specifically for CDA to evaluate person-model fit is the Hierarchy Consistency Index (HCI; Cui & Leighton, 2009; Cui & Li, 2014; Cui & Mousavi, 2015). The HCI is a statistic for evaluating the fit of the observed responses from an examinee with the expected responses from a CDA model based on a comparison between the observed and expected response vectors. The main assumption for the HCI is that if an examinee gives a correct response to an item requiring a set of skills, then the examinee is assumed to have mastered that set of skills and therefore should also respond correctly to items that designed to measure those skills. For example, if an examinee gives a correct response to an item that requires the first and third skill in a CDA that assess four skills (or an item with a skill pattern of [1,0,1,0] in the Q matrix), then the examinee is also expected to respond correctly to items that probe the same set of skills [1,0,1,0], or a subordinate or prerequisite of those skills (e.g., [1,0,0,0] , [0,0,1,0]), which require skills should have been acquired. In this manner, the number of misfitting responses across all items with their corresponding subsets of skills is calculated for each examinee to determine an index of person-fit.

Given *I* examinees were administered with *J* items, the HCI for examinee i is calculated as:

$$HCI_i = 1 - \frac{2\sum_{j=1}^{J}\sum_{g\epsilon s_j} X_{ij}\,(1-X_{ig})}{N}, \qquad (1)$$

where X_j is the examinee's scored response for item *j*, s_j is an index set that includes items requiring the subset of attributes measured by item *j*, and X_g is the examinee's scored response for item *g*. For example, if item *j* is answered correctly, then all items that measure the attributes or a subset of attributes probed by item *j* is represented by index set s_j, where *g* is an item index within s_j. *N* is the number of comparisons made across all s_j. The HCI has a maximum of 1 and minimum of -1, where a high positive HCI value represents good person-fit with the expected response model.

The HCI is a useful index for analyzing person-fit across different types of CDAs, as it requires only the use of the Q-matrix and examinee responses. In this study, we modify the HCI to create an index for analyzing item-model fit. Thus, the purpose of this study is twofold. First, we introduce and define an item-model fit index called the *item consistency index* (ICI). The ICI is used to evaluate the fit of an item related to the underlying cognitive model used to make diagnostic inferences with that item. Second, we present results from two studies to demonstrate both the simulated and practical performance of the ICI across of host of testing conditions typically found in diagnostic assessments.

Item Consistency Index (ICI)

As elaborated earlier, the HCI measures the proportion of misfitting observed examinee responses relative to the expected examinee responses on a diagnostic assessment. This principle can also be extended to evaluate item-fit. With the HCI, the misfitting responses related to each item is summed across all items for each examinee. As described in (1), misfit for examinee i (m_i) can be written as:

$$m_i = \sum_j^J \sum_{g\epsilon s_j} X_{ij}\,(1 - X_{ig})\,. \qquad (2)$$

Alternatively, to evaluate the misfit for item j, the number of misfitting responses from the subset of item j can be summed across all examinees. This modification can be written as:

$$m_j = \sum_i \sum_{g\in S_j} X_{ij}\left(1 - X_{ig}\right), \qquad (3)$$

where X_i is student i's score (1 or 0) to item j, and X_{i_g} is student i's score (1 or 0) to item g. Item g belongs to S_j, a subset of items that require the subset of skills measured by item j. In this manner, for a correct response to item j for examinee i ($X_{i_j} = 1$), one can consider any incorrect responses in S_j to be a misfit for examinee i. The number of misfits is then summed across all examinees.

It should be noted that the HCI only considers students' correct responses for analyzing misfit of a given item ($X_j = 1$). That is, misfit is calculated against the required skills only when students have provided the correct response. While this was adequate for analyzing misfit for person-fit, analyzing item-fit against a cognitive model also requires comparisons to be made when students respond to an item incorrectly ($X_j = 0$). As such, an evaluation of item-fit needs to account for this alternative comparison. For example, suppose an incorrect response was given on our exemplar item that required the skill pattern of [1,0,1,0]. From this item response, we could infer that the examinee does not possess all the necessary skills required to solve this item and, therefore, should respond incorrectly to all items that require the same skill pattern of [1,0,1,0]. Furthermore, the examinee should also respond incorrectly to items that require more skills than the current item (i.e., [1,1,1,0], [1,0,1,1], [1,1,1,1]). These items that require the same skill or a more complex skill pattern can be conceptualized as an alternative subset of item j (S_j^*), and a correct response in any of the items belonging to S_j^* can be conceptualized as a misfit. This outcome can be expressed as:

$$m_j^* = \sum_i \sum_{h \in S_j^*} X_{i_h}(1 - X_{i_j}). \qquad (4)$$

The set of alternative comparisons combined with comparisons from correct responses form the numerator of the ICI. To maintain the same scale of comparison with HCI, the numerator is then divided by the total number of comparisons, which effectively transforms the outcome to a proportion of misfit responses for item j. The proportion is then rescaled to a maximum of 1 and a minimum of -1. The ICI for item *j* is then given as:

$$ICI_j = 1 - \frac{2\sum_i \left[\sum_{g \in S_j} X_{i_j}(1-X_{i_g}) + \sum_{h \in S_j^*} X_{i_h}(1-X_{i_j})\right]}{N_{c_j}}, \qquad (5)$$

where X_{i_j} is student i's score (1 or 0) to item j, 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, S_j^* is an index set that includes items requiring all, but not limited to, the attributes measured by item j, X_{i_h} is student i's score (1 or 0) to item h where item h belongs to S_j^*, and N_{c_j} is the total number of comparisons for item j across all students.

To illustrate the calculation of the ICI, consider a hypothetical administration of a CDA with 15 items and a Q-matrix presented in (6).

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \qquad (6)$$

Suppose this CDA of four skills was administered to an examinee who produced the item response vector (0,0,0,0,0,1,1,0,0,0,0,0,0,0,0). That is, the examinee responded correctly to items 6 and 7 only. To calculate the ICI for item 6, we first consider that the examinee has responded to the item correctly, therefore comparisons should be made with items that require skills that are prerequisites to or same with the original item. In this case, items 2 and 4 belong in S_6. Since both item responses were incorrect, two comparisons were made ($N_{c_6} = 2$) and two unexpected responses were found ($m_6 = 2$) for this examinee. In addition, suppose we wanted to calculate the ICI of item 2 for this examinee. The alternative subset (S_j^*) will be needed since the examinee responded to the item incorrectly. For this instance, seven items form the alternative subset for item 2 ($S_2^* = \{3,6,7,10,11,14,15\}$). Since the examinee responded correctly to item 6 and 7, there were two unexpected responses ($m_2 = 2$) from a total of seven comparisons ($N_{c_2} = 7$). In this manner, the number of unexpected responses and comparisons are summed across all examinees and rescaled to form the ICI.

To demonstrate the performance of this item-model fit index across a variety of different testing situations, a simulation study was conducted to determine the performance of ICI for detecting poor-fitting items. Then, a real data study was conducted to demonstrate how the ICI can be applied in operational testing situations a CDA in Mathematics.

Methods and Results

Study 1: Simulation Study

To evaluate how well the ICI can identify items that fit poorly relative to their underlying cognitive model, a Monte-Carlo study was conducted by simulating responses from a diagnostic test designed to measure seven skills. To determine the performance of the ICI using simulated CDA data, examinee responses were generated under the Bernoulli distribution. In addition to generating examinee responses, different testing conditions were manipulated to probe conditions that may occur in a real CDA administration. Finally, to classify poor-fitting items using ICI, a common evaluation criterion was used to determine which items were fit poorly with the given cognitive model.

The simulation process is similar to the actual steps used in developing CDAs (Gierl, Leighton, & Hunka, 2007), where cognitive model, items, and responses were developed in a sequential manner. First, an existing cognitive model from Cui and Leighton (2009) was used to guide the simulation process. The cognitive model consists of seven skills, with 15 patterns of skill mastery identified as permissible. The patterns of required skills for each item are expressed in the Q-matrix presented in Table A1 in the Appendix. To generate examinee responses, examinees were first assigned to an expected pattern of skill mastery from one of the 15 skill patterns. In addition to the 15 skill patterns, a null pattern [0,0,0,0,0,0,0] was also used to represent examinees who did not master any skills. In total, sixteen expected skill patterns are distributed equally among the sample examinees. To simulate response for an examinee on a given item, the examinee's assigned skill pattern is compared with the skills required by that item as indicated by the Q matrix. A probability of correct response is assigned based on whether the examinee has all the prerequisite skills of the item. Based on this assigned probability, the examinee's response to each item was generated using a Bernoulli function.

