Abstract

The goal of this study was to investigate the usefulness of person-fit analysis in validating student score inferences in a cognitive diagnostic assessment. In this study, a two-stage procedure was used to evaluate person fit for a diagnostic test in the domain of statistical hypothesis testing. In the first stage, the person-fit statistic, the *hierarchy consistency index* (*HCI*; Cui, 2007; Cui & Leighton, 2009), was used to identify the misfitting student item-score vectors. In the second stage, students' verbal reports were collected to provide additional information about students' response processes so as to reveal the actual causes of misfits. This two-stage procedure helped to identify the misfits of item-score vectors to the cognitive model used in the design and analysis of the diagnostic test, and to discover the reasons of misfits so that students' problem solving strategies were better understood and their performances were interpreted in a more meaningful way.

Key words: Person-fit analysis; Cognitive diagnostic assessments; Hierarchy consistency index; Student score validation; Verbal reports

Cognitive diagnostic assessment serves as an important effort to improve the informational value of traditional educational tests by providing more specific information about students' problem solving strengths and weaknesses (Leighton & Gierl, 2007a). It is designed to classify examinees into a set of latent classes or states, often called attribute patterns, which are defined in term of the mastery or non-mastery of a set of attributes (e.g., knowledge and skills) being measured by the test. The diagnostic feedback produced by cognitive diagnostic assessment has the potential to help teachers design effective instructional interventions.

Cognitive diagnostic assessment uses a cognitive model or theory to guide the development of test items and to interpret student performance on tests. Leighton and Gierl (2007b) defined a cognitive model in educational measurement as "a simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills students at different levels of learning have acquired and to facilitate the explanation and prediction of students' performance" (p. 6). Although a cognitive model is aimed at identifying the knowledge and skills required to solve educational tasks in a content domain, a challenge of developing such a model is that students may vary widely in the knowledge and skills they possess and use in solving test items. A cognitive model may not accurately describe the knowledge and skills every student uses to solve items, which poses a potential thread to the validity of the inferences to be made about each student's test-based performance.

Given that the accuracy of the cognitive model underwrites the validity of diagnostic feedback from the test, studies are deemed necessary to investigate ways of evaluating the accuracy of the cognitive model in characterizing the knowledge and skills used by each individual student to solve test items. One way to accomplish this is to conduct a *person-fit study* (Meijer & Sijtsma, 2001) to evaluate whether individual students actually use the knowledge and skills specified in the cognitive model to solve test items. A person fit statistic can be employed to assess how well the pattern of each student’s item responses fits the typical item-score patterns that are consistent with the cognitive model. Although the statistical results are important, the use of a person-fit statistic to identify misfitting item-response vectors is only the first step of a person-fit analysis as it does not provide clear indications of how misfits occur or what types of misfitting response behaviour underlie test performance.

Therefore, the objective of this study was to use data obtained in the domain of statistical hypothesis testing to investigate the usefulness of a two-stage approach to person-fit analysis for evaluating the validity of a cognitive model used in a diagnostic test. In the first stage, we used a person-fit statistic, the *hierarchy consistency index* (HCI; Cui, 2007; Cui & Leighton, 2009), to statistically identify the misfitting student item-score vectors. The HCI is designed specifically to examine the degree to which an observed item response vector fits the cognitive model used in test design and analysis. In the second stage, students’ verbal reports were collected to provide additional information about students’ response processes so as to reveal the actual causes of misfits. This type of information provides relatively detailed pictures of how students actually solve items on tests, which has the potential to help understand the reasons for misfits so that the results from person-fit statistics can be interpreted substantively and meaningfully. By analyzing the person fit of student item responses relative to the expectations of the cognitive model used in diagnostic test design, the appropriateness of the cognitive model can be examined at individual student level to ensure the measurement accuracy of student performance.

The paper is divided into three sections. First, we briefly review the hierarchy consistency index as a person-fit statistic that can be used to identify misfitting student item-score vectors in a cognitive diagnostic test. Second, we conduct a person-fit analysis by first identifying misfitting item-score vectors using the HCI and then uncovering the causes of misfits using students’ verbal reports. The results from the person-fit analysis are presented and discussed in this section. Third, we summarize the paper by discussing the limitations of our approach and the usefulness of person-fit studies in helping shape perspectives on how students solve items and in validating the student score inferences by a test.

Section 1: Hierarchy Consistency Index

In this section we review the hierarchy consistency index (HCI; Cui, 2007; Cui & Leighton, 2009). The HCI is a person-fit statistic designed to explicitly evaluate the degree to which student item responses are consistent with the implications of the cognitive model generated to represent the knowledge, skills, and processes students use in solving items in a test. The goal of the HCI is to identify students who are not being measured well by a cognitive diagnostic test. The logic of the HCI is that a student should not be able to answer an item correctly unless the student has solved its prerequisite items successfully. The HCI ranges from -1.0 to 1.0, with a value close to 1.0 indicating that the student responds consistently with the expectations derived from the cognitive model, and a value close to -1.0 indicating that the student responds unexpectedly or differently from the responses expected under a given cognitive model. With HCI values close to -1.0, the cognitive model should not be used as a basis for making inferences about that student's performance.

