Roduta Roberts, M., Alves, C. B., Chu, M-W., Thompson, M., Bahry, L. M., & Gotzmann, A.

Testing expert-based versus student-based cognitive models for a grade 3 diagnostic mathematics assessment.

**AUTHOR POST PRINT VERSION**

Roduta Roberts, M., Alves, C. B., Chu, M. W., Thompson, M., Bahry, L. M., & Gotzmann, A. (2014). Testing expert-based vs. student-based cognitive models for a Grade 3 diagnostic mathematics assessment**.** *Applied Measurement in Education, 27,* 173-195.

**Abstract**

The purpose of this study was to evaluate the adequacy of three cognitive models, one developed by content experts and two generated from student verbal reports for explaining examinee performance on a Grade 3 diagnostic mathematics test. For this study, the items were developed to directly measure the attributes in the cognitive model. The performance of each cognitive model was evaluated by examining its fit to different data samples: verbal report, total, high-, moderate-, and low ability using the Hierarchy Consistency Index (Cui, 2009) a model-data fit index. This study utilized cognitive diagnostic assessments developed under the framework of construct-centered test design and analyzed using the Attribute Hierarchy Method (Gierl, Wang, & Zhou, 2008; Leighton, Gierl, & Hunka, 2004). Both the expert-based and the student-based cognitive models provided excellent fit to the verbal report and high ability samples, but moderate to poor fit to the total, moderate and low ability samples. Implications for cognitive model development for cognitive diagnostic assessment are discussed.

*Keywords:* cognitive models, cognitive diagnostic assessments, mathematics, hierarchy consistency index, attribute hierarchy method

# Testing Expert-Based vs. Student-Based Cognitive Models for a Grade 3 Diagnostic Mathematics Assessment

Cognitive diagnostic assessment (CDA) is designed to measure a student's knowledge structures and processing skills so as to identify areas of cognitive strengths and weaknesses (Leighton & Gierl, 2007a; Mislevy, 2006). Recently, CDA has received increasing attention from researchers and educational stakeholders for its potential value in providing more formative information to support instruction and student learning. Today, some research in the development of CDAs can be described as having three stages. The first stage concerns the specification and validation of a cognitive model. The second stage uses a principled test design approach, to create items designed to measure the knowledge and skills specified in the cognitive model. The third stage involves the psychometric analysis of observed data.

The cognitive model can serve two purposes. First, a cognitive model provides the link between test score interpretations and cognitive skills. The test developer is in a better position to make valid and defensible claims about student performance in cognitive terms. Second, a cognitive model integrates cognitive and learning psychology with instruction. **For example, detailed information about a student's cognitive strengths and areas requiring improvement could be used to inform instruction with the purpose of improving student learning and performance.** Given these purposes, accurate specification of the cognitive model is important when developing CDAs.

Cognitive models can be created in different ways, by: (1) reviewing theoretical literature or conducting an expert task analysis (conceptualized as the top-down approach), (2) using verbal report data of students answering test items in the target domain (conceptualized as the bottom-up approach), or (3) combining the two approaches. Each approach has different

demands with respect to the test developer's time and resources for developing cognitive models. Regardless of the approach used, previous studies (Gotzmann, Roberts, Alves, & Gierl, 2009; Gotzmann and Roberts, 2010; Leighton, Cui, & Cor, 2009) have shown that a single cognitive model often does not explain the performance of different subgroups of examinees equally well and that multiple models may be warranted. **For example, Gotzmann et al. (2009) found that an expert-based cognitive model that fit reasonably well for a random sample of 5000 examinees, did not predict the performance of different ethnic groups within the same sample. Given this result, the authors recommended multiple models be developed, specific to subgroups, whether it be derived by expert judgment or student verbal reports.** Methodologically, one limitation of these studies is the retro-fitting of existing test items, not originally developed for the purposes of a CDA, to a cognitive model which was developed post-hoc. To date, there are no research studies that compare the performance of expert-based and student-based cognitive models for explaining examinee data where the CDA in question employs principled or construct-centered test development procedures.

**Purpose of the Study**

The purpose of this study was to extend previous research by evaluating the adequacy of three cognitive models, one developed by content experts and two generated from student verbal reports, for explaining examinee performance on a Grade 3 diagnostic mathematics test, where the items are developed to explicitly measure the knowledge and skills in the cognitive model. More specifically, this study compared the model-data fit indices calculated for each model (i.e., expert-based and student-based) with five different data samples (i.e., overall sample, verbal report sample, high ability, moderate ability, and low ability groups). The value of the model-

data fit index provides a source of information on how well the cognitive model accounts for student performance for a given sample of students.

This paper is structured into four sections. In the first section an overview of the conceptual framework, context for the study and research hypotheses are provided. In the second section, the method used to conduct the study is described. In the third section, the student-based cognitive models and results of the psychometric analyses are presented. Finally, in the fourth section, implications of the results are discussed and conclude the paper.

## Section 1: Conceptual Framework and Background

### Conceptual Frameworks for Creating Cognitive Diagnostic Assessments

Recognizing the limitations of early educational testing practices, Snow and Lohman (1989) argued how developments in cognitive psychology could serve to positively inform psychometric practice. Creating assessments that are grounded in substantive cognitive theory should yield inferences that are more interpretable, meaningful, and valid. The cardinal feature of these assessments was that the substantive assumptions regarding the processes and knowledge structures used by the examinee, how these processes and knowledge structures develop and how they differ between more competent and less competent examinees were made explicit.

Current development frameworks for CDA include Evidence Centered Design (ECD; Mislevy, Steinberg, & Almond, 2003), and assessment engineering (AE; Luecht, 2006). ECD is a framework for test development that is analogous to developing an assessment argument, where claims about a student's performance require evidence to support it. By clearly defining and explicitly linking all processes of test development from domain analysis to assessment delivery, the outcomes of an ECD assessment provide the evidence needed to defensibly support inferences about student performance. AE is another framework where engineering-based

principles are used to direct the design and development of assessments so that the results can be meaningfully grounded in the intended measured construct and thus promote valid interpretations of performance. Both frameworks are considered to be construct-centered approaches to test development.

**Developing Cognitive Models for Educational Assessments**

In the case of CDA, specification of the targets of measurement, which are the knowledge, skills, and abilities, should be identified and operationalized in the form of a cognitive model before proceeding to test development. Leighton and Gierl (2007b) define 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 explanation and prediction of students' performance" (pg. 6). For CDA, the cognitive model should be specified in sufficient detail to allow for fine-grained inferences about student performance. It has been previously argued by researchers that the partnership of cognitive psychology and educational measurement is both logical and mutually beneficial since students invoke specific cognitive processes when answering an item (Embretson & Gorin, 2001; Glaser, 2000; Mislevy, 2006). The contribution of cognitive psychology to developing educational tests lies in the method for developing cognitive models that captures both the structure and process of students' knowledge and skills. The *cognitive model of task performance* is argued as being most appropriate when developing CDAs, in order to generate specific and defensible claims about student performance (Leighton & Gierl, 2007).

Theories of task performance can be used to derive cognitive models of task performance in a subject domain. However, the availability of these theories of task performance and

cognitive models in education are limited. Therefore, other means must be used to generate cognitive models. One method is the use of a task analysis of representative test items from a subject domain. A task analysis, or the top-down approach, represents a hypothesized cognitive model of task performance, where the likely knowledge and procedures used to solve the test item are specified. This task analysis is usually completed by a content expert or an individual who is familiar with the domain of interest. The bottom-up approach is another method which involves students thinking aloud as they solve test items to identify the actual knowledge, processes, and strategies elicited by the task (Chi, 1997; Ericsson & Simon, 1993). A cognitive model derived using a task analysis can be validated and, if required, modified using student verbal reports collected from think-aloud studies.

**Empirical Studies Evaluating Cognitive Models**

As argued by Leighton et al. (2009), the test developer's investment of time and resources may outweigh the benefits of developing cognitive models for use in educational tests. For this reason, cognitive models developed from a content expert task analysis are more common than those derived from student verbal reports. Previous studies have shown that the fit of the models developed by context experts to observed examinee data can be reasonably good. **The cognitive model specifies the knowledge, skills, and processes needed to solve a set of test items and the interrelationships among them. Expected examinee response patterns are derived from the cognitive model. The observed examinee response patterns are obtained upon administration of a set of items created using the cognitive model as a blueprint. The model fit index serves as an indicator of how well the observed examinee response patterns match with those predicted by the cognitive model. Better model fit indices allow for the interpretation of the cognitive model accurately representing examinee responses.**

Gotzmann et al. (2009) found that the expert-based model demonstrated good model-data fit for an overall sample of 5000 examinees, but did not predict the performance of sub-samples of examinees representing different ethnic groups well. In a follow up study, Gotzmann and Roberts (2010), observed a similar outcome when the data for low-, moderate-, and high-ability examinees were fit to the expert-based model, when the model showed good model-data fit for the overall sample.

