

---

**Using Judgments from Content Specialists to Develop  
Cognitive Models for Diagnostic Assessments**

**Mark J. Gierl**

**Mary Roberts**

**Cecilia Alves**

**Andrea Gotzmann**

Centre for Research in Applied Measurement and Evaluation  
University of Alberta



**Paper Presented at the Symposium  
“How to Build a Cognitive Model for Educational Assessments”**

Annual Meeting of the National Council on Measurement in  
Education (NCME), San Diego, CA

**April 15, 2009**

---

## USING JUDGMENTS FROM CONTENT SPECIALISTS TO DEVELOP COGNITIVE MODELS FOR DIAGNOSTIC ASSESSMENTS

### INTRODUCTION

Most educational assessments are based on cognitive problem-solving tasks. Cognitive diagnostic assessments (CDA) are designed to model examinees' performances on these tasks and yield specific information about their cognitive strengths and weaknesses. One way to produce this diagnostic summary is with an information-processing approach where the psychology of test performance is modelled to yield scores that measure examinees' cognitive skills. With a cognitive approach, problem solving is assumed to require the processing of information using relevant sequences of operations. Examinees are expected to differ in the knowledge they possess and the processes they apply thereby producing response variability in each testing situation. Because cognitive test performance is both covert and, often, complex, a model is required to link the examinees' problem-solving skills with interpretations of test performance. The development and use of a *cognitive model* provides one approach for identifying and measuring these skills so they can be connected with test performance and test score interpretations. A cognitive model in educational measurement refers to a simplified description of human problem solving on standardized tasks at some convenient grain size or level of detail in order to facilitate explanation and prediction of students' performance, including their strengths and weaknesses (Leighton & Gierl, 2007a). Cognitive models are indispensable in CDA because they provide an interpretative framework that can guide item development so test performance can be linked to specific inferences about examinees' knowledge and skills.

There are many potential benefits to modeling test performance using a CDA. For instance, these types of assessments could increase our understanding of student test performance, given many educational tests are based on cognitive problem-solving tasks. A test score serves as a coarse indicator of how students think about and solve educational tasks because cognitive performance cannot be observed directly (Snow & Lohman, 1989). Often, we assume that students who correctly solve a task use the appropriate knowledge and skills. However, this assumption is rarely substantiated and, in some cases, it may be wrong. Researchers have demonstrated, for instance, that examinees can generate correct answers using knowledge and skills that are unrelated to the target of inference specified in the items (e.g., Brown & Burton, 1978; Leighton & Gokiert, 2008; Norris, Leighton, & Phillips, 2004; Poggio, Clayton, Glasnapp, Poggio, Haack, &

Thomas, 2005). When this disjunction between the target of inference and student performance occurs, test score inferences, including diagnostic inferences, may be inaccurate because the student did not use the knowledge and skills the developer intended to measure.

CDA may also be used to link theories of cognition and learning with instruction. Most large-scale educational tests yield little information for students, teachers, and parents about why some students perform poorly or how instructional conditions can be altered to improve learning (National Research Council, 2001). Increasingly, cognitive theory and research is improving our understanding of student performance on a wide range of academic tasks (Anderson, 2005; Anderson & Shunn, 2000; Bransford, Brown, & Cocking, 1999; Donovan, Bransford, & Pellegrino, 1999; Mayer, 2008; Pellegrino, 1998, 2002; Pellegrino, Baxter, & Glaser, 1999). This enhanced view of thinking has also led to a better understanding of how assessment can be used to evaluate learning and improve instruction. Instructional decisions are made, in part, on how students think about and solve problems. Thus, teachers must draw on and, if possible, develop methods for making students' thinking overt so these cognitive skills can be evaluated. Instructional feedback can then focus on overcoming weaknesses while building on strengths. Cognitive models provide one method for representing thought. Because these models specify the knowledge structures and processing skills required to respond to test items, they can also be used to enhance test score interpretations and to guide instruction when the knowledge and skills specified in the model are identified as weak.

The benefit of developing a CDA using a cognitive model stems from the detailed information that can be obtained about the knowledge structures, processing skills, and their ordering that are used by examinees to produce a test score. Each item is designed to yield specific information about the students' cognitive strengths and weaknesses. If the target of inference is information about students' cognitive skills, then the small grain size associated with these models is required for generating specific information. This specific information can be generated because the grain size of these models is narrow thereby increasing the depth to which both knowledge and skills are measured with the test items. The drawback of developing a CDA according to a cognitive model stems from the paucity of information currently available on the knowledge and skills that characterize student performance in most testing situations. To make matters worse, some cognitive researchers also believe this situation is unlikely to change in the near future as there is little interest

in publishing outcomes from task analyses in the educational and psychological literature (Anderson & Schunn, 2000). Consequently, we have few cognitive models to draw on because little is known about how students actually solve items on educational tests. Hence, the purpose of our study is to describe how content specialists can be used to develop cognitive models for diagnostic assessments. We describe a two-stage procedure for developing these models. We also illustrate our procedure at the elementary and senior high school level using results from two operational testing programs.

### **CONTEXT FOR COGNITIVE MODEL DEVELOPMENT**

Our experience with developing cognitive models for CDA stems from two research projects. The first project, funded by the College Board, was designed to study cognitive models for college readiness. This project ran from Fall of 2004 to the Summer of 2008. Data from the SAT Reasoning Test and the Preliminary SAT<sup>®</sup>/National Merit Scholarship Qualifying Test (PSAT) were used. The PSAT is a co-sponsored program by the College Board and National Merit Scholarship Corporation. The PSAT is a standardized test that provides students with practice for the SAT Reasoning Test. It also allows students to enter National Merit Scholarship Corporation scholarship programs. The PSAT measures critical reading skills, math problem-solving skills, and writing skills. The purpose of the PSAT research was to investigate enhanced diagnostic scoring and reporting procedures so that students would receive more specific information about their strengths and weaknesses on college readiness skills. This enhanced feedback was intended to help students focus their preparation on areas where they wanted to improve their test performance. For the PSAT, we developed cognitive models for Critical Reading, Mathematics, and Writing. However, only the results for Mathematics will be described in this manuscript. Cognitive models were created for four content areas in Mathematics: (a) Numbers and Operations, (b) Algebra, (c) Geometry and Measurement, and (d) Data, Statistics, and Probability.

The second project is funded by the Learner Assessment Branch at Alberta Education. This project started in the Fall of 2008, and is on-going. Diagnostic Mathematics arose from a need identified by the Minister of Education in Alberta and practicing teachers who, together, concluded that a diagnostic tool was needed in mathematics to inform teachers about how students think and solve problems. The Alberta Education K-6 Diagnostic Mathematics project is a computer-based on-line assessment for students in Kindergarten to Grade

6. The goal of the project is to create tests that will provide teachers with diagnostic information so students' cognitive mathematical knowledge and skills can be identified and evaluated. The on-line administration system will include the assessments, score reports, and suggestions for instructional strategies that support students and teachers in becoming competent, literate, mathematical learners. Cognitive models were created 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.