The HCI compares an observed item-score vector to expected item-score vectors that are derived from the cognitive model or a task analysis of test items where the prerequisite relationships among test items are specified. The prerequisite relationship between two items exists when the set of knowledge and skills required by one item is a subset of attributes required by the other item. For example, item one is considered as the prerequisite to item two if item one measures only a subset of knowledge and skills required by item two. If a student answers item two correctly, the student is expected to produce a correct answer to item one. Otherwise, a misfit is found. The HCI for student i is given by

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

where

$S_{correct_i}$ is an index set that includes items correctly answered by student i,

X_{i_j} is student i's score (1 or 0) to item j, where item j belongs to $S_{correct_i}$,

S_j is an index set including items that are prerequisite to item j,

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

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

The term $\sum_{j\in S_{correct_i}}\sum_{g\in S_j} X_{i_j}(1 - X_{i_g})$ in the numerator of the HCI represents the number of misfits between student i's item responses and the expected responses as specified by the prerequisite relationship among items. When student i correctly answers item j, $X_{i_j} = 1$, and the student is expected to also correctly answer item g that belongs to S_j, namely, $X_{i_g} = 1 \left(g \in S_j\right)$. If the student fails to correctly answer item g, $X_{i_g} = 0$, then $X_{i_j}\left(1 - X_{i_g}\right) = 1$, and it is a misfit of the response vector i to the cognitive model. Thus, $\sum_{j\in S_{correct_i}}\sum_{g\in S_j} X_{i_j}(1 - X_{i_g})$ is equal to the total number of misfits. The denominator of the HCI, N_{c_i}, contains the total number of comparisons for items that are correctly answered by student i. When the numerator of the HCI is set to equal the total number of misfits multiplied by 2, the HCI has the property of ranging from -1 to +1, which makes it easy to interpret. When no misfit is found, the numerator of the HCI will be 0 and the HCI will have a value of 1. Conversely, when the response vector completely misfits the cognitive model (i.e., the student correctly answers one item but fails to answer all of its prerequisite items), the numerator of the HCI will be equal to $2N_{c_i}$ and the HCI will be -1. Cui (2007) provides tentative guidelines for interpreting the HCI results. She suggests the use of an HCI value of .8 as the cut point for distinguishing a good and a moderate person fit, and .6 as the cut point for distinguishing a moderate and a poor fit. HCI values of .8 and .6 indicate that on average, of all the pair-wise comparisons of student responses, 10% (i.e., [1-.8]/2) and 20% (i.e., [1-.6]/2) of comparisons do not correspond to the model expectations. The selection of these two cut points is to ensure the power of the HCI in identifying misfitting item response vectors considering that the consequence of failing to identify the misfit of a student response vector is to falsely validate the inferences made from the cognitive model and further lead to incorrect decisions about student performance. However, as pointed out by Cui, these criteria are partly based on subjective judgment so they cannot be considered as infallible.

It should be noted that the calculation of the HCI relies on item complexity as determined by the set of knowledge and skills required for a correct item response. The HCI examines whether the prerequisite relationship among test items implied by a cognitive model is reflected in the student response data. Therefore, the HCI focuses on the evaluation of the fit of observed item responses relative to the cognitive model of a diagnostic test. And the HCI is not specifically designed to examine the fit of student response data relative to the statistical classification model used in the data analysis of cognitive diagnostic assessment. As a result, the calculation of HCI does not directly take into account of item parameters as they rely on the assumption of the specific form of probabilistic model of item responses and subject to estimation errors. Although this is advantageous as the HCI is not affected by the inaccuracy of item parameter estimates, the consequence is that test items of poor quality could lead to low HCI values. Item format might potentially affect the HCI results given that multiple choice questions are more likely subject to guessing and testwiseness. As a result, there are multiple possible explanations for a low HCI value, including a) an inaccurate cognitive model (i.e., the cognitive model fails to provide a valid representation of the knowledge and skills used by the student in solving test items), b) poor test items (e.g., items with low discriminating power or items that fail to measure what they are designed to measure), and c) aberrant response behaviour such as creative responding (i.e., answering simple items incorrectly for the reason of interpreting these items in a unique, creative manner), or random responding (i.e., randomly guessing on multiple-choice items). However, in any of these cases, the cognitive model should not be used as a basis for diagnostic inferences about the student's performance.

Section 2: A Person-Fit Study

In this study, a two-stage procedure was used to evaluate person fit for a diagnostic test in the domain of statistical hypothesis testing. In the first stage, the person-fit statistic, the *HCI*, was used to statistically test the fit of each student response vector relative to the cognitive model that was developed and used in test design and analysis. In this way, misfitting item response vectors were statistically identified. In the second stage, students' verbal reports were collected to help validate the *HCI* results and reveal the actual causes of misfits. Verbal reports are commonly used in cognitive psychological research to gather important evidence for validating cognitive models (Leighton, 2004). In recent years, verbal reports have been used increasingly by researchers of educational measurement to identify the knowledge and skills students actually use when responding to test items (e.g., Hamilton, Nussbaum, & Snow, 1997; Leighton, Cui, & Cor, 2009; Leighton & Gierl, 2007a). By analyzing student verbal reports, a detailed description of students' cognitive strategies is obtained, which can help validate and facilitate interpretation of the person-fit statistic results so the reasons for misfits can be understood. This two-stage procedure helped identify the misfits of item-score vectors to the cognitive model, and to discover the reasons of misfits so that students' problem solving strategies were better understood and their performances were interpreted in a more meaningful way.