In a study by Leighton et al. (2009), they found that the expert-based model fit the data reasonably well across the three samples of students writing the SAT. The samples differed in sample size and ability level (i.e., 21 moderate-high ability students, random sample of 5000 examinees, and a random sample of 100 moderate-high ability examinees). However, their study also compared the performance of a student-based cognitive model to the expert-based model for the same three samples. The student-based model based on verbal reports of 21 moderate-high students accounted for the data of the moderate-high samples better than the expert-based model. A limitation of the Leighton et al. study is that the expert-based and student-based cognitive models were based on the knowledge of one content expert and one preservice teacher, respectively.

The results of these studies suggest that one cognitive model derived by experts or developed from student verbal reports does not accurately represent the knowledge structures and processing skills of certain subgroups of examinees. However, a limitation of both studies is that examination of model-data fit was completed using items (oftentimes only one item per attribute) that were retro-fitted to a cognitive model that was developed post-hoc. That is, the items were not developed to specifically measure the knowledge and skills in the expert-based model. The retro-fitting approach to cognitive model development is less than optimal because

the generated model will be constrained by the attribute specifications that happen to occur among the existing items (if it exists, at all). Consequently, retro-fitted models might not accurately represent students' knowledge and skills required for mastering a certain domain (Gierl & Cui, 2008).

Given the limitations stated previously, the purpose of this study was to extend previous research by evaluating the adequacy of three cognitive models, one developed by content experts (hereafter referred to as the *expert-based cognitive model*) and two generated from student verbal reports (hereafter referred to as the *student-based cognitive models*), for explaining examinee performance on a Grade 3 diagnostic mathematics test, where the items are developed to explicitly measure the attributes, in the cognitive model. The performance of each cognitive model was evaluated by examining its fit to different data samples using a model-data fit index. This study utilized CDAs developed under the framework of construct-centered test design and analyzed using the Attribute Hierarchy Method (AHM; Gierl, Wang, & Zhou, 2008; Leighton, Gierl, & Hunka, 2004). The AHM is a cognitively based psychometric procedure used to classify examinees' test item responses into a set of structured attribute patterns associated with a cognitive model of task performance. Attributes include different procedures, skills and/or processes that an examinee must possess to solve a test item. The AHM provides a framework for designing diagnostic items based on the cognitive model, thus linking students' test performance to specific inferences about knowledge, and skill acquisition.

**Hypotheses.** Based on the results of studies by Gotzmann et al. (2009), Gotzmann and Roberts (2010), and Leighton et al. (2009), the following hypotheses will be tested:

(1) The student-based model will outperform the expert-based model when accounting for the verbal report sample.

(2) There will be a difference between the expert-based and student-based models when accounting for the total sample.

(3) There will be a difference between the expert-based and student-based models when accounting for the high ability sample.

(4) There will be a difference between the expert-based and student-based models when accounting for the moderate ability sample.

(5) There will be a difference between the expert-based and student-based models when accounting for the low ability sample.

### Operational Context: The Cognitive Diagnostic Mathematics Assessment Project

The context for this study is the Cognitive Diagnostic Mathematics Assessment (CDMA) project. CDMA is a curriculum-based set of assessments that can be used throughout the school year to measure students' thinking and learning in mathematics. The goal of the project was to create tests that provided teachers with diagnostic information so students' cognitive mathematical knowledge and skills can be identified and evaluated. The online computer-based administration system included the assessments and score reports. Construct-centered test design procedures in the context of the AHM as a form of CDA were used to create online diagnostic tests were in four content areas: (a) Number, (b) Patterns and Relations, (c) Shape and Space, and (d) Statistics and Probability at two grade levels, 3 and 6. Development of CDMA in Grade 3 began in 2008 and concluded in 2011.

## Section 2: Method

### Participants

The total sample was composed of Grade 3 students from a Western Canadian province. One data sample was a convenience sample composed of 295 students who wrote the online

diagnostic Mathematics assessment in 2010. **This sample was part of a larger assessment initiative with students writing from across the province in public, private, and separate school systems. Schools were located in both rural and urban centres.** The second data sample was composed of 18 Grade 3 students from a Western Canadian school district who participated in the think-aloud portion of the study. These 18 students were sampled from an intact classroom where the teacher volunteered to participate in research projects approved by the district. Collection of the verbal report or think-aloud data occurred in the fall of 2011, once the topic had already been introduced and taught in the classroom. Ethical protocols for data collection were approved and adhered to.

Based on previous studies, five data samples were used for analysis in this study: (1) verbal report student sample, (2) total examinee sample, (3) low ability, (4) moderate ability, and (5) high ability student samples. The low-, moderate- and high- ability samples were created using the total test sample via cluster analysis.

**Instrument**

A 24-item diagnostic assessment was created to measure one expert-based cognitive model. Each attribute (i.e., 8 attributes in total) in the cognitive model was measured by three items. The diagnostic assessment was a computer-based administration. Examples of items from the diagnostic assessment are provided alongside the expert-based model in the results section.

**Procedure**

The study was conducted in three stages. In the first stage expert-based cognitive models were developed as part of an earlier research study (Gierl, Alves, and Taylor-Majeau, 2010). In the second stage, the student-based cognitive model was developed using think-aloud methods. Verbal reports were collected for a sample of Grade 3 students in the same Western Canadian

province and analysed using verbal analysis to develop the student-based model. In the third stage, model-data fit analyses were conducted for both the expert- and student-based models on five different data samples.

**Stage one: Developing expert-based cognitive models.** Model development was directed by three content specialists: an exam manager who oversaw the CDMA project and two examiners with elementary mathematics teaching experience ranging from 15 to 32 years. During this stage, the exam manager created the preliminary cognitive models using the curriculum source documents, teaching experience, and expert judgment to identify the relevant knowledge, skills, and processes (i.e., attributes) to be measured by the test. Then, the two examiners and a panel of seven Grade 3 content specialist teachers scrutinized the proposed cognitive models for their wording, content and ordering of the attributes. The content specialists were experienced classroom teachers with teaching experience ranging from 13 to 30 years. The teachers were required to be familiar with all aspects of the curriculum because mathematics curriculum documents and their expert judgment were used to guide their decisions when organizing the attributes into a cognitive model. This panel of content specialists, hereafter referred to as *content experts* also had experience in developing large-scale tests and were familiar with best practices in large-scale item and test construction.

Throughout the development process, the content experts were instructed to ensure that the models be written at a fine-grain size, ordered by complexity from simple to complex, the attributes contain measurable skills, and the skills in the hierarchy be instructionally relevant. The first characteristic, grain size, requires that the skills specified in the model are written at a level of specificity that allows us “to provide examinees with information concerning whether or not they have mastered each of a group of specific, discretely defined skills, or attributes”

(Huebner, 2010, pg. 1). A second characteristic is the hierarchical ordering of the skills in the cognitive model. Often, a cognitive model reflects a hierarchy of ordered skills within a domain as cognitive processes share dependencies and function within a much larger network of inter-related processes, competencies, and skills (Gierl et al., 2009). The third characteristic concerns the measurability of the skills, meaning skills must be described in a way that would allow a test developer to create an item to measure each skill. That is, if the skill is not stated clearly, it does not entail an observable outcome, and it is difficult to operationalize using a test item. The fourth characteristic is the instructional relevancy of the skill.

Through discussion and consensus, a final set of cognitive models were developed for each strand in the mathematics curriculum. Items were written for each attribute in the cognitive models and then administered to a sample of Grade 3 students across the province. For more detail on the procedures used to create the cognitive models and test items, the reader is referred to Gierl, Alves, and Taylor-Majeau (2010).