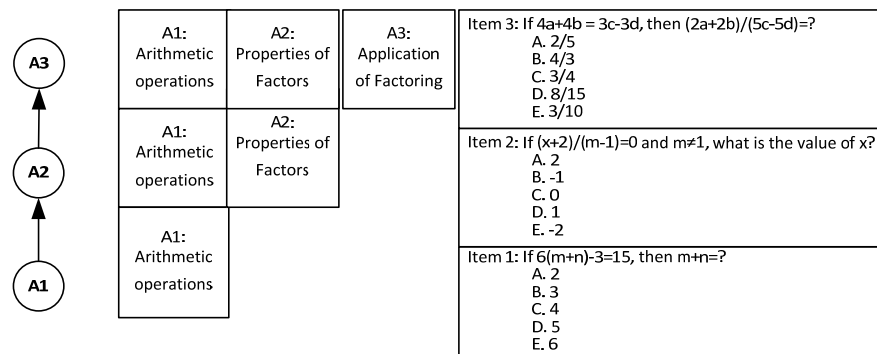
### CHARACTERISTICS OF COGNITIVE MODELS FOR CDA

Cognitive models for CDA have at least four defining characteristics. First, the model contains skills specified at a fine grain size because these skills must magnify the cognitive processes underlying test performance (**FINE GRAIN SIZE**). This grain size must be also be specified consistently so the knowledge and skills can be ordered, in some fashion, within the model and it must reflect the types of diagnostic inferences that will be produced from the diagnostic assessment in the score report. Second, the skills must be measurable (**MEASURABLE**). That is, each skill must be described in way that would allow a developer to create a test item to measure that skill. Third, the skills must be instructionally relevant and meaningful to a broad group of educational stakeholders, including the students, parents, and teachers (**INSTRUCTIONALLY RELEVANT**). Diagnostic skills will be reported to stakeholders as scores, and these scores are intended to guide remediation and instruction. Hence, the score feedback must be communicated clearly. Fourth, a cognitive model will often reflect a hierarchy of ordered skills within a domain because cognitive processes share dependencies and function within a much larger network of inter-related processes, competencies, and skills (**ORDERED SKILLS**). Assessments based on cognitive models can be developed so test items directly measure specific cognitive skills of increasing complexity, thereby allowing students' test item performance to be linked to information about their cognitive strengths and weaknesses using diagnostic and cognitive psychometric models such as the attribute hierarchy method (see Leighton, Gierl, & Hunka, 2004; Gierl, Leighton, & Hunka, 2007).

Figure 1 provides one example taken from a small section of a larger cognitive model developed for producing diagnostic inferences on SAT algebra (cf. Gierl, Wang, & Zhou, 2008). As a prerequisite skill, cognitive attribute A1 includes the most basic arithmetic operation skills, such as addition, subtraction,

multiplication, and division of numbers. In cognitive attribute A2, the examinee needs to have the basic arithmetic skills (i.e., attribute A1) as well as knowledge about the property of factors. In cognitive attribute A3, the examinee not only requires basic arithmetic skills (i.e., attribute A1) and knowledge of factoring (i.e., attribute A2), but also the skills required for the application of factoring. The attributes are specified at a fine grain size; each attribute is measurable; each attribute, and its associated item, is intended to be instructionally relevant and meaningful; and attributes are ordered from simple to more complex as we move from A1 to A3. Other hierarchical structures, in addition to the linear model, can also be selected to model cognitive information processing. Leighton, Gierl, and Hunka (2004) identified four forms of hierarchical structures and described their possible implications for construct representation and test development.

**Figure 1. Three sample items designed to measure three ordered skills in a linear cognitive model.**



### BACKGROUND OF THE CONTENT SPECIALISTS

Three content specialists participated in the PSAT project. Ten content specialists (1 examination manager, 1 Grade 3 examiner, 1 Grade 6 examiner, 3 Grade 3 content specialists, and 4 Grade 6 content specialists) worked on the Alberta Education K-6 Diagnostics Mathematics project. All participants were experienced classroom teachers or university instructors (see summary in Appendices A to C). Collectively, the three content specialists for PSAT had 72 years teaching experience; the three Alberta Education Diagnostic Mathematics staff had 67 years of teaching experience; the seven Alberta Education Diagnostic Mathematics content specialists had 134 years of experience. This foundation in student cognition, learning, and instruction among the content specialists was essential for cognitive model development because they identified fine-grained knowledge and skills required to solve problems in mathematics, ordered these skills

within each model, and described these cognitive attributes in a way that was both instructionally relevant and meaningful to a broad group of educational stakeholders. The content specialists also had a range of experiences in the development of large-scale educational tests. Aside from one Grade 3 Alberta Education content specialist who had no previous experience, all remaining specialists worked with testing agencies to develop items for large-scale assessments (and many had even more experience as markers, consultants, and/or exam managers in a host of large-scale testing projects). Hence, the specialists were familiar with the processes and “best practices” in large-scale item and test construction. This foundation in testing was essential for cognitive model development because the content specialists were required to ensure the skills they identified and ordered for each model were measurable, meaning that the skills could be assessed with test items.

### **PROCEDURES FOR DEVELOPING COGNITIVE MODELS**

The development and validation of the cognitive models for the diagnostic assessment projects occurred in two stages. The first stage consisted of developing the initial cognitive models. This stage was conducted to produce the cognitive models across different content areas in mathematics. The content specialists used research and curricular documents to inform their model development efforts, but they did not base their initial models on a review of existing test items. The second stage was designed to evaluate the cognitive models developed in the first stage by having the content specialists map existing test items onto the cognitive models. This stage was conducted so the content specialists could modify their initial models if additional or different knowledge and skills were detected when reviewing actual test items. The skill-by-item alignment task also provided information about how well existing items measured the knowledge and skills in the cognitive models.

#### **Stage #1: Developing the Initial Cognitive Models**

**PSAT Mathematics.** For PSAT Mathematics, stage #1 was completed in two steps. In the first step, the authors of this manuscript developed preliminary cognitive models. Our preliminary models were created in one month. This development work was undertaken so the content specialists would have a starting point for

creating their cognitive models.<sup>1</sup> To create the preliminary cognitive models, two College Board research papers—*Developing Skill Categories for the SAT Math Section* by O’Callaghan, Morley & Schwartz (2004) and the *Performance Category Descriptions for the Critical Reading, Mathematics, and Writing Sections of the SAT* (2007), also known as the SAT scale anchoring study—provided the starting points for creating the preliminary models. O’Callaghan et. al. (2004) described five cognitive skill categories identified by content specialists, after reviewing large numbers of previously administered SAT Mathematics items. Their cognitive skill categories ranged from simple to complex: (a) applying mathematical knowledge 1 (AMK1), (b) applying mathematical knowledge 2 (AMK2), (c) reasoning (REAS), (d) managing complexity (MC), and (e) creating representations and insight (CR/I). This categorization of ordered cognitive skills was applied to the four areas of Numbers and Operations, Algebra, Geometry and Measurement, and Data, Statistics, and Probability.