Participants

Students enrolled in a graduate-level introductory statistics course at a Canadian university were recruited for this study. Students were told that their participation in the study was completely voluntary and there was absolutely NO consequence if they chose not to participate. Of the 38 students who enrolled in the course, 18 students responded to recruitment and agreed to participate. However, only 13 students (10 female and 3 male) actually attended the interview sessions. Students who participated in the study were from a variety of faculties and departments in the social and health sciences. Participating students were compensated with 20 dollars.

Instrument Development

A diagnostic test was created in the domain of hypothesis testing. To do so, a cognitive model must be first specified. Leighton and Gierl (2007b) identified three categories of cognitive models in educational measurement, including cognitive models of test specification, of domain mastery, and of task performance. Of these three categories, a cognitive model of domain mastery is often used to help develop classroom-based tests. It should be noted that a model of domain mastery does not usually specify the detailed thinking or reasoning processes of human problem solving, as emphasized typically by a cognitive model in computer science or cognitive psychology (e.g., Anderson’s ACT-R theory). Rather, according to Leighton and Gierl, a model of domain mastery features a comprehensive listing of knowledge and skills that are believed to conceptualize mastery within a domain. Such a model is useful for educational purposes as a profile of the mastery or nonmastery of knowledge and skills can inform teachers and students of the content areas that need improvement and therefore direct instruction decisions and remediation efforts.

To generate a model of domain mastery, two instructors of the statistical course were asked to identify the key knowledge (i.e., useful concepts and their properties) and skills (i.e., application/manipulation of knowledge) in the area of hypothesis testing. Six key categories of knowledge and skills that characterize domain mastery in the area of statistical hypothesis testing were first identified. These categories included prerequisite knowledge (e.g., the knowledge of mean, standard deviation, and etc.), knowledge of basic concepts in hypothesis testing (e.g., the knowledge of type I and II errors, Power, and et al.), the ability to conduct z test, single-sample t test, independent-samples t test, and related-samples t test. The interrelationships among the skill categories were also specified to facilitate the person-fit analysis using the HCI. The specification of the interrelationships among knowledge and skills was consistent with the findings from cognitive research (e.g., Kuhn, 2001; Vosniadou & Brewer, 1992), which have indicated that knowledge and skills do not operate in isolation but belong to a network of interrelated competencies. The six skill categories were ordered into a hierarchical cognitive model from basic to complex based on the expectation that complex skills should not be possessed unless basic skills have been mastered. The ordering of knowledge and skills was based on their logical properties and was consistent with the way most statistics textbooks ordered their concepts in the area of hypothesis testing.

Next, key sub-skills and their prerequisite relationships were specified within each skill category to provide a finer-grained description of the knowledge and skills required for successful performance in the task domain. In this way, detailed feedback about students' strength and weaknesses can be provided based on the cognitive model. In total, 26 sub-skills were identified across the six knowledge and skill categories. Within each of the six categories, different sub-skills can be ordered into a hierarchy. The full hierarchy displaying the prerequisite relationship of sub-skills both *among* and *within* the six categories is presented in Figure 1. According to Figure 1, all sub-skills related in category 1 are the prerequisite to all sub-skills in category 2, which are in turn the prerequisite to all sub-skills in category 3, and so on. The prerequisite relationship among sub-skills within each category is also presented. For example, category 1 has a total of four sub-skills, including the concepts of mean, standard deviation, standard error and probabilities. The prerequisite relationship among these four sub-skills is also shown in Figure 1. Because the model specified at this stage has not been validated using the empirical evidence of student responses, we considered it to be a hypothesized model. The prerequisite relationships might not hold in practice or for different populations of students. Therefore, a person-fit analysis was conducted to collect empirical evidence to validate the cognitive model and its structure.

To measure each of the 26 sub-skills specified in the cognitive model, 26 items were then selected from an item bank (which is provided as part of the instructor's manual for the textbook used in the course). That is, one item was purposely selected to measure each sub-skill specifically. And these items were mostly well-defined tasks in the sense that each item was associated with a smaller number of solution paths for getting the correct response. Of these items, 10 were multiple-choice questions and 16 were short-answer questions. Of the six knowledge and skill categories, items in categories one and two were all multiple-choice questions intended to measure students' understanding of key statistical concepts. Items in categories three to six were all short-answer questions and some of them required students to do some simple forms of hand calculations. All items were dichotomously scored.

Although the selection of test items from an item bank is a common practice for classroom assessments, we suspect that the development of items in the bank was not subject to a rigorous validation process. Therefore, we presumed that the 26 items selected for the statistical test might not measure exactly what they were designed to measure or they contained some sources of ambiguity. It was hoped that the person-fit analysis could begin to identify potential sources of item ambiguity and invalidity.

Procedure

Students who agreed to participate were interviewed in a quiet room at a Canadian university. They were asked to think aloud as they responded to the set of items. Standard think-aloud interview procedures outlined by Ericsson and Simon (1993) were used. During the interview, students were provided instructions about the think-aloud requirement. Before the actual task, participants were given a chance to practice thinking aloud using sample exercises. Once training was completed, students were asked to think aloud while solving test items and prompted to keep thinking aloud if they remain silent for more than 10 seconds. Students' verbalizations were audio recorded so that their verbal report data were preserved and later transcribed. Students also had the opportunity to debrief with the interviewer. All students finished the tasks within approximately 1 hour and 10 minutes.