**Stage two: Developing the student-based cognitive model.**

***Collecting the verbal report data.*** Research to date on the use of verbal reports in assessment of mathematics problem solving in elementary school-aged children has been sparse (Robinson, 2001). Much of the education research utilizing think-aloud methodology has been in other subject areas such as spelling (Steffler, Varnhagan, Friesen, & Treiman, 1998), reading (Pressley & Afflerbach, 1995) or scientific reasoning (Klahr, Fay, & Dunbar, 1993). However, Robinson's (2001) study provided important evidence supporting the use of think-aloud methods for generating valid verbal reports with children as young as six years old in the subject area of mathematics.

A sample of 18 students (12 male and 6 female) were recruited for the verbal report study. These 18 students represented a range of mathematics proficiency. To ensure that adequate verbalizations were generated, the students were recruited from a classroom whose teacher encouraged students to think-out aloud when reasoning through math word problems in class.

Using standard protocol analysis procedures (Ericsson & Simon, 1993), students were individually interviewed in a quiet, semi-private space and asked to think-aloud as they solved items from the diagnostic test. Concurrent verbalization provided information on what knowledge and concepts the student was attending to and working with at the time of problem solving. Students were provided with six questions, three on paper and three on the computer, to practice thinking aloud prior to the actual think-aloud session using the diagnostic test items. Not all six questions were completed if the student felt he or she was ready to proceed with the actual assessment. Standard probes such as "keep talking" were used if the student was silent for longer than 20 seconds. After the students completed the items in the test, they were asked to recall retrospectively how they solved each item. Information collected retrospectively about the knowledge and skills used to solve the item serves as an internal check by which to compare the verbalizations collected *while* the student was solving the item. As with the original administration of the diagnostic assessment, there was no time limit for completing the assessment. Each session lasted anywhere between 30 minutes and 1 hour and 10 minutes in length, including the practice questions, with the average session lasting 45 minutes.

***Coding the verbal report data and creation of the student-based cognitive models.*** The interviews were audio-recorded and transcribed. The verbal reports were reduced to facilitate coding and then coded by item. Where coding of the concurrent report was not possible due to

lack of verbalization, the retrospective report was used to inform the knowledge used at the time when solving the given problem.

The verbal reports were analyzed using verbal analysis (Chi, 1997). Three raters in total analyzed the transcripts to develop the final model. Two raters with experience in collecting verbal reports and familiarity with assessment of mathematics initially reviewed the transcripts to identify the key problem solving strategy, knowledge, and/or skills used by the student to solve the items. These two individuals coded 6 out of the 18 transcripts independently to come up with their own set of preliminary codes where the codes reflected a substantive summary of the actual knowledge or skill the students used when solving a particular problem. Coding and writing of the attribute descriptors for the student-based model were done without reference to the expert-based model. Then, the two coders came together to discuss and compare codes. Any discrepancies on the knowledge and skills used to solve the problem as presented in the transcript were discussed, debated, and a consensus reached. A set of working codes was developed based on coding of the common set of six transcripts. At this point, one of these two raters recoded the six transcripts and coded the remaining 12 using the code.

Another coder, who was also familiar with verbal reports and mathematics, was given the set of common codes. This coder coded all 18 transcripts independently. Inter-coder agreement was calculated for a random sample of six transcripts (33% of the total number of transcripts). The initial inter-coder agreement for the six transcripts was 90% then increased to 100% after further discussion and debate regarding discrepancies.

Two student-based cognitive models were created based on the knowledge and skills identified within the verbal reports across the 18 students. Each cognitive model was created across the 18 students because most of the students demonstrated similar solution paths to the

correct answer. Evidence of whether the student used a previously identified knowledge and skill located lower in the hierarchy while solving a given item directed the ordering of the attributes. For example, consider a student solving an item measuring attribute 4. If the student verbalized knowledge that was measured by attribute 3 then that knowledge and skill is a precursor to the knowledge and skills verbalized while solving an item measuring attribute 4. Judgment was also used to examine the logical ordering of the attributes. For example, the decision to keep addition and subtraction of facts as one attribute or split into two attributes was a judgment informed by experience with teaching elementary school level mathematics. Finally, the mean p-values across the sets of items measuring an attribute also informed the creation of the hierarchy. It was assumed that higher p-values corresponded to attributes lower in the hierarchy and lower p-values corresponded to attributes higher in the hierarchy.

Once the student-based cognitive models were constructed (see Results: Student-Based Cognitive Models), the 24 test items were aligned to each of the attributes in the model. For the first student-based model, one attribute in student-based model was measured by two items, another attribute in the model was measured by four items, and the remaining attributes were measured by three items. For the second student-based model, three attributes were measured by two items and the remaining attributes were measured by three items. The decisions to align particular items to their attributes was completed on a logical basis by two raters and then independently by a third rater. There was 100% agreement among the three raters with the alignment of the items to each of the attributes in both student-based cognitive models.

**Stage three: Data analysis and summary.** After specification of the expert-based and student-based cognitive models, an analysis of student response data from the administration of the diagnostic Mathematics test was completed using the AHM. The analysis included item

analysis and calculation of a model-data fit index called the Hierarchy Consistency Index (HCI; Cui & Leighton, 2009) for each cognitive model (3) with each data sample (5) for a total of 3 X 5 = 15 HCI values. The AHM and HCI analyses were completed using code developed with Mathematica 6.0.

The HCI is a person-fit statistic that can provide meaningful information on the fit of observed student response patterns relative to expected response patterns derived from the cognitive model.

The HCI for examinee $i$ is calculated as follows:

$$HCI_i = 1 - \frac{2\sum_{j \epsilon S correct_i} \sum_{g \epsilon S_j} X_{ij}(1 - X_{ig})}{N_{ci}}$$

where,

$S_{correct\ i}$ includes items that are correctly answered by student $i$,

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

$S_i$ includes items that require the subset of attributes measured by item $j$,

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

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

The HCI values are calculated for each student and the mean or median value is taken across students for each cognitive model. **Because the HCI is a person-fit statistic, it is not affected by sample size**. HCI values range from -1 to +1 where an HCI of 0.6 or higher indicates moderate model-data fit and values greater than 0.80 indicate excellent model-data fit (Cui & Leighton, 2009). The mean or median and standard deviation of the HCI are indicators of overall model-data fit. Low HCI values indicate a large discrepancy between the observed examinee

response patterns and the expected examinee response patterns. Because the HCI is calculated for each student, students who do not fit the model (i.e., HCI < 0.6) can be identified. In addition, cognitive models that have good model-data fit overall can be evaluated for several subgroups to ensure valid application of the cognitive model for all examinees.

The first group consisted of 18 Grade 3 students taking the verbal reports in the Fall 2011; this group is hereafter called *Verbal Report*. The second group consisted of a total of 295 Grade 3 students answering the test online in the Fall 2010; this group is called *Total*. These 295 students were also split into three different level of mathematic ability (*High, Moderate, and Low*), representing the three other investigated groups. The three groups were split using the K-means clustering method. The goal of the K-means algorithm is to find the best division of *n* entities (i.e., the total number of students according to their total score) into *k* groups (i.e., three groups), so that the total distance between each group's members and its corresponding centroid, representative of the group, is minimized (Lin, Koh, & Chen, 2010). Information concerning the score range and the number of students participating in each of the five groups is presented as follows in Table 1.

[Insert Table 1]

## Section 3: Results

In this section, we present the expert-based cognitive model and the student-based cognitive models followed by the psychometric analyses. We conclude this section by summarizing our evaluation of the five hypotheses.

**Expert-Based Cognitive Model**

Figure 1 shows the expert-based cognitive model with eight attributes organized in a linear hierarchy. Attribute 1, *express number words in symbolic form,* is the most cognitively simple whereas attribute 8, *represent a number in more than one way,* is the most cognitively complex. Figure 2 shows an example of a test item that was developed to measure attribute 1 in the model. Figure 3 shows an example of another test item developed to measure attribute 8 in the model. Both items illustrate the difference in the level of cognitive complexity as reflected in the cognitive model which guided its development.

[Insert Figure 1]

[Insert Figure 2]

[Insert Figure 3]

**Student-Based Cognitive Models**

Figure 4 shows the first of two student-based cognitive models created from the student verbal reports. This cognitive model also has eight attributes organized in a branching hierarchy. Attribute 1, *recognize numerical form of written numbers,* is the most cognitively simple whereas attribute 8, *compare two numbers to establish equivalence,* is the most cognitively complex. This model is similar to the expert-based model where it does have a linear hierarchical structure however there are two branches (attributes 3 and 5).