To simulate the effectiveness of ICI under different testing conditions, three factors were manipulated. First, the number of items representing each skill pattern in the CDA was varied by three levels. If a CDA is lengthened by including multiple items probing the same set of skills, then the reliability of each corresponding skill measured is expected to increase (Gierl, Cui, & Zhou, 2009). In our study, the number of items in the CDA varied by one, two, or three items representing each possible skill pattern. These three levels of variation on a total of 15 skill patterns resulted in test lengths of 15, 30, and 45 items, respectively.

Second, unlike the related person-fit HCI which is independent of sample size, the ICI is based on the proportion of misfit responses from *all* examinees. Therefore, different sample sizes may affect the outcome of the ICI. Three levels of sample sizes were manipulated: 800, 1600, and 2400. Since the 15 skill patterns and a null pattern are distributed equally among the examinees, the numbers of examinees representing each skill pattern are 50, 100, and 150, respectively.

Third, an important feature for an item-model fit index is to detect items that fit poorly with the expected response determined by cognitive model. This concept is contaminated when the ICI is influenced by misfitting items related to the skills of the original item. To investigate whether the proportion of poor-fitting items have an effect on the ICI, the proportion of poor-fitting items were manipulated at three levels proportional to the test length: 5%, 10% and 25%. In Cui and Leighton (2009), a well-fitting item was deemed to have a 10% chance for slips, where an examinee without mastery of the necessary skills will have a 10% chance of responding correctly while an examinee who has mastered the necessary skills will have a 90% chance of responding correctly. While there can be many reasons for an item to fit poorly with the underlying cognitive model (e.g., model misspecification, item quality, option availability), generally a poor-fitting item yields a response that is aberrant from the cognitive model. To simulate a poor-fitting item, items responses were generated close to random. Table 1 contains the probabilities of correct response given the level of item fit (good or poor fit) and whether the examinee possesses the required set of skills. Taken together, three manipulated factors with three levels each yielded a total of 27 conditions as shown in Table A2 of the Appendix.

Table 1. Correct response probability given the level of item fit and whether the examinee possesses the required set of skills

	Item Fit	
Required skills	Good	Poor
Present	0.9	0.6
Not present	0.1	0.4

To evaluate the effectiveness of the ICI for detecting poor-fitting items, a criterion is needed for the ICI to differentiate between poor- and well-fitting items. A classification approach was used to measure the precision of the ICI in this study. A cut-score criterion, set at an ICI value of 0.5, was used to illustrate the classification characteristics for poor-fitting items. For example, if an item was calculated to have an ICI value of less than 0.5, then that item was deemed to fit poorly with the expected responses from the cognitive model. This preliminary criterion for dichotomizing item fit was needed because no point of comparison currently exists in determining an appropriate level of fit with an existing cognitive model. Further, an ICI value of 0.5 for any item translates to roughly 75% of the responses on a given item fitting with the expected skill pattern as defined by the cognitive model. Using this initial cut-score, we could then classify items as poor- or well-fitting.

To ensure the classification results were consistently produced, each of the 27 testing conditions was replicated 100 times. The dependent variables for the simulation study included the average proportion of correctly identified poor-fitting items and misclassification of well-fitting items across all conditions. The simulation environment, the implementation of the ICI, and the replication of results were programmed in R (R Core Development Team, 2011), and are available from the first author.

Table 2 contains a summary of the mean ICIs for each condition. The mean ICIs were calculated separately for the poor- and well-fitting items. The overall mean for poor-fitting items was 0.30 whereas the mean ICI for well-fitting items was 0.53. Three observations must be noted from the results in Table 2. First, test length tended to have a positive impact on the values of ICI. For example, CDAs with only one item measuring each skill pattern (i.e., test length=15) had consistently lower ICIs compared to CDAs with two or three items measuring each skill (i.e., test length=30 or 45). Second, as expected, the magnitude of the mean ICI differences between poor and well-fitting items tended to decrease when an increase in poor-fitting items included in the ICI. Third, the means of ICI were relatively stable across different sample sizes for each condition.