Results of Statistical Analysis Using the HCI

The mean and standard deviation of the total score on the 26 items across the 13 students were 18.54 and 4.82, respectively. Item difficulties ranged from .46 to 1.00. Student proportion correct scores for the full test and each of the six knowledge and skill categories are presented in Table 1. Each student's set of responses to test items were then examined by using the HCI to test its statistical fit relative to the expectations under the cognitive model. Higher HCI values are expected for students who used model-specified knowledge and skills to solve test items, while low HCI values indicate that students likely used different sets of knowledge and skills from those specified in the cognitive model when solving test items. The HCI was calculated for each of the 13 students based on their responses to the 26 items in the test. The HCI values ranged from .29 to .85, with a mean of .65 and a standard deviation of .17. The mean of the HCI values for the full test can be used as an indicator for the overall model-data fit. The mean of HCI value equal to .65 indicated that, on average, of all the pair-wise comparisons of student responses, 17.5% (i.e., [1-.65]/2) of comparisons did not correspond to the model expectations. With Cui (2007)'s guidelines for interpreting the HCI, an overall moderate fit of student item responses to the cognitive model was found.

In addition, to find where the misfits occurred, the HCI values were also calculated for each of the six knowledge and skill categories based on the prerequisite relationships among the sub-skills in the category (see Figure 2). The HCI values are presented in Table 2. The mean of HCI values ranged from .44 to 1.00 across the six categories. HCI results suggested that students' response vectors tended to show relatively higher degrees of consistencies with expected patterns for categories 3 to 6 comparing with categories 1 and 2. One possible explanation is that items in categories 1 and 2 are all multiple choice questions, which are more likely subject to guessing and testwiseness. This led to more inconsistencies between observed and expected student responses.

Of the six sub-skill categories, category 1 was associated with the lowest mean HCI value, .44. Out of the 13 students, seven students' HCI values were .33 or below. Given that a large proportion of student item-response vectors showed low HCI values, it is likely that either the cognitive model of sub-skill category 1 does not provide a valid representation of student knowledge and skills, or test items do not measure what they are designed to measure or fail to discriminate well between students who have and students who have not mastered the knowledge and skills that the items are designed to measure.

It should be noted that the HCI value for the full test was not a simple average of the HCI values for the six knowledge and skill categories. This is because the HCI value for the full test not only considered the consistency of student item responses within each category but also took into account the relationships among the categories as presented in Figure 1. For this reason, response vectors that showed perfect fit across all six categories did not necessarily produce an HCI value of 1.0 for the full test. Student 4 was such a case.

The coding of verbal reports

Each of the 26 items in the test was selected from the item bank to measure specifically one of the 26 sub-skills outlined in the cognitive model. A task analysis was conducted to identify all, if any, alternative knowledge and skills that would be used by students to solve each item. This information serves as guidance for the coding of student verbal reports. The first author, blind to the

HCI values associated with each item response vector, analyzed verbal report data to identify the knowledge and skills students actually used to solve each item. Student verbal reports were coded by comparing the knowledge and skills actually used by students in solving test items with those expected from the cognitive model and the task analysis. For each item, if a student used the model-specified knowledge and skills to solve the item, the verbal reports on the item were coded as 1. If a student used one of the alternative strategies other than what was expected by the cognitive model, the verbal reports were coded as 0. For example, a student stated "*...Probability is one of the things I have a real problem with, and I'm just taking a guess at this one*". If partial knowledge and skills specified in the model were used in solving the item, then the verbal reports were coded as .5. For example, a student reported "*Well, it definitely is [a] and it's [b], so it's got to be [d]"* when responding to a multiple-choice question in which option [d] was "all of the other three choices are correct" and was unsure about option [c]. It should be noted that the correctness of item responses did not necessarily affect the coding of verbal reports. Rather, verbal reports were coded by evaluating the agreement between knowledge and skills students actually used to solve test items and the expected knowledge and skills based on the cognitive model. As long as a student uses the model-specified knowledge and skills to reason and solve test items, his or her verbal reports would be coded as 1 regardless whether the student's final answer is correct or wrong. The coding scores of student verbal reports for each knowledge and skill category were presented in Table 3. Although the coding scores across the six knowledge and skill categories can be simply added to serve as the overall coding scores for the full test, they do not reflect the inconsistencies of student responses as imposed by the prerequisite relationship among categories and therefore tend to overestimate the overall fit of cognitive model to student item responses. As a result, the coding scores for the full test are not reported in Table 3.

Overall comparison of verbal report codings with the HCI results

Generally speaking, the codings of the verbal reports showed a consistent pattern with the *HCI* results across the six subtests. That is, student item responses to categories 3 to 6 showed a higher level of consistency to the cognitive model than responses to categories 1 and 2. For categories 1 and 2, the correlations between the codings of student verbal reports and the *HCI* values were found to be .52 (df=11, p=.04) and .62 (df=11, p=.01), respectively. For categories 3 to 6, both the *HCI* and verbal reports indicated a high level of consistency of the actual knowledge and skills used in solving items and those specified by the cognitive model across all students. The mean HCI values range from .94 to 1.00 across the four categories, and the average verbal report codings all equal to 4 (i.e., 100%). Due to the low variability of the *HCI* values and the codings of verbal reports among the students, the correlation was not calculated for categories 3 to 6.