[Insert Figure 4]

Figure 5 shows the second of two student-based cognitive models created from the student verbal reports. This cognitive model has nine attributes organized in a hierarchy with both convergent and divergent branches. The major difference for this second student-based model is the splitting of attribute 7 in student-based model #1 into two separate attributes (i.e., 7 and 8) in student-based model #2. Both models were considered to be equally viable options based on an analysis of the verbal report data.

[Insert Figure 5]

**Psychometric Analysis**

In this section psychometric results for the three cognitive models, one developed by content experts and the two created from student verbal reports, are presented. This section is organized into two parts. In the first part we present the characteristics of the items, including item difficulty and item discrimination. In the second part we discuss the fit of the models relative to the observed student response vectors of the five groups using the Hierarchy Consistency Index (HCI).

**Group and Item Characteristics**

Group performance on the 24 item mathematics test is summarized in Table 2. The highest mean student performance (with standard deviation in parentheses) was 20.94 (SD = 3.10) for the *Verbal Report* group. As expected, the smallest mean student performance was found on the *Low* ability group, 6.83 (SD=2.49).

[Insert Table 2]

Table 3 summarizes the difficulty and discrimination level of each item measured as the percentage of correct answers (hereafter called p-values) and biserial correlation. The biserial computed over the combined samples is presented, instead of for each group, due to the potential for restriction of range.

[Insert Table 3]

As shown in Table 3, in general, for the four first groups–Verbal Report, Total, High, and Moderate–a pattern of higher p-values on the initial items (which measure the most basic skills) and lower p-values on the last items (which measure the most complex skills) is observed. This pattern serves as one indicator of the alignment between test items and the cognitive model. The highest mean item difficulty was found for the Verbal Report group (M=0.87), followed by the High ability group (M=0.84). Overall, the items seem highly discriminative (average biserial = 0.58) and, except for item 24 (Biserial=0.19), all items have acceptable[1] biserial correlations. As shown in Table 3 the lowest p-values were for items measuring Attribute 8, and the highest p-values were for items measuring Attribute 1. Because Attribute 8 is believed to be most complex attribute according to the conceptualized cognitive models, and Attribute 1 the most basic and pre-requisite attribute, these results were expected.

**Overall, we judged the content of the items to align well with the attributes in the cognitive model. However, some items appeared to have p-values inconsistent with the other items measuring the same attribute. These items were inspected to investigate the**

[1] Biserial correlation higher than 0.25–0.30 can be deemed acceptable for differentiating examinees (Alberta Education, 1999).

**potential reasons for misalignment. One reason could have been the small differences in the wording of the response options. For example, the Total p-value for Item 5 was 0.36 but for Items 4 and 6 the corresponding p-values were 0.63 and 0.67 respectively. Upon inspection, the pattern of alternatives for Item 5 were different from Items 4 and 6. Item 5 is the only item that contained two alternatives that did not use the word "hundred" in the response options. The remaining alternatives were chosen more frequently by the lower ability students. Another possible reason for p-value misalignment was the need for students to activate animation for presentation of response options. As an example, Item 8 showed a different pattern with p-values for the low ability group at 0.53 being higher than for Items 7 and 9 with p-values of 0.23 and 0.20 respectively. Upon inspection of Item 8, we discovered that the item could be answered without needing to play any animation. This small difference in presentation could have contributed to the item being easier relative to Items 7 and 9 for low ability students.**

The attribute difficulty, which is the average of the items that directly measure the corresponding attribute, is presented in Figure 6.

[Insert Figure 6]

By comparing the results among the three ability levels–High, Moderate, and Low–the percentage of correct responses is noticeably and consistently larger for the high ability students compared to the other groups. For all five groups, in general, the average difficulty slightly increased from one attribute to the other, except for Attributes 3 which seems easier than the preceding attribute. The same is true for Attribute 7 for the Total, High, and Moderate groups.

**Model-fit**

The HCI results were compared for the Verbal Report, Total, High, Moderate, and Low ability samples. Using the HCI, it is possible to investigate the degree to which an observed examinee response pattern is consistent with the specified cognitive model. The median, mean, and standard deviations of the HCI values for the expert-based cognitive model according to the five groups are summarized in Table 4.

[Insert Table 4]

Using the expert-based linear cognitive model (see Figure 2), a median HCI value of 0.89 was obtained for the Verbal Report sample of 18 students. Cui (2009) suggested that the median HCI values greater than 0.60 indicate moderate fit, whereas values greater than 0.80 suggest excellent fit between the students' response patterns and the expected response patterns based on the hierarchical relationship among attributes, as represented in the cognitive models. If Cui's guideline is considered, then the median HCI value (0.81) is considered excellent for the High ability group as well. In comparison, low median HCI values were obtained when the same expert-based cognitive model was fit to the observed response vectors of the Total and Low ability groups (median HCI=0.25 and -0.32, respectively).

[Insert Table 5]

As shown in Table 5, the highest median HCI values were obtained when the data from the Verbal Report and the High ability group were used (0.95 and 0.85, respectively). These

values indicate a strong fit between the cognitive model and the response data, suggesting that the hierarchical arrangement of attributes adequately predicted the response patterns of high ability students[2]. Conversely, the cognitive model poorly predicted the observed response vectors of the low ability students and the Total sample (median HCI = -0.26 and 0.29, respectively).

A median HCI value of 0.48 was obtained when the data for the Moderate group was considered. This median HCI value suggests that the fit between the cognitive model and the response data for this sample was weak to moderate.

[Insert Table 6]

For the second student-based model, Table 6 indicates a strong fit between the cognitive model and the response data obtained from the Verbal Report and Total sample of students (median HCI = 0.95 and 0.87, respectively). However, the cognitive model did not seem to predict response patterns for the Low ability students and Total sample well, with the median HCI being -0.25 and 0.31, respectively. For the sample of students with Moderate ability, a HCI value of 0.51 was found, suggesting that the fit between the cognitive model and the response data for this sample was moderate.

[Insert Figure 7]

Taking the five groups into consideration, the median HCI values suggest that the model-data fit was satisfactory for the Verbal Report and High ability group for the three types of

[2] As the average performance of the students participating in the Verbal Report procedure was slightly above the average performance of the High ability group (see Table 2), Verbal Report group can be considered to perform similarly to high ability students.

cognitive model, Expert- and Student-based #1 and #2. Although, the HCI values are similar, they favor the Student-based model, more so with Student model #2.[3]

Concerning the hypotheses of the study, a brief summary is provided together with the empirical evidence supporting our conclusions.

(1) The student-based model performed better than the expert-based model for the verbal report sample, as hypothesized. The median HCI values for the student-based models (0.95) were higher than the median HCI for the expert-based model (0.89).

(2) There was a difference between the expert-based and student-based models when accounting for the Total sample as hypothesized. The median HCI values for the student-based models (0.29 and 0.31) were higher than the median HCI for the expert-based model (0.25).

(3) There was a difference between the expert-based and student-based models when accounting for the High ability sample as hypothesized. The median HCI values for the student-based models (0.85 and 0.87) were higher than the median HCI for the expert-based model (0.81).

(4) There was a small difference between the expert-based and student-based models when accounting for the moderate ability sample. The median HCI values for the student-

[3] **We conducted supplementary analyses as suggested by one reviewer and dichotomized the data sample into high- and low- ability students. We then fit each data sample to the expert-based and student-based models. Overall, we observed reduced model-data fit and increased variability of the HCIs for the high-ability group across both expert and student models. The overall model-data fit and variability of HCIs did not change much with the low ability group. We interpret this finding as the high ability group having increased variability with regards to problem solving processes when splitting the sample into two groups instead of three. The low ability group remained relatively homogenous in their problem solving processes reflecting little change in overall model-data fit. We chose to keep our original analysis presentation, and discussion with three ability groups to provide a clearer demonstration of increasing model-data fit moving from low to high ability groups.**

based models (0.48 and 0.51) were slightly higher than the median HCI for the expert-based model (0.47).

(5) There was a difference between the expert-based and student-based models when accounting for the low ability sample as hypothesized. The median HCI values for the student-based models (-0.26 and -0.25) were slightly higher than the median HCI for the expert-based model (-0.32).