One problem with the skill categories proposed by O’Callaghan et al. was the that grain size was too coarse to be diagnostic, hence some refinement was required. To refine the skill grain size and to draw on appropriate skills measured by the SAT Mathematics items, we used the “performance descriptions”, which are relatively specific cognitive skills statements, from the SAT scale anchoring study. The SAT scale anchoring study was conducted by the College Board to identify what students know and can do at specific score intervals on the SAT. Skills in the scale anchoring study were identified by dividing the SAT score scale into intervals; then, the items that students could successfully complete in each score interval were identified; finally, skills in each interval were identified by asking content specialists to describe the knowledge and skills required by students to solve the items correctly. With a more complete set of skills at a finer grain size, we recoded each skill category identified by O’Callaghan et al. For example, under the score scale of 200-290 in the content area of Numbers and Operations, the fine-grain skill “identify factors of whole numbers” was classified under the skill category AMK1. All skills across the score scale intervals, content areas, and skill categories were classified in this manner. Next, the fine-grain skills were ordered in increasing cognitive complexity from simple to complex. Although the grain size of the skills within each cognitive category for each content area were relatively specific, we still found that the skills were difficult to order because many

---

<sup>1</sup> In our experience, a starting point—such as the development of our preliminary cognitive models—provides the basis for productive discussion among the content specialists in the complex task of cognitive model development, particularly when time is limited.

of them were still too broad and, oftentimes, skills were missing in our hierarchical ordering. Hence, the grain size of the cognitive skills was refined even further. At this point, we used our judgment to add and modify skills to create ordered cognitive skill categories in each content area to produce the final version of our preliminary cognitive models for PSAT Mathematics.

In the second step in Stage #1, three content specialists nominated by the College Board reviewed our preliminary cognitive models with the intention of making appropriate modifications, given a particular emphasis on the identification of the appropriate skills and on the ordering of these skills. They were also asked to evaluate each cognitive model for its measurability and instructional relevance. That is, the content specialists were instructed to modify our initial models in light of the four characteristics of cognitive models for CDA. Step 2 in Stage #1 was completed in two working days. All three content specialists had extensive mathematics backgrounds as well as teaching and test development experience (see Appendix A).

Prior to our meeting with the three content specialists for the PSAT study, two of our preliminary cognitive models from step 1 in Stage #1 were sent to each specialist for an independent review. That is, each specialist was given the “homework” assignment of reviewing two preliminary models and providing us with feedback on the skills and their ordering, along with suggestions for model revisions, before the three content specialists met as a group. This initial independent review provided the foundation for discussion and ensured that each specialist was familiar with the tasks they would perform.

As expected, when the three content specialists finally met, there was much discussion about how best to proceed with the revisions to the cognitive models. This discussion was fueled, in part, by their criticisms of our preliminary cognitive models. They identified three problems: (a) our preliminary models lacked key process skills that were used in the SAT scale anchoring study<sup>2</sup> (i.e., problem solving, representation, reasoning, connections, and communication), (b) the use of cognitive skills categories AMK1, AMK2, REAS, MC, and CR/I in our preliminary models were ambiguous and not representative of the skills used in the SAT scale anchoring study, and (c) the ordering of the skills in our preliminary cognitive models increased in difficulty, but did not always maintain a dependent, hierarchical ordering. Hence, the first major revision was to eliminate the O’Callaghan cognitive skill categories and rename the models simply as A to n, where n

---

<sup>2</sup> The three College Board content specialists who worked with us on the PSAT Diagnostic Mathematics project also participated as content experts on the SAT Scale Anchoring Study.

was the total number of models in a content area (the content in each cognitive model was to be labeled at a later date, but this activity was not completed before the end of the PSAT Project). The content specialists also agreed to structure the skills in the models as a developmental progression of skills. The content specialists began with revisions to the cognitive models in the areas of Number and Operations and Algebra using their “homework” results before completing Geometry and Measurement and Data, Statistics, and Probability.

Throughout the review, skills for each model within each content area were evaluated for clarity of description. The content specialists scrutinized the wording of skill descriptors to ensure it would be clear and meaningful to teachers. Any important skills that were deemed to be missing were drawn from the document, *College Board Standards for College Success: Mathematics and Statistics* (2006). This document outlines the knowledge and skills that students should master for college success beginning with Middle School Math I and concluding with High School Precalculus. Any relevant, measurable, and instructionally relevant process skills were also added to the cognitive models. When revising the order of the skills, the content specialists also felt the cognitive model must reflect the different connections among the content and skills within the domain of Mathematics. Modeling this connection was accomplished by reusing a number of skills both within and across content areas. The outcome from steps 1 and 2 in Stage #1 was a strong first draft of the cognitive models that characterized student performance on PSAT Mathematics. In total, 39 cognitive models containing 134 ordered diagnostic skills seen as measurable and instructionally relevant were identified across four content areas in PSAT Mathematics (see Table 2).

**Table 2. Summary of Cognitive Models in PSAT Mathematics by Content Area**

Content Areas	# Cognitive Models	# Skills
Numbers and Operations	9	28
Algebra	14	46
Geometry and Measurement	11	40
Data, Statistics, Probability	5	20
Total	39	134

**Alberta Education K-6 Diagnostic Mathematics.** The development of the cognitive models for the Alberta Education Diagnostic Mathematics project was also completed in two steps within Stage #1. Cognitive models were developed for Grades 3 and 6. However, unlike the PSAT project, Diagnostic Mathematics was directed by three full-time staff members—an exam manager was responsible for the entire Diagnostic

Mathematics project and one examiner at each grade level. All three Alberta Education staff had extensive mathematics teaching experience as well as range of knowledge, skills, and experiences in developing achievement test items (see Appendix B). Our research staff from the University of Alberta worked with the three Alberta Education examiners to guide cognitive model development. In the first step of Stage #1, the exam manager developed the preliminary cognitive models for two content areas at Grade 3. This preliminary work was undertaken so the examiners and content specialists who were responsible for validating the preliminary cognitive models would have a starting point for their discussions. The exam manager created her preliminary cognitive models using the content, knowledge, and skills specified in the provincial curriculum which, in Canada, is common for the four western provinces of British Columbia, Alberta, Saskatchewan, and Manitoba and the three northern territories, the Yukon, the Northwest Territories, and Nunavut. Education across these seven jurisdictions is guided by the Western and Northern Canadian Protocol (WNCP) for the Collaboration in Basic Education. In the WNCP, the mathematics curriculum from Kindergarten to Grade 9 is described in the document, *The Alberta K-9 Mathematics Program of Studies with Achievement Indicators* (2007). The program of studies identifies beliefs about mathematics, general outcomes, specific outcomes, and achievement indicators agreed on by the seven provinces and territories. Learning outcomes in the program of studies for K-9 are organized into four content areas: Number, Patterns and Relations, Shape and Space, and Statistics and Probability. By reviewing the program of studies and drawing on her own knowledge, beliefs, and experiences, the Alberta Education exam manager developed the preliminary cognitive models in the “Number” and “Patterns and Relations” content areas at Grade 3.