Evaluating the validity of the cognitive model with the HCI and verbal reports

The cognitive model specifies the key knowledge and skills required to solve items correctly in the test domain and the prerequisite relationships among these knowledge and skill. Relatively low HCI values and verbal report codings were found in skill categories 1 and 2. This result indicates that both the HCI and student verbal reports suggest the presence of the misfits of student observed responses relative to the cognitive model in skill categories 1 and 2. These two categories focused on measuring students' understanding of statistical concepts while other categories placed more attention on testing students' procedural skills of conducting hypothesis tests. The inspection of the verbal reports revealed that some conceptual questions were more likely to elicit alternative solution paths and therefore students showed great diversity in the way they solved these questions. For example, one question related to one-tailed hypothesis test asked whether the following statement is correct: *"If the null hypothesis is rejected using a one-tailed test, then it certainly would be rejected if the researcher had used a two-tailed test with the same alpha level.*" Some students recalled the definition of one- and two-tailed tests from the textbook and then use the graphical approach to compare the rejection regions of the two tests; some students related the statement to the power of a hypothesis test and recalled the factual knowledge that one-tailed test can increase the power of a test; some other students solved the questions by creating concrete examples and comparing the critical values of one- and two-tailed tests. In comparison, students showed much less diversity in solving questions that tested their procedure skills of conducting hypothesis tests. The majority of students followed closely the computational steps of hypothesis testing. This finding indicates that it might be relatively easier for a cognitive model to specify the sequence of steps used by students in solving an item that probes procedure skills and capture the prerequisite relationships among these procedural skills, which results in a better fit to student responses.

The HCI values for the full test also examined the fit of student responses relative to the overall model in which the relationships among the six categories is specified. The inspection of verbal reports for students with relatively low HCI values for the full test revealed that some students had difficulty in answering the conceptual questions in categories 1 and 2 but showed competence in performing statistical hypothesis tests using procedure skills in categories 3 to 6. This finding showed that the prerequisite relationship between the conceptual understanding of hypothesis tests and the actual execution of hypothesis testing procedures may not hold for these students.

In addition, as discussed earlier, perfect fit was found for Student 4 within each of the six sub-skill category. However, misfits of student responses relative to the full model were found. By further examining her response vector and verbal reports, we found that student 4 successfully solved items related to category 4, single-sample t test, but failed to answer questions associated with category 3, z test, which was not consistent with the prerequisite ordering between these two categories as illustrated in Figure 1. In fact, several students showed a similar pattern of responses with correct answers to items measuring t test but incorrect responses to z-test items. Although the ordering between z test and t test is theoretically sound considering that t test is a simple variation of z test, the prerequisite relationship might not be as strong as previously thought. One possible explanation for this outcome is that some students may have mastered how to mechanically follow procedures before fully understanding the theoretical principles behind them.

Evaluating the quality of test items with the HCI and verbal reports

To evaluate the quality of test items, student verbal reports were examined to see whether each item was measuring what it was supposed to measure and whether the item tended to elicit different knowledge and skills other than those specified in the cognitive model. To accomplish this, the proportion of students who used model-specified knowledge and skills in solving each item was calculated. Out of 26 items, 24 items were answered with model-specified knowledge and skills by over 75% of the students. For the remaining two items (items 1 and 2), the proportion of students who used model-specified knowledge and skills to answer each item was 54% and 69%, respectively. Both items belonged to sub-skill category 1, which partly explained why this category was associated with the lowest mean HCI value across the six categories.

The contents of these two items along with student verbal reports were reviewed to find out why students tended to use different knowledge and skills to solve these two items. We found that both items were multiple choice questions and they were subject to guessing and testwiseness. For example, item 1 asked students to identify which statement correctly describes the concept "statistical mean" from the four options. The item is presented below.

A sample mean is _____.

a. The average score of the sample
b. An unbiased estimate of a population mean
c. The most commonly used measure of central tendency
d. All of the other three choices are correct

By reviewing student verbal reports, we found that the majority of students had difficulty in evaluating whether the statement "a sample mean is an unbiased estimate of the population mean" was correct. Although this statement was closely related to what the item was designed to measure, it involved a concept (i.e., unbiased statistics) that was not specified in the cognitive model. In addition, this item tended to lend itself to alternative problem-solving strategies such as testwiseness. For example, one student stated*"...Okay, well I know that it's for sure the average sample and the most commonly-used, so it must be D"*. In comparison, another student reported that "*...an unbiased estimate of a population mean. [pause] Well, I'm not sure about this; I'm just going to be on the safe side and circle (a), what I'm sure, pretty sure about, the average score of the sample. So it's (a)."* Clearly, testwiseness was an unexpected factor contributing to the variance of student responses to this item, and, as a result, the fit of the cognitive model to student responses was compromised.