Table 2. Summary of the mean ICIs across the three variables manipulated in the simulation study

Sample Size	Proportion of Poor Fitting Items	Test Length	Mean ICI	
			Poor Fitting Items	Well-Fitting Items
800	5%	15	0.24	0.49
	5%	30	0.22	0.57
	5%	45	0.30	0.59
	10%	15	0.31	0.48
	10%	30	0.29	0.56
	10%	45	0.38	0.58
	25%	15	0.37	0.43
	25%	30	0.29	0.56
	25%	45	0.32	0.51

1600	5%	15	0.21	0.41
	5%	30	0.22	0.56
	5%	45	0.29	0.59
	10%	15	0.27	0.44
	10%	30	0.29	0.57
	10%	45	0.38	0.58
	25%	15	0.36	0.41
	25%	30	0.29	0.56
	25%	45	0.32	0.51
2400	5%	15	0.24	0.55
	5%	30	0.23	0.58
	5%	45	0.30	0.59
	10%	15	0.32	0.53
	10%	30	0.30	0.57
	10%	45	0.38	0.58
	25%	15	0.32	0.53
	25%	30	0.29	0.56
	25%	45	0.32	0.51

Items were also classified based on the cut-score criterion. This simulation process was repeated 100 times, with the correct classification rate, or power, being the likelihood of correctly identifying a poor-fitting item using the ICI across the conditions in the simulation study. The power values for the 27 conditions are shown in Table 3. The conditions with the highest power were found in CDAs with the longest test-length (45), specifically with conditions that had the largest proportion of poor-fitting items (25%). Under those conditions, the highest power was 0.99, meaning that for the ICI criterion of 0.50, 99% of all poor-fitting items were correctly classified across 100 replications. The lowest power values were found in conditions with the smallest sample size (800), where a power of 0.67 was found for a 30-item CDA with 5% of poor-fitting items and 1600 examinees.

Table 3. Power of ICI for identifying poor-fitting items

Test Length	Sample Size	Proportion of Poor-Fitting Items		
		5%	10%	25%
15	800	0.68	0.76	0.92
	1600	0.93	0.89	0.95
	2400	0.79	0.79	0.92
30	800	0.67	0.73	0.79
	1600	0.77	0.74	0.81
	2400	0.73	0.72	0.79
45	800	0.76	0.80	0.99
	1600	0.77	0.83	0.99
	2400	0.76	0.81	0.99

Table 4 summarizes the likelihood of a well-fitting item being misclassified by the ICI as a poor-fitting item in each condition. The lowest misclassification rates were associated with CDAs that have the longest test-length (45) and the smallest proportion of poor-fitting items (5%). Under those conditions, the lowest misclassification rate was 15%. The highest error rates were observed with the shortest test length (15), where misclassification was 78%.

Taken together, the simulation study results highlight important trends and outcomes that can be used to interpret how accurately the ICI identifies poor-fitting items. The power values of ICI were erratic when the number of items probing each skill pattern was small, but stabilized as the number of items representing each skill pattern increased. For example, each increase in test length resulted in a decrease in the variation of power values among the same proportion of poor-fitting items and between different sample sizes. This finding suggests that the reliability of using the ICI to classify poor-fitting items is related to the reliability of the CDA as a whole. Moreover, the proportions of misclassification were approximately 2.5 times higher in CDAs with a single item representing each test skills as compared to the other two levels. This outcome further supports the conclusion that as skills are measured more accurately, the ICI better distinguishes poor- from well-fitting items.