Then, in the second step of Stage #1, the cognitive models developed by the exam manager, were scrutinized by the two examiners and seven content specialists. Step 2 of Stage #1 was completed in 1.5 working days. The examiners and content specialists were all active teachers with a range of backgrounds and experiences in teaching and test development (see Appendix C). Although one Grade 3 content specialist had no test development experience with Alberta Education, the remaining six content specialists were highly experienced provincial test developers. Prior to meeting as a group, each content specialist was asked to review the initial cognitive models independently as their “homework” assignment. During this review, the content specialists were instructed to evaluate each model, with a particular emphasis on the

identification of the appropriate skills and their ordering, as well as evaluating each model for its measurability and instructional relevance. To facilitate this review, the content specialists were provided with the Alberta K-9 Mathematics Program of Studies document and a description of the Diagnostic Mathematics project. This initial independent review provided the foundation for discussion when the three Grade 3 content specialists met with the Grade 3 examiner and the four Grade 6 content specialists met with the Grade 6 examiner.

When the content specialists met as a group, the preliminary cognitive models developed by the Alberta Education exam manager were modified substantially. The first key change was the introduction and development of “knowledge” and “skill” statements by the content specialists. These statements were later used by the content specialists to generate cognitive skills at a fine grain size that, eventually, would be included in the cognitive models. For example, in the content area “Number”, for the general outcome “Developing Number Senses”, with the specific outcome “demonstrating understanding of place value for numbers greater than 1 thousand and less than 1 million”, the knowledge and skills initially identified by the content specialists included:

- Grouping in tens
- Patterns, each place value is 10 times the value of the place to the right
- A number has many different meanings depending on its place value
- Zero is a place holder
- Read and write numerals for numbers of any magnitude ranging from ten thousandths to billions
- Use expanded form to explain a number and its relationship to other numbers
- Relate the value of a number using its place value location in relation to other place value units

These knowledge and skills served as the first step in specifying finer grain-sized cognitive attributes for each specific outcome presented in the Alberta K-9 Mathematics Program of Studies. Next, the wording of the knowledge and skill statements was refined so the grain size was consistent across all content areas and that the statements were instructionally relevant and meaningful to teachers. Any important skills that were deemed to be missing were identified and added by the content specialists. A preliminary ordering of the skills was also proposed. Finally, the skills were explicitly ordered from simple to most complex for each specific outcome in the Alberta K-9 Mathematics Program of Studies to create the initial cognitive models. The wording of the skills in each model was, again, reviewed and revised. In total, seven iterations were

recorded for Grade 3 model development while nine iterations were required for the Grade 6 models. Like the content specialists for the PSAT project, the Alberta Education content specialists believed the cognitive models must reflect different connections among the content and skills within the domain of Mathematics. Hence, a number of skills were used repeatedly for cognitive models both within and across content areas. The outcome from steps 1 and 2 in Stage #1 was a strong first draft of the cognitive models that characterized student performance in Alberta Education Mathematics curriculum at Grades 3 and 6. In total, 26 cognitive models containing 178 ordered diagnostic skills seen as measurable and instructionally relevant were identified across four mathematics content areas in Grade 3. Similarly, 26 cognitive models containing 150 ordered diagnostic skills seen as measurable and instructionally relevant were identified across the four content areas in Grade 6. The results are summarized in Table 5.

**Table 5. Summary of Cognitive Models in Alberta Education Diagnostic Mathematics by Grade and Content Area**

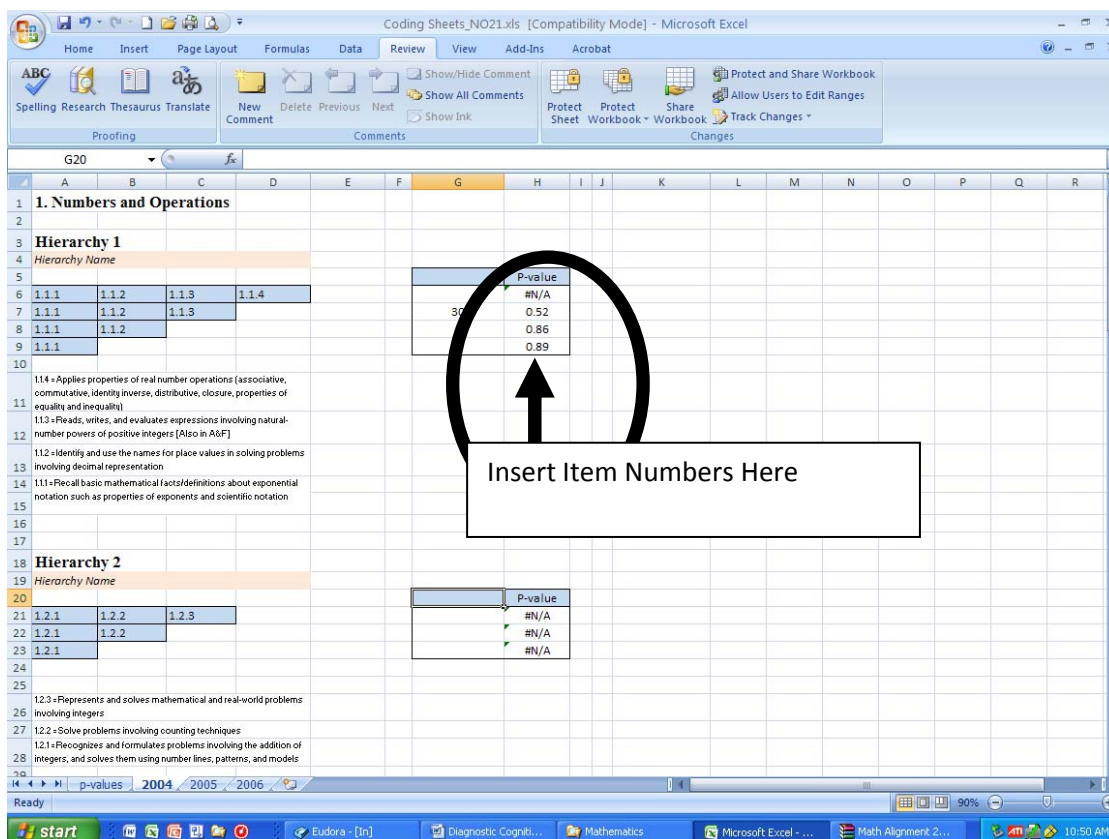
Content Area	# Cognitive Models	# Skills
Grade 3		
Numbers	13	92
Patterns & Relations	4	32
Shape & Space	7	42
Statistics & Probability	2	12
Total	26	178
Grade 6		
Numbers	9	64
Patterns & Relations	4	22
Shape & Space	9	46
Statistics & Probability	4	18
Total	26	150