Evaluating student aberrant response behaviour with the HCI and verbal reports

Verbal reports were examined to assess whether aberrant response behaviour (e.g., random guessing or high level of test anxiety) had occurred, especially for students whose item responses were associated with relatively low HCI values. For example, student 10 produced the lowest HCI value for the full test, .29. Upon inspection of her HCI values for different skill categories, it was found that misfits were more significant in skill category 1 where the HCI value was -.33. For skill category 1, student 10 answered one relatively complex item correctly but failed to get its two prerequisite items right. Student verbal reports were then examined to identify the causes of the identified misfit. It was found that the student really struggled with the two prerequisite items as evidenced by her comments such as "*I don't know. No, I don't know, so I'm just going to go with the* —." Now, the question is why the student was able to produce a correct response to a more complex item when she failed to solve its prerequisite items successfully. The answer to this question became apparent as she stated "*…I think this is the standard error; yes. So there is no thinking involved in this question; this is what I just remember. Yes."* Student verbal reports not only explained why misfits occurred but also suggested that the quality of the item needed to be improved. Another student also produced the same pattern of responses to the three items. That is, the student answered correctly the relatively complex item but failed to answer its two prerequisite items. Upon inspection of her verbal reports, we found that the reason that she couldn't answer the two prerequisite items correctly was due to the high level of anxiety at the beginning of the interview as evidenced by her comments "*…I'm all nervous; I can't really concentrate very well... It's so silly. It happened to me when I was doing my mid-term; I just get all nervous.*" After a couple of questions, the studentcalmed down and started to engage in the problem solving process.

Section 3: Summary and Discussion

The goal of this paper was to investigate the usefulness of person-fit analysis in assessing the fit of student responses to the cognitive model used in the test design and interpretation of cognitive diagnostic assessment. Although the potential usefulness of person-fit analysis in validating student score inferences has been recognized and documented, to a large degree, this approach to test validation has not yet been applied to real testing situations. In this study, we used real data to demonstrate that results from person fit analysis have great potential to serve as an important source of evidence for testing the validity of cognitive model used in test development and interpretation. A two-stage procedure was used for investigating person fit where substantive and statistical approaches are unified to provide evidence about whether students use the knowledge and skills specified in the cognitive model to solve test items.

Our person-fit analysis procedure begins with the use of the person-fit statistic, the HCI, to statistically evaluate the degree to which the cognitive model fitted student observed responses. The evaluation of the misfit of an item-response vector relative to the cognitive model is focused on assessing whether students' actual item response patterns match the expected response patterns based on the prerequisite relationships among test items. Following the statistical analysis for testing person fit, verbal report data are used to validate the statistical results from the HCI and to find out the actual causes of the misfits. Given that low HCI values could be due to ill-specified cognitive models, low-quality test items, and/or student aberrant response behaviours, the results of the HCI and verbal reports are analyzed and compared at three different levels. At the cognitive model level, the goal is to evaluate the validity of the cognitive model used in test design by investigating whether model-specified knowledge and skills are actually used by students in answering test items. At the test item level, the analysis focus on examining whether items tend to elicit different knowledge and skills other than those specified in the cognitive model. And at the individual student level, the HCI and verbal reports are examined and compared to reveal whether aberrant response behaviours (e.g., random guessing or high level of test anxiety) have occurred.

This two-stage procedure for person fit was applied to real student response data from a diagnostic test in the domain of statistical hypothesis testing. For this sample, both the HCI results and verbal report data indicate the inconsistency between the ordering of knowledge and skills in the cognitive model and the actual patterns of student responses. Because our sample is small and not constructed to be representative of the population, it is unwise to generalize our results to the larger population. However, our results do provide insight into the difficulties that students in an introductory statistical course often encounter. That is, students tend to have difficulty in understanding the key concepts in statistical hypothesis testing. For some students, execution of statistical procedures resulted in the correct answer without true understanding of the prerequisite concepts. The results of this study suggest that increased instructional and assessment emphasis on foundational conceptual knowledge is needed. Additionally, making explicit the connections between statistical concepts to specific features of statistical procedures through instruction may facilitate more meaningful learning for the student. One example of this is connecting the concepts of sampling, sampling distributions, and probability to determining statistical significance of a test statistic. Results from this study also indicate that the use of diagnostic assessment as a formative tool has the potential to help tailor and monitor the effectiveness of instruction in addressing student misconceptions while enhancing student learning by providing feedback on the development of knowledge of statistical concepts.

Our person fit analysis also helped identify items that lend themselves to testwiseness or failed to elicit high-level thinking. Our next step in the future study will be to modify the cognitive model and test items according to the results from the HCI and verbal report data. For example, some concepts might be overlooked from our cognitive model and instrument, such as the concept of statistical distribution. Another modification for a future study is rewriting all the multiple choice items in short answer format to remove the confounding effect of item format on HCI results. The revised test will be administered to students, and person fit and model-data fit will be re-evaluated. Our hypothesis is that improved person fit and overall model-data fit should be observed.

This study illustrated the usefulness of person-fit analysis in validating the cognitive model used in a small-scale classroom diagnostic test. We believe that person-fit analysis should be beneficial for large-scale assessment as well, especially at the stage of model and test development. It helps evaluate the underlying cognitive model and the quality of test items, and assists identifying student aberrant response behaviors. In large-scale assessment, the sample size requirement for person-fit analysis would be considerably large in order for results to be generalizable. In addition, for high-stake assessments, the content analysis of the test domain and the development of the cognitive model must undergo a more rigorous process potentially involving panel discussion and focus group participation.

In terms of future directions, more research is called for regarding the development and evaluation of different person-fit statistics for cognitive diagnostic assessment. The *HCI* focuses on examining the fit of observed student response patterns relative to the theoretical cognitive model of diagnostic assessment. However, the *HCI* does not directly evaluate the fit of student responses relative to the statistical classification model used in the data analysis of student response patterns.