## Section 4: Summary and Discussion

The purpose of this study was to evaluate the adequacy of three cognitive models, one developed by content experts and two generated from student verbal reports for explaining examinee performance on a Grade 3 diagnostic mathematics test. For this study, the items were developed to explicitly measure the attributes in the cognitive model. The performance of each cognitive model was evaluated by examining its fit to different data samples: verbal report, total, high-, moderate-, and low ability using the HCI, (Cui, 2009) a model-data fit index. This study utilized CDAs developed under the framework of construct-centered test design and analyzed using the AHM (Gierl, Wang, & Zhou, 2008; Leighton, Gierl, & Hunka, 2004). Methodologically, this study extended upon Leighton et al.'s (2009), and Gotzmann and Roberts' (2010) studies by using data from an administration of a CDA where three items measured each attribute in the cognitive model instead of only one. In this way, a clearer evaluation of the performance of the expert-based and student-based cognitive models for explaining student performance could be done. Additionally, the use of a panel of content experts to create the expert-based model helped to bolster the external validity of the model.

**Results of the study showed that the student-based models and the expert-based model demonstrated a similar pattern of model fit across the five data samples. Both the**

**expert-based and the student-based cognitive models provided excellent fit to the high ability and verbal report samples, but moderate to poor fit to the total, moderate and low ability samples. Generally speaking, as ability increases, the model-data fit increases as well. The poor fit between the cognitive models and the responses of the low ability examinees is not surprising. The student-based cognitive models, as the point of reference, were generated from students who scored, on average, moderately high on the assessment. Therefore, the student-based model could be thought of as a model of thinking for moderately high ability students, for this diagnostic assessment. This assumption may substantially affect how well cognitive models, derived from students of one ability level, predict responses for students of other ability levels. Research on expertise has shown that expert and novice problem solvers differ in many ways not just in the amount of substantive knowledge, but also with how this knowledge is structured (Chi, Glaser, & Farr, 1988; Leighton, et al., 2009; Mislevy, 1994).**

**The student-based models shared similar attributes with the expert-based model however, the *structures* of the student-based models were different than the expert-based model (i.e., branching vs. linear cognitive model). An improvement in the model-data fit index with the branching hierarchies suggested that a more complex cognitive model structure may be a more accurate reflection of how these knowledge and skills are organized. Having the student models differ from the expert-based models is consistent with what is predicted by theories of expertise where these differences are demonstrated in studies using verbal reports comparing performances of novices and experts (Ericsson & Simon, 1993; Pressley & Afflerbach, 1995). However, this study did not employ think-alouds with the expert teachers to generate the expert-based**

**model was created through consensus of experts' judgements, based on knowledge of the curriculum and extensive teaching experience, to identify the relevant the knowledge and skills and their ordering when creating the cognitive model. Therefore, the expert-based model could be conceptualized as a hypothetical model of how students would likely solve problems and not necessarily how an expert solved the problem. The similarity of some of the attributes within the models suggest that expert teachers did identify the knowledge and skills students used while problem solving. Following this, an interpretation could be made that these students engaged in thinking and problem solving using knowledge and skills that are consistent with what was predicted by the content expert teachers but within a different and more complex model structure. This finding can help explain how both expert-based and student-based models had the same patterns of fit across the data samples, but why the student-based models performed better that the expert-based models, especially for the high ability group.**

Given that one cognitive model does not represent subgroups of student performance equally well and that more truthful representations of student cognition can be created from student verbalizations, the recommendation to use students from a particular subgroup to create cognitive models to represent that subgroup is logical. In the context of this study, the expert-based model did not predict these subgroup performances very well. Examining the composition of the high, moderate, and low examinees in the total sample, this finding may not be surprising. 45% of the total sample was classified as low ability, 35% were classified as moderate ability, and only 20% were classified as high ability. It is important, however, to temper the development of cognitive models for the low and moderate ability groups with what is feasible to accomplish with respect to model development. Lower ability examinees often engage in "buggy" thinking,

may have misconceptions, or have no strategy at all (i.e., guessing; Tatsuoka & Tatsuoka, 1983). This variability in knowledge, skill, and strategy poses a significant challenge when considering development of student-based cognitive models. There is a great benefit to overcoming these challenges in being able to form stronger validity arguments when making inferences about what students know and are able to do.

**Our work with creating the cognitive models has parallels with Gagne's analysis of behavioral objectives (Gagne, 1968, 1977). Gagne speaks to the use of an information processing task analysis to identify the sequence of knowledge and processes used when problem solving and a learning task analysis to identify essential and supporting prerequisite knowledge and skills. The development of the expert-based models were derived, in part, using these forms of task analysis. These models are hypothetical representations of the cognitive processes students use when solving a set of items until empirically validated using student verbal reports. Our work differs from the work of Gagne in two ways. First, in the context of assessment and validation, our cognitive models serve as an inferential bridge that allows us to make valid and defensible claims about a student's cognitive strengths and weaknesses. In order to do this, the cognitive model underwriting the assessment must have empirical psychological evidence to support it (Leighton and Gierl, 2007b). That evidence is can be provided with student verbal reports. Second, it has been argued that the provision of empirical psychological evidence is a requirement, not an option as written by Gagne, when opting to use cognitive models in CDA (Leighton and Gierl, 2007b).**

**Our work with cognitive models, more specifically, and CDA more generally, acknowledges the importance of aligning curriculum content, instructional objectives, and**

**assessment (Gagne, 1962). Gagne's work sharing similarities with some procedures used with CDA and cognitive models, shows one way to link two areas for the benefit of supporting student learning.**

One limitation of this work is the potential for bias due to convenience sampling. **Fitting the models to data collected from a sample purposefully constructed to represent all Grade 3 students in the province could result in variation in the HCIs.** Another potential limitation of this study is the expert-based cognitive models, although the work of a panel of experts, nevertheless represents one set of models out of a potential number of other viable sets of models. This limitation could be overcome in future studies by using two panels of expert teachers and integrating their judgments in a systematic way when creating the cognitive models. **Last, at this time, there is no objective way to test for the significance of the HCI statistic (Y. Cui, personal communication, July 3, 2013). However, the use of the HCI in this study has provided a basic mechanism for comparing models and generating insight into the relative capability cognitive models created by experts and by students for explaining student performance.**

There are at least three lines of future research that can be conducted to extend this study. One future study can involve a comparison of expert-based and student-based cognitive models in a domain other than Mathematics. Process models in a domain such as mathematics may be more easily generated due to the linearity in some problem solving activities with certain groups of students. This kind of problem solving process may not be the case in domains such as reading or social studies. A second follow up research study should be conducted to investigate the feasibility and effectiveness of developing student-based cognitive models to represent performance of student subgroups. Development of the student-based cognitive model from

verbal reports requires selection of subjects who can verbalize their thought processes while solving a given task. The theory behind protocol analysis suggests that verbal reports may not capture thinking processes if there is mastery of a skill (i.e., automaticity in solving the problem) or if the cognitive load of the task is too great. The collection of verbal reports especially with low and low to moderate ability students and its usability for creating cognitive models in the context of CDA requires further study. Developing cognitive models for lower ability examinees is critically important, given that this is the intended target population of diagnostic assessment. A third follow-up study could examine possible modifications to estimating model-data fit by accounting for guessing in student responses. This modification could also assist with evaluating cognitive models more precisely in the absence of operational cognitive models for low ability examinees.

## Conclusion

This study can serve as an example of an integration of cognitive psychology and educational measurement and assessment right from the specification of the targets of measurement, to the development of the cognitive models, to the creation of the test items and analysis of student responses. At present, this study is one of a very few number of research studies employing the think-aloud method with elementary age students for the purposes of supporting test development. The cognitive models generated with this population could potentially serve as the basis for further research on theories of task performance in an academic domain. The results of this study can assist in the process of creating and evaluating cognitive models for use in diagnostic assessments. Moreover, the results of this study can provide increased understanding of how students think when solving math problems to test developers, teachers, and administrators to assist in student learning and teaching.

## References

Alberta Education (1999). *Alberta Education Annual Report 1998-1999*. Edmonton, AB: Alberta Education.

Alves, C. B. (2012). *Making Diagnostic Inferences about Student Performance on the Alberta Education Diagnostic Mathematics Project: An Application of the Attribute Hierarchy Method*. Unpublished Doctoral Dissertation. University of Alberta: Edmonton, Alberta, Canada.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of Learning Sciences, 6,* 271-315.

Chi, M. T. H., Glaser, R., & Farr, M. (Eds.). (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum.