### Stage #2: Skill-by-Item Alignment Task

**PSAT Mathematics.** The purpose of Stage #2 was to align existing PSAT items to the skills outlined in the cognitive models developed in Stage #1. This task was conducted so the content specialists could modify their initial cognitive models in light of additional or different knowledge and skills that became apparent by reviewing operational test items. The skill-by-item alignment task also provided us with valuable information about how well existing items measured the knowledge and skills in the cognitive models. The same three content specialists who participated in Stage #1 also worked on Stage #2. Each specialist was presented with the math items from the 2005 and 2006 administration of the PSAT and asked to align each item with one or

more skills in the cognitive models. The alignment meeting required two working days, and was conducted four months after the cognitive models were first created. As in the previous stage, each content specialist was asked to complete the alignment task independently for one content area as “homework” prior to the group meeting. To guide the skill-by-item alignment task, the content specialists were also provided with the p-value or proportion correct value for each item calculated from a random sample of 5000 students who wrote these forms of the exam. The cognitive models and p-values were placed in an EXCEL item coding spreadsheet. Recall, cognitive models reflect a hierarchy of ordered skills ranging from least to most complex. Hence, the p-values were intended to provide some empirical feedback for the content specialists on the accuracy of their ordering of these skills. The content specialists placed the item numbers into each sheet, as shown in Figure 2. They were also asked to order the items in difficulty from easiest to hardest (i.e., from highest p-value at the bottom of each box to lowest p-value at the top of the box), if possible.

**Figure 2. The EXCEL item coding sheet used for the skill-by-item alignment study in Stage #2.**



Two activities occurred during the review of each model. First, the content specialists reviewed the description of the skills and scrutinized the ordering of the skills. The review produced some small but

important changes to the cognitive models. This finding demonstrates that reviewing operational test items is helpful in cognitive model development because each item provides a concrete example of the knowledge and skills that can be measured on a test. The ordering of operational items with p-values also demonstrates how the skills can be arranged to measure different cognitive proficiencies for each model. Twelve of the 39 (31%) cognitive models were changed by either adding or removing skills and/or changing the ordering of the skills. In addition, 23 of the 134 (17%) initial skills were modified by either adding or removing words, phrases, and/or descriptions. According to the content specialists, these changes to the skills and their ordering were motivated by the need to ensure that each model contained skills that were specified at a consistent grain size to yield measurable and instructionally relevant results. A sample of the final five cognitive models produced by the content specialists in the content area Numbers and Operations is provided in Appendix D.

Second, the content specialists aligned all items from the 2005 (38 items) and 2006 (38 items) administrations of the PSAT to the skills in their final cognitive models. The coding results must be interpreted carefully for two reasons. First, some of the skills were measured in multiple models, thus items could be aligned to skills more than once. Also, some of the items were believed to measure the same skill, with varying degrees of skill-by-item fit (as will be described in the next paragraph), hence multiple items could be coded for the same skill. With these coding conventions in-mind, the 38 items from 2005 were aligned to the 134 skills 73 times, which is a 55% alignment match. That is, one or more of the items on the PSAT measured at least one of the skills in our cognitive models 55% of the time. The remaining 45% of the skills were not measured by any of the available PSAT items. The alignment match ranged from a low of 25% in Data, Statistics, and Probability to a high of 93% in Numbers and Operations. Similarly, the 38 items from 2006 were aligned to the 134 skills in the cognitive models 106 times, which is a 79% alignment match. The alignment match ranged from a low of 35% in Data, Statistics, and Probability to a high of 100% in Numbers and Operations. The results are summarized in Table 6.

**Table 6. Skill-by-Item Alignment using Items from the 2005 and 2006 PSAT Administrations**

Content Areas	# Hierarchies	# Skills	# Items matched to Skills
<b>2005</b>			
Numbers & Operations	9	28	26 (92.9%)
Algebra & Functions	14	46	25 (54.4%)
Geometry & Measurement	11	40	17 (42.4%)
Data, Statistics, & Probability	5	20	5 (25.0%)
<b>Total</b>	<b>39</b>	<b>134</b>	<b>73 (54.5%)</b>
<b>2006</b>			
Numbers & Operations	9	28	32 (> 100%)
Algebra & Functions	14	46	42 (91.3%)
Geometry & Measurement	11	40	25 (62.5%)
Data, Statistics, & Probability	5	20	7 (35.0%)
<b>Total</b>	<b>39</b>	<b>134</b>	<b>106 (79.1%)</b>

A coding scheme for evaluating the quality of the skill-by-item alignment was also developed. Each skill-by-item combination was rated in one of four ways: The PSAT item and the cognitive skill were judged to be a good fit (GF); the PSAT items measured only a subset of the cognitive skill (I/S); only a subset of the skill was measured by the PSAT item (S/I); or item and skill were a poor fit (PF). The quality of the alignment results are revealing. In 2005, for instance, of the 55% of the items that aligned to the skills, only 23% were judged to be a good fit; 12% contained items that measured only a subset of the cognitive skill (I/S); conversely, only a subset of the skill was measured by 43% of the items (S/I); and 22% of the skill-by-item alignments were judged to be a poor fit. We can also combine categories I/S, S/I, and PF because these three categories indicate inadequate skill-by-item fit. In this case, of the 55% of the items aligned to the skills, 23% were judged to be a good fit while the remaining 77% provided an inadequate skill-by-item fit using the 2005 items. In 2006, of the 79% of the items that aligned to the skills, only 36% were judged to be a good fit; 9% contained items that measured only a subset of the cognitive skill (I/S); in other cases, only a subset of the skill was measured by 23% of the items (S/I); and 32% of the skill-by-item alignments were judged to be a poor fit. In other words, of the 79% of the items aligned to the skills in 2006, 36% were judged to be a good fit while the remaining 64% provided an inadequate skill-by-item fit. The results are summarized in Table 7.