Additional research is needed to develop person-fit statistics to specifically evaluate the consistency of student actual item responses and the probabilities of a correct response calculated based on the probabilistic model of item responses. In addition, the *HCI* can be used with conjunctive cognitive models where the mastery of all the knowledge and skills measured by an item is required for successful performance. Person-fit statistics are needed for compensatory cognitive models, which make the assumption that high ability on one skill can compensate for low ability on other skills. In addition, the *HCI* are useful in test domains where students are expected to gain knowledge and skills gradually from simple to complex. In these domains, relatively simple knowledge and skills must be possessed in order for students to move to the next stage of learning in which more complex knowledge and skills are involved. However, for content domains where prerequisite hierarchical relationships among skills and tasks are not apparent, the *HCI* cannot be used and therefore new person-fit statistics need to be developed.

Currently, person-fit statistics are studied largely by researchers and therefore remain an area of potential for improving measurement practice. This study represented an endeavour to evaluate person fit in an applied setting. More systematic studies must be undertaken before person-fit analysis is ready for routine use in the analysis of student item response data.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Author. Washington, DC.

Cui, Y. (2007). *The hierarchical consistency index: A person-fit statistic for the attribute hierarchical method.* Unpublished Doctoral Dissertation. University of Alberta, Department of Educational Psychology.

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

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. Ed.). Cambridge, MA: MIT Press.

Hamilton, L.S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*, 181-200.

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

Kuhn, D. (2001). Why development does (and does not occur) occur: Evidence from the domain of inductive reasoning. In J. L. McClelland & R. Siegler (Eds.), *Mechanisms of Cognitive Development: Behavioral and Neural Perspectives.* (pp. 221-249). Hillsdale, NJ: Erlbaum.

Leighton, J. P. (2004). Avoiding misconceptions, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23,* 6-15.

Leighton, J. P., Cui. Y., & Cor, M. K. (2009). Empirically testing cognitive models generated from verbal reports: An application of the attribute hierarchy method. *Journal of Applied Measurement in Education, 22,* 229-254.

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

Leighton, J. P., & Gierl, M. J. (2007b). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*, 3–16.

Lohman, D.F. (2000). Complex information processing and intelligence. In R.J. Sternberg (Ed.), *Handbook of intelligence* (pp. 285-340). NY: Cambridge University Press.

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

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

Figure 1. A cognitive model of domain mastery in the area of statistical hypothesis testing

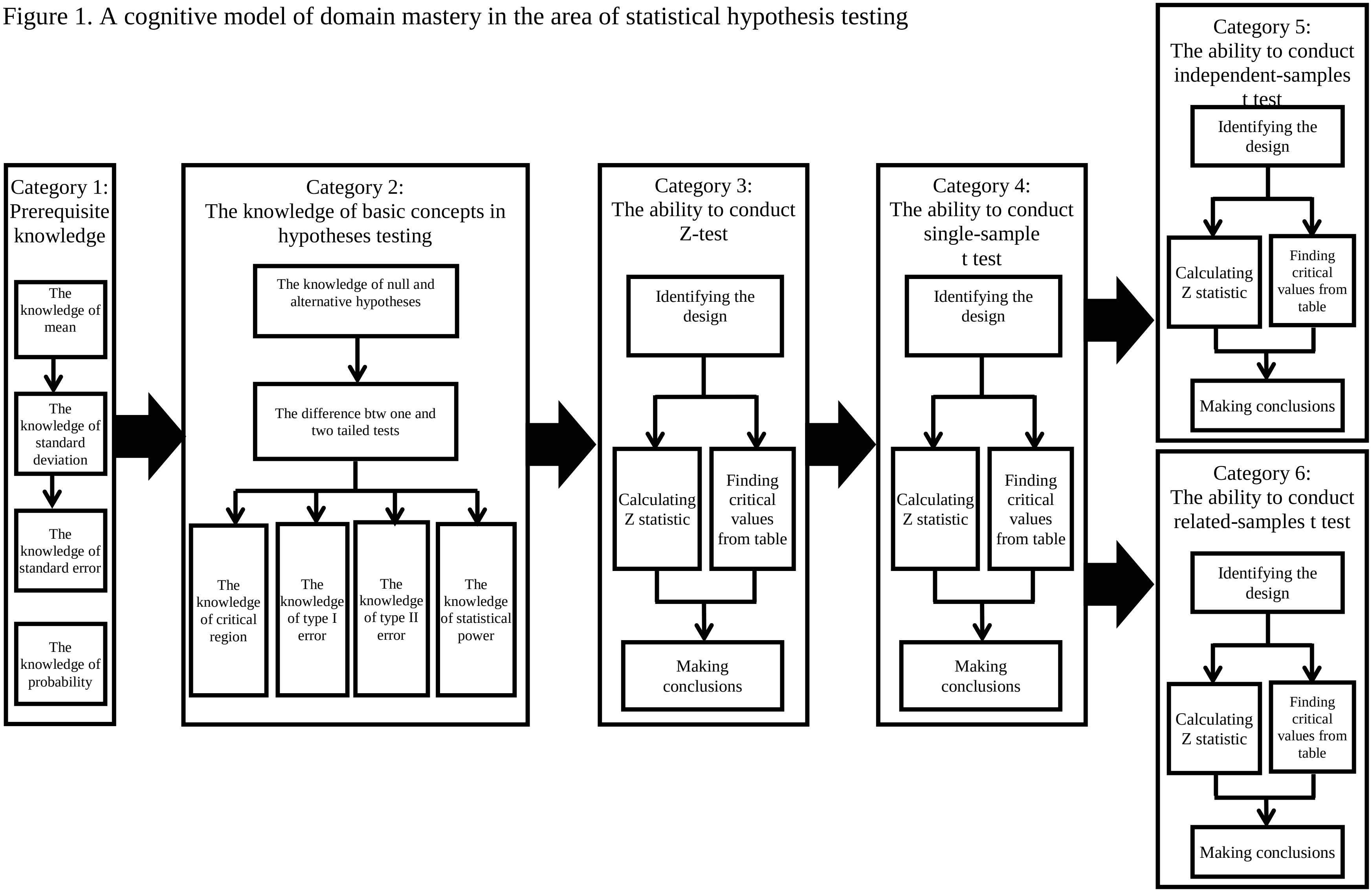
Category 1:
Prerequisite knowledge
The knowledge of mean
The knowledge of standard deviation
The knowledge of standard error
The knowledge of probability
Category 2:
The knowledge of basic concepts in hypotheses testing
The knowledge of null and alternative hypotheses
The difference btw one and two tailed tests
The knowledge of critical region
The knowledge of type I error
The knowledge of type II error
The knowledge of statistical power
Category 3:
The ability to conduct Z-test
Identifying the design
Calculating Z statistic
Finding critical values from table
Making conclusions
Category 4:
The ability to conduct single-sample t test
Identifying the design
Calculating Z statistic
Finding critical values from table
Making conclusions
Category 5:
The ability to conduct independent-samples t test
Identifying the design
Calculating Z statistic
Finding critical values from table
Making conclusions
Category 6:
The ability to conduct related-samples t test
Identifying the design
Calculating Z statistic
Finding critical values from table
Making conclusions