Table 4. Misclassification rate of ICI in identifying well-fitting items

Test	Sample	Proportion of Well-Fitting Items		
Length	Size	5%	10%	25%
15	2400	0.28	0.35	0.66
	1600	0.78	0.65	0.72
	800	0.50	0.50	0.66
30	2400	0.16	0.20	0.22
	1600	0.28	0.20	0.27
	800	0.27	0.22	0.24
45	2400	0.15	0.18	0.33
	1600	0.17	0.19	0.34
	800	0.15	0.19	0.33

There were no obvious trends that the sample size manipulated across the three levels yielded important differences among the power or misclassification of well-fitting items. This finding suggests that the sample sizes used in this study do not yield important ICI differences across our study conditions. This outcome could also suggest that the representation of approximately 50 examinees per skill pattern may be sufficient for evaluation of the ICI.

When the proportion of poor-fitting items was manipulated, the power increased with the proportion of poor-fitting items in the CDA, where the overall power rose as the proportion of poor fitting item increased. An increase of poor-fitting items also yielded more misclassification of well-fitting items. This finding suggests that poor-fitting item responses contribute to an overall decrease in the magnitude of ICI, where the resulting errors are reflected using the classification criterion of 0.50.

Study 2: Use Case Application

The purpose of the second study is to demonstrate how the ICI can be used to identify poor-fitting items in an operational CDA. The ICI was used to evaluate item-model fit for a CDA program designed to assess students' knowledge and skills in Grade 3 mathematics. From this CDA program, 324 students responded to an 18-item CDA (see Gierl, Alves, & Taylor-Majeau, 2010).

The CDA we used was designed to evaluate student mastery for subtraction skills. Each item was designed to yield specific diagnostic information in a hierarchy of cognitive skills were the first skill was the easiest (Subtraction of two consecutive 2 digit numbers) and the last skill was most difficult (Subtraction of two 2 digit numbers using the digits 1 to 9 with regrouping). The CDA was developed as follows. First, a cognitive model of task performance was created by specifying the cognitive skills necessary to master subtraction in Grade 3. The domain of subtraction was further specified into a set of six attributes related in a linearly hierarchical manner by a group of subject matter experts. The attributes produced a total of seven unique patterns of skill mastery (six plus null). Three items were created by content experts to probe student mastery on each attribute to ensure adequate representativeness of each skill pattern resulting in eighteen items for this CDA. The test was administered to students in 17 Grade 3 classrooms. A list of the attributes and the Q-matrix for the 18-item CDA are shown in Table A3 and Table A4 of the Appendix, respectively.

Three hundred and twenty four student responses were collected, which would yield approximately 45 students per skill pattern if the patterns were distributed equally across the skills. Participating teachers would first instruct on the topics relevant to subtraction within their classrooms, and then administer the CDA to students at a convenient time within two-week of instruction. The CDA was delivered using an online computer-based testing system. Students were presented with CDA items that contain both an item stem to prompt for a typed-response and an interactive multimedia component that provided additional information for students to understand the item. From this administration process, responses were collected, formatted and scored dichotomously. As the participation of this CDA was voluntary, students with greater than two missing responses were removed from the analysis to minimize unmotivated responses (as the completion of the CDA was not mandatory). For the purposes of demonstrating the ICI, only the scored student responses were used.

The results are summarized first at the test level and then at the item level. Overall, the results were ideal at the test level. The median HCI, which is used to quantify the fit of the responses to the expected model of response on a CDA, was 0.81. With a cut-off of 0.70 as the quality criterion for CDA designs (Gierl, Alves, & Taylor-Marjeau, 2010), this result suggests that the student responses fit with the expect model of response for subtraction. As the purpose of this CDA is to identify non-mastery students in order to refine and enhance instruction, the majority of students were expected to master the CDA.

At the item level, Table 5 provides a summary of the results from the subtraction CDA. The p-values of each item and the discrimination value (i.e., point-biserial correlation) are presented along with the ICI values. Three findings should be noted from these results. First, the ICI was not correlated with either the difficulty or discrimination values. This result supports the idea that item-model fit is summarizing a different outcome from the classically defined notion of difficulty and discrimination. Second, with items created in a principled manner, with three items representing each skill pattern, the real data results support the results of the simulation study. Further, as p-values decrease, ICI values increase because the items change from measuring simple to more complex skills. Third, using the cut-score criterion of 0.50 from the simulation study, only three items were deemed to have poor item fit (Items 1, 2, 3). The poor ICI values for these items may suggest a problem at the attribute level (see Table A3 in the Appendix for the description of the skills assessed). It is important to note that without the ICI conventional scoring and psychometric approaches would not have identified issues of misfit at the attribute level, where items one through three are performing nominally at the item level. Although subject matter experts did not evaluate the cognitive model in the light of the student results, a follow-up study may find that a reorganization of the attributes may yield better fitting responses.