**Table 7. Quality of the Skill-by-Item Alignment using Items from the 2005 and 2006 PSAT Administrations**

Content Area	# Skills	# Match	Skill-Item Fit			
			GF	I/S	S/I	PF
<b>2005</b>						
Numbers & Operations	28	26	7.7%	15.4%	53.8%	23.1%
Algebra & Functions	46	25	32.0%	0.0%	32.0%	36.0%
Geometry & Measurement	40	17	23.5%	29.4%	41.2%	5.9%
Data, Statistics, & Probability	20	5	60.0%	0.0%	40.0%	0.0%
Total	134	73	23.3%	12.3%	42.5%	21.9%
<b>2006</b>						
Numbers & Operations	28	32	38.9%	16.7%	44.4%	77.8%
Algebra & Functions	46	42	35.7%	25.0%	39.3%	50.0%
Geometry & Measurement	40	25	78.9%	0.0%	21.1%	31.6%
Data, Statistics, & Probability	20	7	85.7%	0.0%	14.3%	0.0%
Total	134	106	35.8%	9.4%	22.6%	32.1%

**Alberta Education K-6 Diagnostic Mathematics.** The purpose of Stage #2 was to align existing Alberta Education mathematics achievement test items to the skills outlined in the cognitive models developed in Stage #1. The same Alberta Education content specialists who participated in Stage #1 also worked on Stage #2. Each specialist was presented with the math items from the 2006 and 2007 administration of the achievement tests and asked to align each item with one or more skills in the cognitive models. The alignment task was conducted one month after the cognitive models were developed. It was completed in two working days. Each content specialist was asked to complete the alignment task independently for one content area as “homework” prior to the group meeting. As with PSAT Mathematics, the content specialists were provided with the p-values for each item calculated from a random sample of 5000 students who wrote each form of the test to guide the skill-by-item alignment task. The cognitive models and p-values were placed in an EXCEL item coding spreadsheet. The content specialists were asked to place the item numbers into each sheet. They were also asked to order the items in difficulty from easiest to hardest, whenever possible.

Two activities occurred during the review of each model. First, the content specialists reviewed the description of the skills and scrutinized the ordering of the skills. This review produced some small but important changes to the cognitive models. In Grade 3, 17 wording changes were made to the cognitive models (the wording changes were distributed as follows: 9 in Number, 4 in Patterns & Relations, 4 in Shape

& Space, 0 in Statistics & Probability). In addition, 14 cognitive models were changed by either adding or removing skills (6 changes in Number, 4 in Patterns & Relations, 3 in Shape & Space, 1 in Statistics & Probability). In Grade 6, 9 wording changes were made to the cognitive models (3 changes in Number, 0 in Patterns & Relations, 3 in Shape & Space, 2 in Statistics & Probability). Nine cognitive models were changed by either adding or removing skills (3 changes in Number, 1 in Patterns & Relations, 2 in Shape & Space, 3 in Statistics & Probability). One new cognitive model was also created. These results demonstrate that the skill-by-item alignment task is helpful in refining the cognitive models. The content specialists justified these changes to the wording or to the addition and removal of skills because they wanted to ensure that each model contained skills that were specified at a consistent grain size to yield measurable and instructionally relevant results. A sample of the final five cognitive models developed by the Alberta Education content specialists in the content area Numbers and Operations is provided in Appendix E.

Second, the content specialists aligned all items from the 2006 (40 items in Grade 3; 50 items in Grade 6) and 2007 (40 items in Grade 3; 50 items in Grade 6) administrations of the provincial achievement tests to the skills in their final cognitive models. The coding scheme used in the College Board study was also applied: The item-by-skill alignment was judged to be a good fit (GF); the item measured only a subset of the skills (I/S); only a subset of the skills were measured by the item (S/I); or the item and skill were a poor fit (PF). As in the College Board analysis, some of the skills were measured in multiple models, thus items could be aligned to skills more than once. Also, some of the items were believed to measure the same skill, with varying degrees of skill-by-item fit, hence multiple items could be coded for the same skill.

For Grade 3, 40 items from 2006 were aligned to the 200 skills 41 times, which is a 21% alignment match. In other words, one or more of the items on the Alberta Education Mathematics achievement test in Grade 3 measured at least one of the skill in the cognitive models only 21% of the time. The remaining 79% of the skills were not measured by any of the available math achievement test items. The alignment match ranged from a low of 9% in Shape and Space to a high of 54% in Statistics and Probability. Similarly, the 40 items from 2007 were aligned to the 200 skills in the cognitive models 40 times, which is a 20% alignment match. The alignment match ranged from a low of 6% in Shape and Space to a high of 77% in Statistics and Probability. The Grade 3 results are presented in Table 8.

**Table 8. Skill-by-Item Alignment in Grade 3 using Items from the 2006 and 2007 Administrations**

Content Areas	# Hierarchies	# Skills	# Items matched to Skills	
			2006	2007
Numbers	13	87	24 (27.6%)	19 (21.8%)
Patterns and Relations	4	46	5 (10.9%)	8 (17.4%)
Shape and Space	7	54	5 (9.3%)	3 (5.6%)
Statistics and Probability	2	13	7 (53.8%)	10 (76.9%)
Total	26	200	41 (20.5%)	40 (20.0%)

For Grade 6, 50 items from 2006 were aligned to the 162 skills 31 times, which is a 19% alignment match. That is, one or more of the items on the Alberta Education Mathematics achievement test in Grade 6 measured at least one of the skill in the cognitive models 19% of the time. The remaining 81% of the skills were not measured by any of the available math achievement test items. The alignment match ranged from a low of 12% in Shape and Space to a high of 32% in Statistics and Probability. The 50 items from 2007 were aligned to the 162 skills in the cognitive models 67 times, which is a 41% alignment match. The match ranged from a low of 26% in Shape and Space to a high of 68% in Statistics and Probability (see Table 9).

**Table 9. Skill-by-Item Alignment in Grade 6 using Items from the 2006 and 2007 Administrations**

Content Area	# Hierarchies	# Skills	# Items matched to Skills	
			2006	2007
Numbers	9	72	12 (16.7%)	29 (40.3%)
Patterns & Relations	5	28	8 (28.6%)	14 (50.0%)
Shape & Space	9	43	5 (11.6%)	11 (25.6%)
Statistics & Probability	4	19	6 (31.6%)	13 (68.4%)
Total	27	162	31 (19.1%)	67 (41.4%)

The quality of each skill-by-item combination was also evaluated. For Grade 3 in 2006, of the 21% of the items that aligned to the skills, 42% were judged to be a good fit; 44% contained items that measured only a subset of the cognitive skill (I/S); conversely, only a subset of the skill was measured by 5% of the items (S/I); and 10% of the skill-by-item alignments were judged to be a poor fit. If we combine categories I/S, S/I, and PF, then we conclude that of the 21% of the items aligned to the skills, 42% were judged to be a good fit while the remaining 58% provided an inadequate skill-by-item fit. In 2007, of the 20% of the items that aligned to the skills, only 17% were judged to be a good fit; 71% contained items that measured only a subset of the cognitive skill (I/S); in other cases, only a subset of the skill was measured by 5% of the items (S/I); and 5% of the skill-by-item alignments were judged to be a poor fit. That is, of the 20% of the items

aligned to the skills in 2007, 17% were judged to be a good fit while the remaining 83% provided an inadequate skill-by-item fit. The results are summarized in Table 10.