Table 1

Proportion correct scores for the full test and for the subtests associated with each knowledge and skill category

		Subtest associated with each knowledge and skill category					
Student	Full test	S_1	S_2	S_3	S_4	S_5	S_6
1	0.96	0.75	1.00	1.00	1.00	1.00	1.00
2	0.85	0.75	0.83	1.00	0.50	1.00	1.00
3	0.88	0.75	0.83	0.75	1.00	1.00	1.00
4	0.46	0.00	1.00	0.50	0.00	0.00	1.00
5	0.54	0.50	0.83	0.50	1.00	0.00	0.25
6	0.73	0.50	1.00	1.00	0.75	0.50	0.50
7	0.58	1.00	0.67	0.00	0.50	1.00	0.25
8	0.62	0.75	1.00	0.25	0.75	0.25	0.50
9	0.73	0.75	1.00	0.75	1.00	0.25	0.50
10	0.50	0.25	0.33	0.00	0.50	1.00	1.00
A	0.92	0.75	0.83	1.00	1.00	1.00	1.00
B	0.54	0.25	1.00	0.50	0.25	0.00	1.00
C	0.96	0.75	1.00	1.00	1.00	1.00	1.00
Mean	0.96	0.75	1.00	1.00	1.00	1.00	1.00

Note: Students 1 to 10 are female; students A to C are male.

Table 2

HCI values for the full test and for the subtests associated with each knowledge and skill category

		Subtest associated with each knowledge and skill category					
Student	Full test	S_1	S_2	S_3	S_4	S_5	S_6
1	0.84	0.33	1.00	1.00	1.00	1.00	1.00
2	0.77	0.33	1.00	1.00	1.00	1.00	1.00
3	0.69	1.00	0.71	1.00	0.60	1.00	1.00
4	0.46	1.00	1.00	1.00	1.00	1.00	1.00
5	0.52	0.33	1.00	1.00	1.00	1.00	1.00
6	0.63	-0.33	1.00	1.00	1.00	0.60	1.00
7	0.80	1.00	-0.14	1.00	1.00	1.00	1.00
8	0.68	1.00	1.00	1.00	1.00	0.60	1.00
9	0.65	0.33	1.00	1.00	0.60	1.00	1.00
10	0.29	-0.33	0.43	1.00	1.00	1.00	1.00
A	0.73	-0.33	1.00	1.00	1.00	1.00	1.00
B	0.47	0.33	1.00	1.00	1.00	1.00	1.00
C	0.85	1.00	1.00	1.00	1.00	1.00	1.00
Mean	0.64	0.44	0.85	1.00	0.94	0.94	1.00

Note: Students 1 to 10 are female; students A to C are male.

Table 3

Coding scores for student verbal reports

	Subtest associated with each knowledge and skill category					
Student	S_1 (4 items)	S_2 (6 items)	S_3 (4 items)	S_4 (4 items)	S_5 (4 items)	S_6 (4 items)
1	3	5.5	4	4	4	4
2	3	6	4	4	4	4
3	4	6	4	4	4	4
4	4	5	4	4	4	4
5	4	6	4	4	4	4
6	3.5	6	4	4	4	4
7	4	5	4	4	4	4
8	3	5.5	4	4	4	4
9	3.5	6	4	4	4	4
10	2	4.5	4	4	4	4
A	3.5	6	4	4	4	4
B	2.5	6	4	4	4	4
C	4	6	4	4	4	4
Mean	3.38	5.65	4	4	4	4

Note: subtests 1, 3, 4, 5, and 6 have 4 items each; subtest 2 has 6 items.