Table 5. Summary of the results from the subtraction CDA

Attribute	Item Number	P-Value	Discrimination	ICI
1	1	0.76	0.58	0.22
	2	0.78	0.87	0.39
	3	0.80	0.96	0.46
2	4	0.84	0.89	0.64
	5	0.87	1.11	0.72
	6	0.85	0.94	0.65
3	7	0.86	1.06	0.76
	8	0.80	0.68	0.65
	9	0.84	1.01	0.75
4	10	0.77	0.79	0.73
	11	0.72	0.78	0.72
	12	0.75	0.82	0.73
5	13	0.74	0.82	0.78
	14	0.77	0.92	0.79
	15	0.79	0.98	0.80
6	16	0.35	0.56	0.81
	17	0.34	0.57	0.81
	18	0.33	0.53	0.80

Discussion

The purpose of this study is to introduce a statistic for determining item-model fit with CDA. The item consistency index (ICI), an extension of a person-fit index for CDA called the Hierarchy Consistency Index (HCI), is a standardized outcome that measures the ratio of misfitting responses relative to the total number of response across all examinees on a given item. Similar to the HCI, the requirements for evaluating item-model fit using the ICI is an item-by-attribute definition of skill mastery called the Q-matrix in addition to the student response vectors. The ICI has a maximum value of 1, which suggests all students responded identically to an expected skill pattern, and a minimum value of -1, which suggests item responses were the exact opposite to what the expected skill patterns suggest. We present two use cases to demonstrate the properties of the ICI under simulation. In addition, we demonstrate the applicability of the ICI through the use of real data to highlight how the ICI can be applied to identify poor-fitting items on a CDA. These two proof-of-concept applications demonstrate how the ICI can be applied in the real world and call for future studies to establish better evaluation criterion for the ICI.

Results from the simulation study provided some general insights on how the ICI performs as a method for detecting item misfit in CDA across a range of testing conditions. Using a cut-score classification method to determine poor-fitting items, the ICI was able to identify the majority of the poor fitting items across different simulated conditions. Although the item-model fit is described in a range by the ICI, the use of a cut-score to classify poor fitting items provided a simple outcome to interpret for evaluating how the ICI will perform in a given testing scenario. In addition, results from the simulation study demonstrated a few assumptions that must be met for the ICI to detect item misfit accurately. The number of items used for each skill pattern and the total number of poor fitting items were two features that affected ICI performance. The implication from these findings demonstrate that although CDA demands a different paradigm of scoring and statistical approaches, traditional issues such as consistency of the responses for a given set of skill can still be problematic in estimating item-model fit. From our simulation results, we suggest the use of three items per attribute or more per skill pattern to ensure adequate ICI detection. This finding is consistent with the research in establishing an adequate reliability in measuring attributes of skills (Gierl, Cui, & Zhou, 2009), where the authors stated that the idea of a short yet diagnostic test will not likely yield results with sufficient reliability.

Sinharay and Almond (2007) noted that tests with many poor-fitting items indicate a problem with the overall model, whereas tests with few poor-fitting items indicate problems lie in the items themselves. In our simulation, we demonstrated that the ICI will produce similar results, where an increase of poor-fitting items in a CDA will lower the precision of the ICI. This finding may be linked to the fact that as more poor-fitting items are introduced, these items affect the fit of items requiring the same set of skills leading to an overall decrease in magnitude of ICIs. Table A5 in the Appendix illustrates this effect, where the mean ICI for well- and poor-fitting items under the 45-item simulation decreases as the proportion of ICI increases. In sum, a rigorous and principled test development process is needed for CDA to ensure all test items are created with minimal deviation from the expected set of skills they were designed to probe. Otherwise, poor model-fit results will lead to poor diagnostic outcomes.