Creswell, J. W. (2008). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research.* Upper Saddle River, NJ: Prentice-Hall.

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

Embretson, S. E. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp.107-135). New York: Plenium Press.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38,* 343-368.

Ercikan, K., and Roth, W-M. (2007). What good is polarizing research into qualitative and quantitative? *Educational Researcher, 35*(5), 14-23.

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

Gagne, R. M. (1962). The acquisition of knowledge. *Psychological Review, 4*(69), 355-365.

Gagne, R. M. (1968), Contributions of learning to human development. *Psychological Review, 3*(75) 177-191.

Gagne, R. M. (1977). Analysis of behavioral objectives. In L. J. Briggs (Ed.), *Instructional design: Principles and applications* (pp. 115-146). Englewood Cliffs, NJ: Educational Technology Publications.

Gierl, M. J., 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,* 318-341.

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research and Perspectives, 6(4),* 263-268.

Gierl, M. J., Leighton, J. P., Wang, C., Zhou, J., Gokiert, R., & Tan, A. (2009). *Validating cognitive models of task performance in algebra the SAT*© (Research Report 2009-3). New York: The College Board.

Gierl, M. J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *Journal of Technology, Learning, and Assessment, 6* (6). Retrieved December 01, 2010, from http://www.jtla.org.

Glaser, R. (Ed.). (2000). *Advances in instructional psychology: Educational design and cognitive science, Vol. 5*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Gotzmann, A., Roberts, M., Alves, C., & Gierl, M. J. (2009). *Using cognitive models to evaluate ethnicity and gender differences*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Gotzmann, A., & Roberts, M. R. (2010). *Do cognitive models consistently show good model-data-fit for students at different ability levels?* Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research & Evaluation, 15* (3), 1-7.

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology, 25,* 111-146.

Leighton, J. P., Cui, Y., & Cor, M. K. (2009). Testing expert-based and student-based cognitive models: An application of the Attribute Hierarchy Method and the Hierarchy Consistency Index. *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.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule Space Approach. *Journal of Educational Measurement, 41,* 205-237.

Lin, C., Koh, J., & Chen, A. L. P. (2010). *A Better Strategy of Discovering Link-Pattern Based Communities by Classical Clustering Methods*. In Proceedings of PAKDD, 1, 56-67.

Luecht, R. M. (2006). Engineering the test: From principled item design to automated test assembly. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Dallas, TX.

Mislevy, R. J. (1994). Test theory reconceived. (CSE Technical Report 376). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles. Retrieved from http://www.cse.ucla.edu/products/Reports/TECH376.pdf

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257-305). Washington, DC: American Council on Education.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-62.

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

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading.* Hillsdale, NJ: Erlbaum.

Robinson, K. M. (2001). The validity of verbal reports in children's subtraction. *Journal of Educational Psychology, 93(1),* 211-222. doi:10.1037//0022-0663.93.1.211

Snow, R. E. & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: American Council on Education, Macmillan.

Steffler, D. J., Varnhagan, C. K., Friesen, C. K., & Treiman, R. (1998). There's more to children's spelling than the errors they make: Strategic and automatic processes for the one syllable words. *Journal of Educational Psychology, 90,* 492-505.

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

Table 1

*Score Range and Number of Students for Each Group*

| Group | Score range | Number of students |
|---|---|---|
| 1. Verbal Report | 0-24 | 18 |
| 2. Total | 0-24 | 295 |
| 3. High | 18-24 | 60 |
| 4. Moderate | 11-17 | 103 |
| 5. Low | 0-10 | 132 |

Table 2

*Descriptive Statistics of the Mathematics Test for Each Group*

| | Verbal Report | Total | High | Moderate | Low |
|---|---|---|---|---|---|
| Mean | 20.94 | 12.06 | 20.08 | 14.11 | 6.83 |
| SD | 3.10 | 5.62 | 1.82 | 1.94 | 2.49 |
| Minimum | 14.00 | 0.00 | 18.00 | 11.00 | 0.00 |
| Maximum | 24.00 | 24.00 | 24.00 | 17.00 | 10.00 |
| N | 18 | 295 | 60 | 103 | 132 |

Table 3

*Item Difficulty and Discrimination for Each Group*

| Attribute | Item | Difficulty (p-value) | | | | | Discrimination |
|---|---|---|---|---|---|---|---|
| | | Verbal Report | Total | High | Moderate | Low | (Biserial) |
| A1 | Item 01 | 1.00 | 0.82 | 0.97 | 0.91 | 0.69 | 0.51 |
| A1 | Item 02 | 1.00 | 0.75 | 0.95 | 0.84 | 0.59 | 0.48 |
| A1 | Item 03 | 1.00 | 0.61 | 0.95 | 0.83 | 0.28 | 0.74 |
| A2 | Item 04 | 1.00 | 0.63 | 0.97 | 0.87 | 0.29 | 0.77 |
| A2 | Item 05 | 0.67 | 0.36 | 0.73 | 0.49 | 0.10 | 0.66 |
| A2 | Item 06 | 0.89 | 0.67 | 0.90 | 0.83 | 0.44 | 0.55 |
| A3 | Item 07 | 1.00 | 0.59 | 1.00 | 0.82 | 0.23 | 0.83 |
| A3 | Item 08 | 1.00 | 0.72 | 0.97 | 0.83 | 0.53 | 0.55 |
| A3 | Item 09 | 1.00 | 0.60 | 1.00 | 0.86 | 0.20 | 0.85 |
| A4 | Item 10 | 0.94 | 0.48 | 0.78 | 0.46 | 0.36 | 0.43 |
| A4 | Item 11 | 0.83 | 0.57 | 0.97 | 0.75 | 0.24 | 0.69 |
| A4 | Item 12 | 0.89 | 0.51 | 0.95 | 0.77 | 0.11 | 0.87 |
| A5 | Item 13 | 0.89 | 0.49 | 0.97 | 0.63 | 0.16 | 0.78 |
| A5 | Item 14 | 1.00 | 0.39 | 0.82 | 0.50 | 0.12 | 0.71 |
| A5 | Item 15 | 1.00 | 0.49 | 0.90 | 0.54 | 0.27 | 0.55 |
| A6 | Item 16 | 0.89 | 0.56 | 0.87 | 0.50 | 0.46 | 0.35 |
| A6 | Item 17 | 0.89 | 0.39 | 0.77 | 0.27 | 0.32 | 0.37 |
| A6 | Item 18 | 0.89 | 0.32 | 0.63 | 0.26 | 0.22 | 0.45 |
| A7 | Item 19 | 0.78 | 0.44 | 0.83 | 0.53 | 0.20 | 0.61 |
| A7 | Item 20 | 0.83 | 0.38 | 0.87 | 0.41 | 0.14 | 0.66 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| A7 | Item 21 | 0.83 | 0.46 | 0.90 | 0.51 | 0.21 | 0.64 |
| A8 | Item 22 | 0.56 | 0.24 | 0.45 | 0.20 | 0.17 | 0.36 |
| A8 | Item 23 | 0.72 | 0.23 | 0.42 | 0.18 | 0.18 | 0.32 |
| A8 | Item 24 | 0.44 | 0.35 | 0.53 | 0.30 | 0.31 | 0.19 |
| Grand-mean | | 0.87 | 0.50 | 0.84 | 0.59 | 0.28 | 0.58 |
| | SD | 0.15 | 0.16 | 0.17 | 0.24 | 0.16 | 0.19 |
| | N | 18 | 295 | 60 | 103 | 132 | 313 |

Table 4

*HCI Values for the Expert Cognitive Model by Group*

| | Verbal Report | Total | High | Moderate | Low |
|---|---|---|---|---|---|
| Median | 0.89 | 0.25 | 0.81 | 0.47 | -0.32 |
| Mean | 0.87 | 0.19 | 0.81 | 0.44 | -0.28 |
| SD | 0.12 | 0.54 | 0.14 | 0.25 | 0.39 |
| N | 18 | 295 | 60 | 103 | 132 |

Table 5

*HCI Values for Student Cognitive Model #1 by Group*

| | Verbal Report | Total | High | Moderate | Low |
|---|---|---|---|---|---|
| Median | 0.95 | 0.29 | 0.85 | 0.48 | -0.26 |
| Mean | 0.92 | 0.22 | 0.82 | 0.45 | -0.23 |
| SD | 0.08 | 0.53 | 0.14 | 0.25 | 0.41 |
| N | 18 | 295 | 60 | 103 | 132 |