**Table 10. Quality of the Skill-by-Item Alignment in Grade 3 using Items from the 2006 and 2007 Administrations**

Content Area	# Skills	# Match	Skill-Item Fit			
			GF	I/S	S/I	PF
<b>2006</b>						
Numbers	87	24	29%	58%	4%	8%
Patterns and Relations	46	5	40%	20%	20%	20%
Shape and Space	54	5	60%	20%	0%	20%
Statistics and Probability	13	7	71%	29%	0%	0%
Total	200	41	41.5%	43.9%	4.9%	9.8%
<b>2007</b>						
Numbers	87	19	6%	88%	6%	12%
Patterns and Relations	46	8	38%	50%	13%	0%
Shape and Space	54	3	33%	67%	0%	0%
Statistics and Probability	13	10	20%	80%	0%	0%
Total	200	40	17.1%	70.7%	4.9%	4.9%

For Grade 6 in 2006, of the 19% of the items that aligned to the skills, 39% were judged to be a good fit; 45% contained items that measured only a subset of the cognitive skill (I/S); only a subset of the skill was measured by 3% of the items (S/I); and 13% of the skill-by-item alignments were judged to be a poor fit. When we combine categories I/S, S/I, and PF, we conclude that of the 19% of the items aligned to the skills, 39% were judged to be a good fit while the remaining 61% provided an inadequate skill-by-item fit. In 2007, of the 41% of the items that aligned to the skills, 30% were judged to be a good fit; 49% contained items that measured only a subset of the cognitive skill (I/S); only a subset of the skill was measured by 10% of the items (S/I); and 10% of the skill-by-item alignments were judged to be a poor fit. In other words, of the 41% of the items aligned to the skills in 2007, 30% were judged to be a good fit while the remaining 70% provided an inadequate skill-by-item fit. The results are summarized in Table 11.

**Table 11. Quality of the Skill-by-Item Alignment in Grade 6 using Items from the 2006 and 2007 Administrations**

Content Area	# Skills	# Match	Skill-Item Fit			
			GF	I/S	S/I	PF
<b>2006</b>						
Numbers	72	12	42%	58%	0%	0%
Patterns & Relations	28	8	25%	50%	0%	25%
Shape & Space	43	5	20%	60%	0%	20%
Statistics & Probability	19	6	67%	0%	17%	17%
Total	162	31	38.7%	45.2%	3.2%	12.9%
<b>2007</b>						
Numbers	72	29	38%	55%	7%	0%
Patterns & Relations	28	14	8%	75%	17%	17%
Shape & Space	43	11	43%	43%	14%	57%
Statistics & Probability	19	13	42%	42%	17%	8%
Total	162	67	29.9%	49.3%	10.4%	10.4%

### SUMMARY AND DISCUSSION

Judgments from content specialists are used extensively for developing educational assessments. Their knowledge of the curriculum and their understanding of the content is indispensable in the test development process. However, developing items for cognitive diagnostic assessments (CDA) poses some unique challenges, in part, because these types of tests are based on specific, ordered cognitive skills outlined in a cognitive model. A cognitive model in educational assessment refers to 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 (Leighton & Gierl, 2007a). Hence, the focus is on characterizing how students think and solve problems on tests. To make the transition from content-based inferences to diagnostic problem-solving inferences both the content specialists' qualifications and their item development tasks must be considered. We drew on the expertise of content specialists who had extensive teaching and test development experience. The teaching experience provides the foundation for understanding student thinking, learning, and instruction necessary for identifying and ordering instructionally-relevant skills in cognitive models; the test development experience provides the foundation needed to create items to measure these skills. In other words, the content specialists' task was guided by our requirements for the

diagnostic cognitive model—the models must be specified at a fine grain size with ordered and measurable cognitive skills that are instructionally relevant for teachers. We also described a two-stage procedure for developing cognitive models using our experiences and results from two operational testing programs, PSAT Mathematics and Alberta Education K-6 Diagnostic Mathematics projects. The cognitive models produced using these procedures are illustrated in Appendices D and E.

### **Implications for Creating Cognitive Diagnostic Assessments**

Gierl and Cui (2008) noted that the successful application of CDA in educational testing would require new test development procedures and practices, in part, because the cognitive models that guide CDA have specific requirements. We identified and operationalized four cognitive model requirements in this study: grain size, measurability, instructional relevance, and ordered skills. One consequence of these requirements is that the **common practice of retrofitting** diagnostic and cognitive models to existing achievement testing data is likely to yield unsatisfactory diagnostic inferences about students' cognitive strengths and weaknesses. Retrofitting can be described as the addition of a new technology or feature to an older system. Similarly, we might consider cognitive diagnostic retrofitting as the application of a new statistical or psychometric model to student response data from an existing testing system that uses traditional test development procedures and practices. We contend that conducting CDA through retrofitting will yield few successful applications because the cognitive models that guide the psychometric analyses in CDA have specific requirements about the structure of the testing data and that this structure is unlikely to exist without some form of principled test design.

The results from our study provide some evidence to support this claim. For instance, 38 PSAT items from 2005 were aligned to the 134 skills 73 times, which is a 55% alignment match, meaning that one or more of the items on the PSAT measured at least one of the skill in our cognitive models 55% of the time. The remaining 45% of the skills were not measured by any of the available PSAT items. The skill-by-item alignment result was better in 2006 as the 38 items were aligned to the 134 skills in the cognitive models 106 times, which is a 79% alignment match. For the Alberta Education Mathematics achievement tests, skill-by-item alignment produced a range of fit from a low of 20% for the Grade 3 2007 administration to a high of 41% for the Grade 6 2007 administration. These uneven distribution of skill-to-items across grades and tests

occurs for the simple, and probably obvious, reason that item development for these tests was not guided by a cognitive model. The consequence of this uneven distribution is potentially severe for CDA because many skills will rarely or never be measured using traditional achievement test items. Hence, the diagnostic inferences that can be made from these items are poor. And yet this limitation should be expected whenever item development proceeds without an explicit cognitive model because large-scale tests like the PSAT and the Alberta Education Mathematics achievement test were neither intended nor developed to evaluate hypotheses about the specific cognitive bases of student performance.

But even when cognitive skills happen to align with some of the existing items, the fit is precarious. The best fit we found occurred with the 2006 PSAT items. Of the 79% of the items that aligned to skills, 36% were judged to be a good fit. The remaining 64% of the items provided either marginal or poor fit. The weakest alignment occurred with the 2007 Alberta Education Grade 3 items. Of the 20% of the items that aligned to the skills, only 17% were judged to be a good fit. From these results we conclude that the cognitive analysis of existing tests using retrofitting procedures will, invariably, produce a tenuous fit between the model (assuming that the model can be identified, initially) and the test data because the tests were not designed from an explicit cognitive framework.