The second study provided a snapshot on the utility of the ICI when applied to an operational CDA. Using a set of carefully designed CDA items, the ICI detected three consecutive poor-fitting items at the beginning of the assessment. This finding suggests that the ICI can not only be used for evaluating item-model fit, but can also be used for evaluating the consequences of test design at the item, attribute, or the cognitive model level. In our example, the three items flagged as poor fitting measure the same attribute revealing that the attribute may be mis-specified in the cognitive model. In addition, the independence of ICI from the difficulty and discrimination values suggest that item model-fit for CDA provides a unique measure of how an item is able to accurately predict performance. Hence, the definition of a good item for CDA may not only be how well an item is able to distinguish poor-performers from good-performers, but also how consistently an item can elicit responses that match the expected response patterns specified in the cognitive model (i.e., Q-matrix).

Item-model fit is challenging to measure, especially when cognitive inferences are involved in the test design. Items have to be aligned with the cognitive skills in the Q-matrix, skills have to be defined and organized in a systematic manner, and examinee responses have to match the expected skill patterns. The ICI can provide a source of evidence for identifying poor-fitting items or poor models for Q-matrix based CDA.

Implications for Future Research

By introducing and demonstrating an item-model fit index for CDA, our study provides two practical implications for the development of diagnostic assessments in addition to a new measure of item-fit. The ICI has the benefit of applicability, meaning that it can be used with a Q-matrix based CDA for determining the relationship between items and skills. Using the Q-matrix, item and examinee responses can be compared to provide a measure of item model-fit. While research on CDA has prompted a plethora of diagnostic scoring methods, one common starting point is the use of the Q-matrix in defining the skills and item requirements. Because item development, validation, and administration all depend on the veracity of the Q-matrix, evidence for validating the cognitive model is paramount. The ICI offers some initial evidence that can be used for validating the definition of skills through item response patterns to determine the relative fit between an item and its set of required skills defined in the Q-matrix.

While the ICI provides a new statistical method for scrutinizing CDA development, the second study highlighted the fact that the most crucial part of a well-designed CDA remains with item development. The importance of item development is, sometimes, neglected in CDA. Although CDA scoring methods can account for different levels of skill contributions, the link between how a skill is measured with how the skill is presented in the form of an item remains largely a subjective interpretation of the test developer and content specialist who create the CDA. To reliably measure a set of skills, multiple items are needed. Yet creating parallel items is often time consuming and expensive. Ensuring that each item is uniformly developed with the same set of skills is one critical activity in test development for CDA that ensures examinees receive useful diagnostic feedback.The ICI is co-dependent with all items requiring a related set of skills. Therefore, to ensure adequate item model-fit, every item in the CDA must adhere to a high level of quality and alignment relative to the expected skill the item is designed to measure.

Through introducing an item model-fit index for CDA, we have demonstrated how such measure can be applied to identify problematic items that are aberrant from the expected response model. This initial study provides directions of future research as further investigation is needed to apply and validate the use of this index. We also suggest three directions of future research. First, more research is needed to ensure different structures of knowledge represented by the Q-matrix can be evaluated with the ICI to identify misfitting items. The number of possible skill pattern representation increases exponentially as the number of evaluated skills increases, therefore more research is needed to ensure ICI provides an appropriate measure for different organization of skills. Second, guidelines to interpret ICIs are needed so we can accurately identify and distinguish adequate and problematic items. As the ICI provides a scaled measure of item model-fit, interpretations of the index has not yet been established and is required to determine the adequacy threshold of item model-fit. Third, as the reliability of CDA measures is highly dependent on the defined skills, more research is needed to determine which model structure is ideal in the application of the ICI. Our analysis relies on non-compensatory attributes, meaning skills are independently defined, acquired and cannot be moderated by existence of other skills. This will likely limit the ICI in measuring item fit for testing complex skills but not for general skills such as elementary mathematics. More research is needed to evaluate appropriate use cases of the ICI.

References

Bock, R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Cui, Y., & Leighton, J. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(4), 429-449.

Cui, Y, & Li, J. C.-H. (2014). Evaluating person fit for cognitive diagnostic assessment. *Applied Psychological Measurement, 39,* 223-238.