Table 6

*HCI Values for Student Cognitive Model #2 by Group*

| | Verbal Report | Total | High | Moderate | Low |
|---|---|---|---|---|---|
| Median | 0.95 | 0.31 | 0.87 | 0.51 | -0.25 |
| Mean | 0.94 | 0.23 | 0.83 | 0.48 | -0.23 |
| SD | 0.07 | 0.54 | 0.14 | 0.26 | 0.41 |
| N | 18 | 295 | 60 | 103 | 132 |

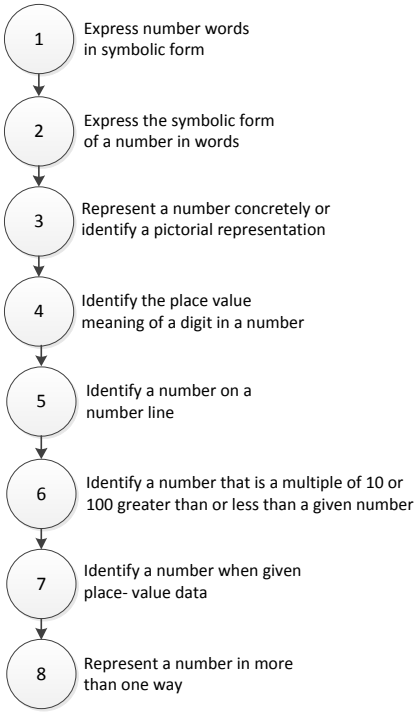
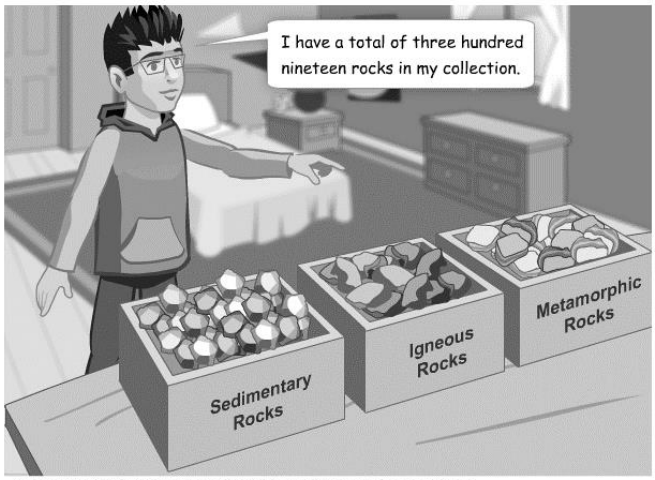



*Figure 1*. The expert-based attribute hierarchy and skill descriptors for the strand of Number: Develop number sense.

3.1.2.1.B




**The number three hundred nineteen can also be shown as**

**A.** 309
**B.** 319
**C.** 390
**D.** 391

*Figure 2.* An item measuring attribute 1, express number words in symbolic form, in the expert-based cognitive model.

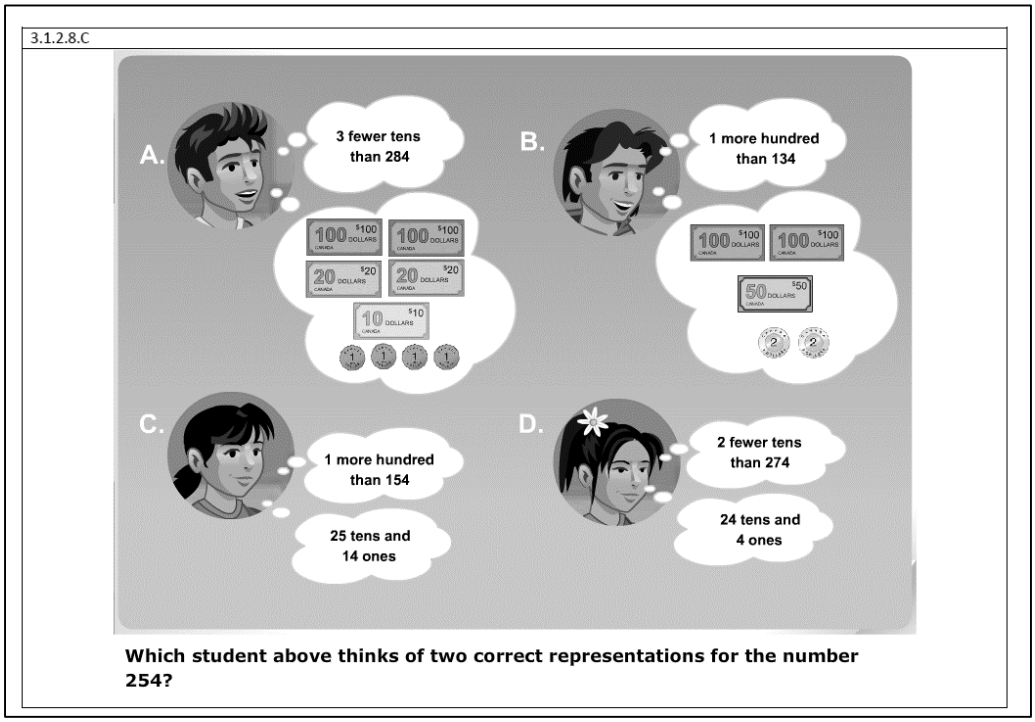



*Figure 3.* An item measuring attribute 8, represent a number in more than one way, in the expert-based cognitive model.

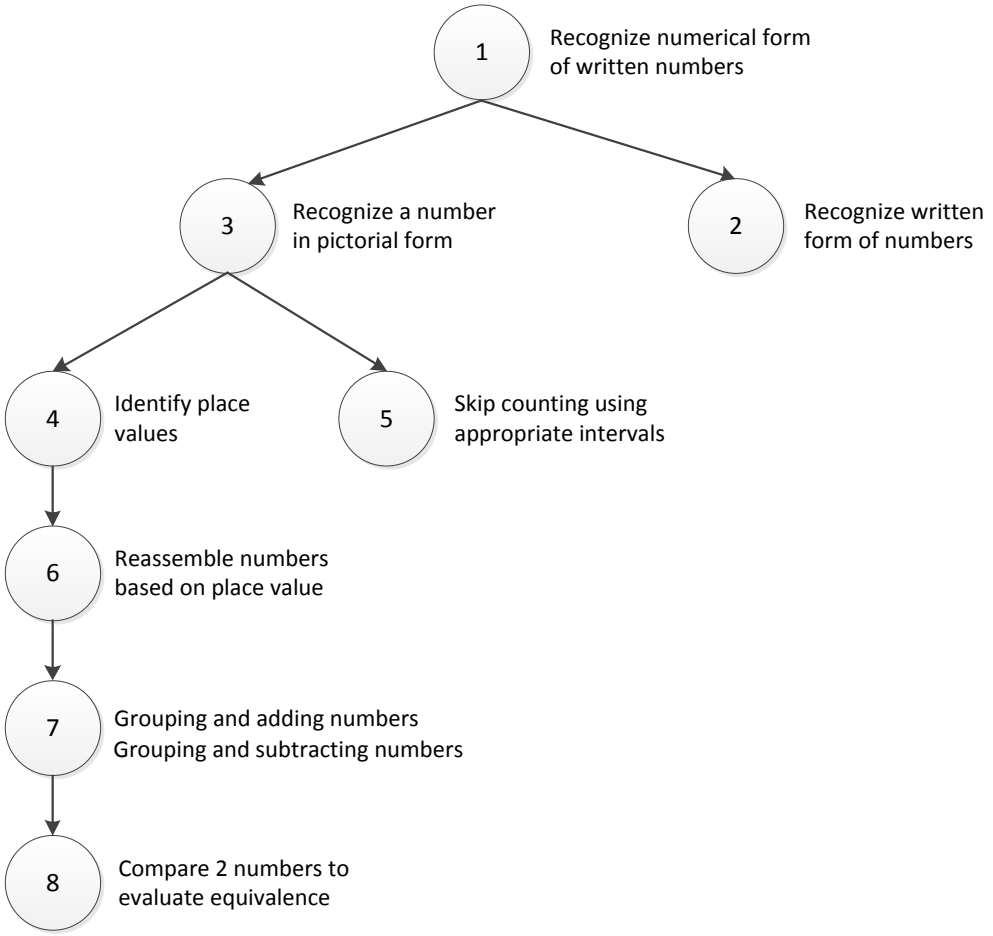



*Figure 4.* The student-based attribute hierarchy #1 and skill descriptors for the strand of Number: Develop number sense.

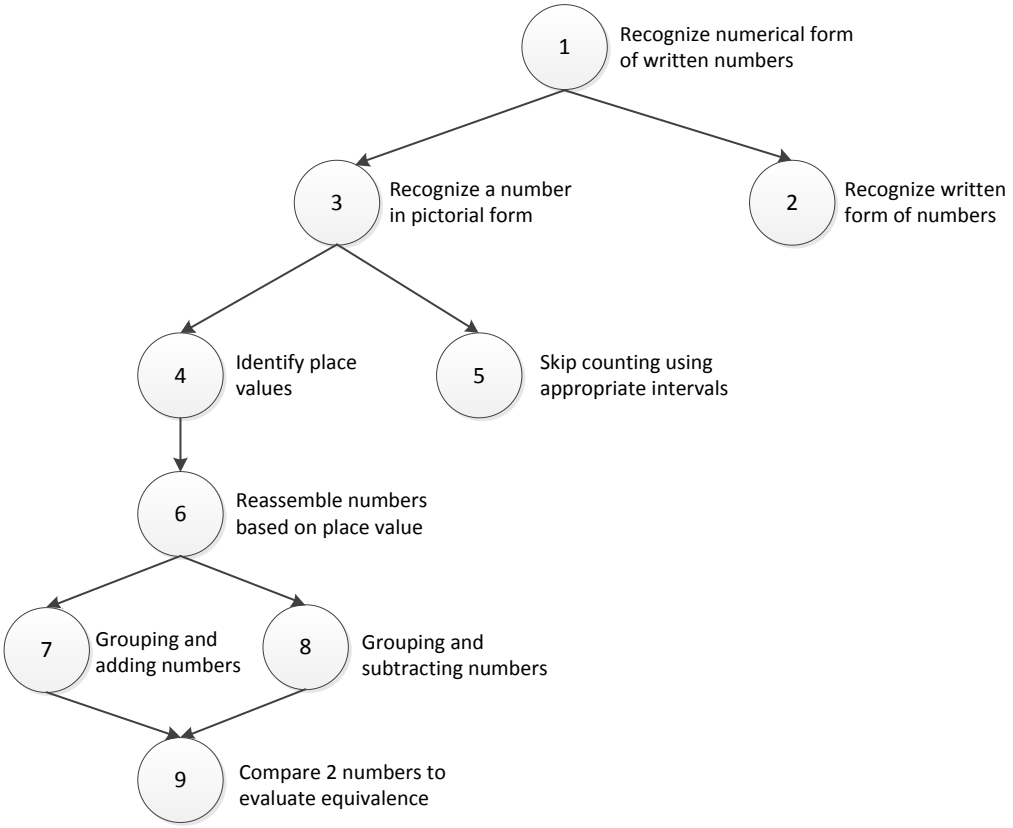



*Figure 5.* The student-based attribute hierarchy #2 and skill descriptors for the strand of Number: Develop number sense.

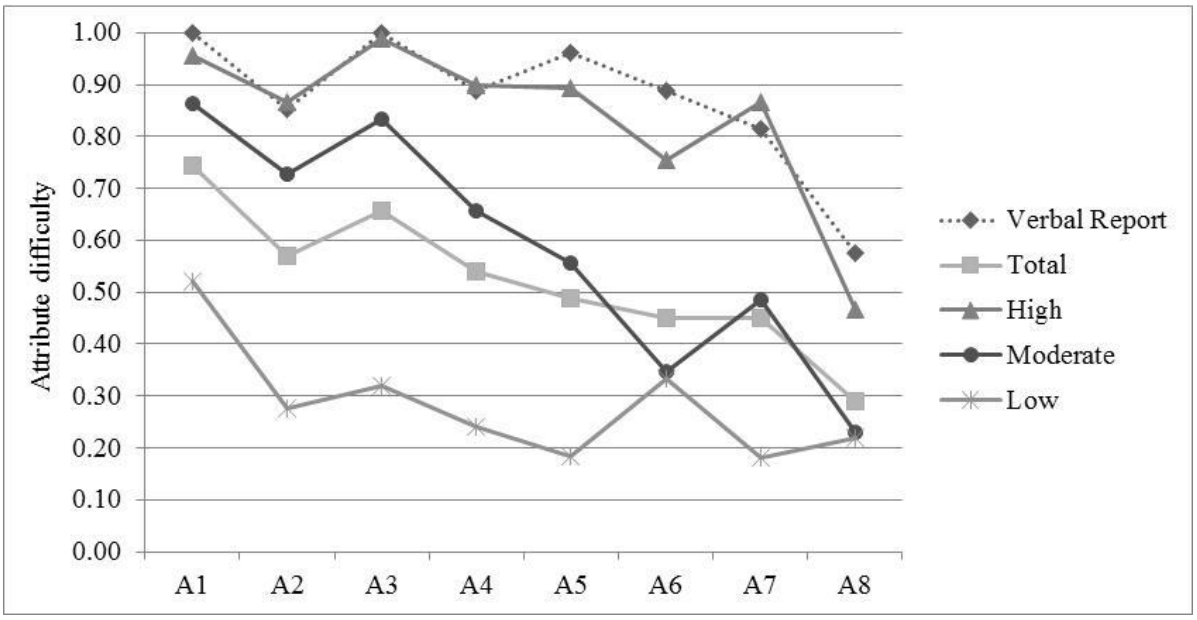



*Figure 6*. Attribute difficulties for each of the five groups.

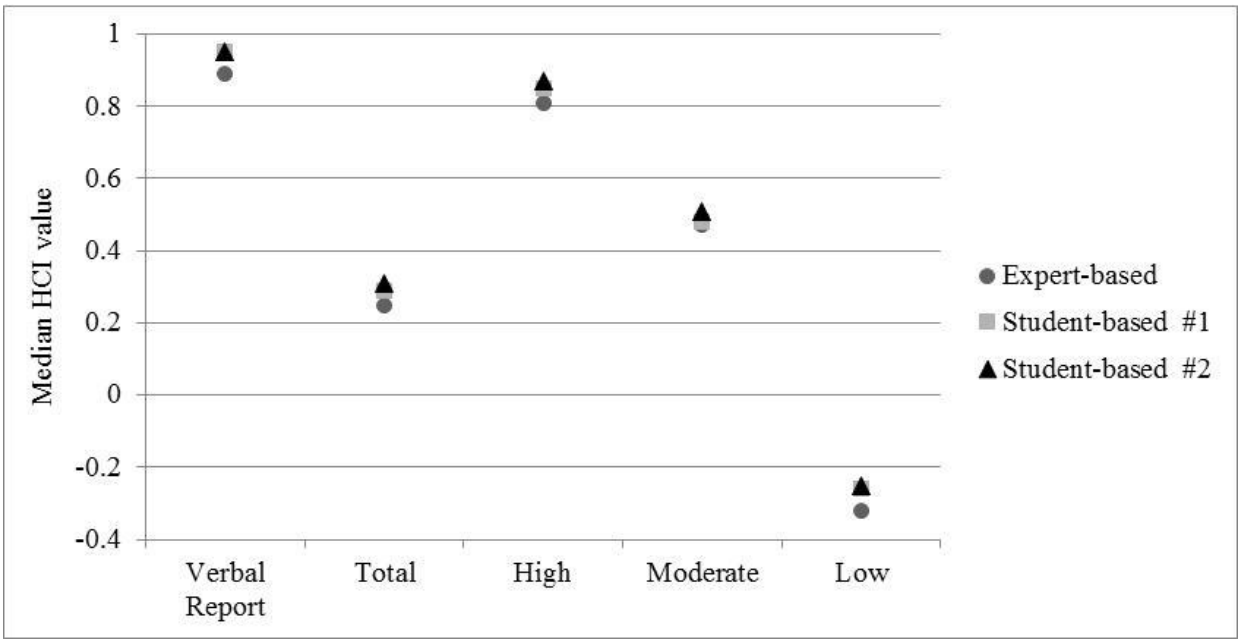



*Figure 7*. Median HCI value for each group as function of cognitive model type.