Cui, Y, & Mousavi, A. (2015). Explore the usefulness of person-fit analysis on large scale assessment. *International Journal of Testing, 15*, 23-49.

Gierl, M., Leighton, J., & Hunka, S. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 242-274). Cambridge, MA: Cambridge University Press.

Gierl, M., Cui, Y., & Zhou, J. (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(3), 293-313.

Gierl, M., Alves, C., & Taylor-Majeau, R. (2010). Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Knowledge and Skills in Mathematics: An Operational Implementation of Cognitive Diagnostic Assessment. *International Journal of Testing, 10*(4), 318-341.

Jang, E. (2005). A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL (Doctoral dissertation). University of Illinois at Urbana-Champaign, IL, USA.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27(4), 289-298.*

R Development Core Team (2011). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. .

Reise, S. (1990). A Comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*(2), 127-137.

Rost, J., & von Davier, M. (1994). A conditional item-fit index for Rasch models. *Applied Psychological Measurement, 18*(2), 171-182.

Sinharay, S., Puhan, G., & Haberman, S. (2009, April). Reporting diagnostic scores: Temptations, pitfalls, and some solutions. Paper presented at the National Council on Measurement in Education, San Diego, CA, USA.

Sinharay, S., & Almond, R. (2007). Assessing fit of cognitive diagnostic models a case study. *Educational and Psychological Measurement. 67*(2), 239-257.

Wang, C., Shu, Z., Shagn, Z., & Xu, G. (2015). Assessing Item-Level Fit for the DINA Model. *Applied Psychological Measurement*, 1-14.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

APPENDIX A

Table A1. The Q-matrix and skill patterns used for the simulation of CDA responses

	Skill						
Pattern	1	2	3	4	5	6	7
1	1	0	0	0	0	0	0
2	1	1	0	0	0	0	0
3	1	1	1	0	0	0	0
4	1	1	0	1	0	0	0
5	1	1	1	1	0	0	0
6	1	1	0	1	1	0	0
7	1	1	1	1	1	0	0
8	1	1	0	1	0	1	0
9	1	1	1	1	0	1	0
10	1	1	0	1	1	1	0
11	1	1	1	1	1	1	0
12	1	1	0	1	0	1	1
13	1	1	1	1	0	1	1
14	1	1	0	1	1	1	1
15	1	1	1	1	1	1	1

Table A2. Variables manipulated in the simulation

	Level		
Conditions	1	2	3
Test length	15	30	45
Sample size	800	1600	2400
Proportion of poor-fitting items	5%	10%	25%

Table A3. Description of the skills assessed in the CDA for subtraction in Grade 3

Cognitive Attribute #	Skill Descriptor: Apply a mental mathematics strategy to subtract
6	Two 2 digit numbers using the digits 1 to 9 with regrouping
5	Two 2 digit doubles (e.g., 24, 36, 48, 12)
4	Two 2 digit numbers where only the subtrahend is a multiple of 10
3	Ten from a 2 digit number
2	Two 2 digit numbers where the minuend and subtrahend are multiples of 10
1	Two consecutive 2 digit numbers (e.g., 11, 22, 33)

Table A4. Q-matrix of the CDA for subtraction in Grade 3

	Skill					
Pattern	1	2	3	4	5	6
1	1	0	0	0	0	0
2	1	0	0	0	0	0
3	1	0	0	0	0	0
4	1	1	0	0	0	0
5	1	1	0	0	0	0
6	1	1	0	0	0	0
7	1	1	1	0	0	0
8	1	1	1	0	0	0
9	1	1	1	0	0	0
10	1	1	1	1	0	0
11	1	1	1	1	0	0
12	1	1	1	1	0	0
13	1	1	1	1	1	0
14	1	1	1	1	1	0
15	1	1	1	1	1	0
16	1	1	1	1	1	1
17	1	1	1	1	1	1
18	1	1	1	1	1	1

Table A5. Summary of the mean ICI in extreme situations when *n*=2400

Item Quality	Proportion of Poor-Fitting Items			
	0%	25%	50%	100%
Well-Fitting Items	0.61	0.49	0.39	n/a
Poor-Fitting Items	n/a	0.33	0.28	0.15