To overcome the limitations associated with retrofitting data to diagnostic and cognitive psychometric models, we advocate a more principled approach to test design and analysis in CDA. Principled test design and analysis in CDA adopts all of the existing standards of practices in test development. But it also has some additional requirements: A cognitive model must be identified and evaluated; items must be developed to measure the knowledge and skills in the cognitive model; and confirmatory diagnostic or cognitive psychometric models are applied to the data to generate the diagnostic scores and reports. We contend that a cognitive model, of some type, will always be needed to develop items and analyze student response data, generate scores, and guide score interpretations for CDAs because this form of testing is designed to identify and evaluate students' cognitive skills at a fine grain size. Ordering will be required to orchestrate the skills into a coherent whole to create the model of test performance. To link the cognitive model with test design, the skills must be measurable. The skills must also be instructionally relevant because students, teachers, and parents may want to use the examinees' scores on these skills to guide and inform instruction.

### Next Steps in Cognitive Model Development

We are expanding the cognitive models for the Alberta Education K-6 Diagnostic Mathematics Project in two ways. First, we are developing three parallel items to measure each skill in every cognitive model. Multiple items per skill are needed to enhance the reliability estimate for diagnostic scoring (cf. Gierl, Cui, & Zhou, in press). The focus for Spring 2009 field testing will be to evaluate skills in the content area of “Number” at both Grades 3 and 6. The “Number” content area in Grade 3 has 13 cognitive models with 87 skills. Hence, 261 new diagnostic items were created. The “Number” content area in Grade 6 has nine cognitive models with 72 skills, resulting in the development of 216 new diagnostic items. Because a cognitive model reflects a hierarchy of knowledge, processes, and strategies, the items measuring these skills at each level in the model were developed at different difficulty levels. For instance, a model with five skills would contain items with difficulty levels or p-values at five different intervals varying from easy to difficult: 0.90 (range: 1.00 to 0.81), 0.70 (range: 0.80 to 0.61), 0.50 (range: 0.60 to 0.41), 0.30 (range: 0.40 to 0.21) and 0.10 (range: 0.20 to 0.00). In other words, a p-value target of 0.90 means that 90% of the students in the field test sample who write the items at this level in the cognitive model should get the items correct. The p-values for the items at this level could range from 100% to 81%, and still be permissible for measuring this skill in the model (see Figure 3 for an illustration).

**Figure 3. The item p-value requirements for a five-skill cognitive model ranging from easy to difficult.**

P-Value Target:	90		70		50		30		10	
	SKILL #1		SKILL #2		SKILL #3		SKILL #4		SKILL #5	
P-Value Range:	100	81	80	61	60	41	40	21	20	0

Second, we are validating our cognitive models with psychological evidence from the population to which inferences will be targeted using verbal reports and protocol analysis (these methods will be discussed in more detail by Dr. Jacqueline Leighton in her presentation at this symposium titled, “*Exploratory and Confirmatory Methods for Cognitive Model Development*”). That is, content specialist create models of cognitive *intentions* as they identify the knowledge and skills students’ are *expected* to use when solving items. However, the content specialists’ intentions should be evaluated to determine if students actually

solve items in a manner outlined in the cognitive models. One method for evaluating the models is to study the cognitive processes used by students as they respond to diagnostic items. These processes can be identified by asking students to think aloud as they solve the diagnostic items and studied using protocol analysis to identify the information requirements and processing skills elicited by the tasks (Ericsson & Simon, 1993; Leighton, 2004; Leighton & Gierl, 2007b; Royer, Cisero, & Carlo, 1993; Taylor & Dionne, 2000). For example, Gierl et al. (2008) used verbal reports and protocol analysis to validate cognitive models of algebra performance by asking a sample of students to think aloud as they solved SAT algebra items designed to measure each skill in a cognitive model. Gierl et al. concluded the content specialists' cognitive models provided an excellent approximation to actual student performance. In fact, verbal report data did not result in any structural changes to the cognitive models. The verbal report data did, however, allow the authors to develop a more concise description of each skill which aided in diagnostic score reporting. Hence, verbal reports and protocol analysis should be considered a necessary and beneficial step in cognitive model development. Verbal report data will be collected for the cognitive models in the Alberta Education K-6 Diagnostic Mathematics Project during late Spring or early Fall of 2009.

### **Acknowledgements**

The research reported in this study was conducted with funds provided to the first author by the College Board and by the Learner Assessment Branch at Alberta Education. We would like to thank the College Board and Alberta Education for their support. However, the authors are solely responsible for the methods, procedures, and interpretations expressed in this study. Our views do not necessarily reflect those of the College Board or Alberta Education.

## REFERENCES

- Alberta Education (2007). *The Alberta K-9 Mathematics Program of Studies with Achievement Indicators*. Edmonton, AB: Alberta Education.
- Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, 29, 313-341.
- Anderson, J. R., & Shunn, C. D. (2000). Implications of the ACT-R learning theory: No magic bullets. In R. Glaser (Ed.), *Advances in instructional psychology: Educational design and cognitive science*, Vol. 5. (pp. 1-33). Mahwah, NJ: Erlbaum.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.
- Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematics skills. *Cognitive Science*, 2, 155-192.
- The College Board. (2006). *College Board Standards for College Success: Mathematics & Statistics*, New York, NY: The College Board.
- The College Board. (2007). *Performance category descriptions for the critical reading, mathematics, and writing sections of the SAT*. New York, NY: The College Board.
- Donovan, M. S., Bransford, J. D., & Pellegrino, J. W. (1999). *How people learn: Bridging research and practice*. Washington, DC: National Academy Press.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.
- 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, 263-268.
- Gierl, M. J., Cui, Y., & Zhou, J. (in press). Reliability of attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement*.
- Gierl, M. J., Leighton, J. P., & Hunka, S. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic*

- assessment for education: Theory and applications.* (pp. 242-274). Cambridge, UK: Cambridge University Press.
- 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<sup>®</sup>. *Journal of Technology, Learning, and Assessment*, 6 (6). Retrieved [date] from <http://www.jtla.org>.
- 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., & Gierl, M. J. (2007a). 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. (2007b). Verbal reports as data for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications.* (pp. 146-172). Cambridge, UK: Cambridge University Press.
- 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.
- Leighton, J.P. & Gokiert, R.J. (2008). Identifying test item misalignment using verbal reports of item misinterpretation and uncertainty. *Educational Assessment*, 13, 215-242.
- Mayer, R. (2008). *Learning and Instruction* (2<sup>nd</sup> ed.). Upper Saddle River, NJ: Pearson.
- O'Callaghan, R.K., Morley, M.E., & Schwartz, A. (2004). *Developing skill categories for the SAT Math section.* Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, and R. Glaser (Eds.). Board on Testing and Assessment, Center for Education. Washington, DC: National Academy Press.
- Norris, S.P., Leighton, J.P., & Phillips, L.M. (2004). What is at stake in knowing the content and capabilities of children's minds? A case for basing high stakes tests on cognitive models. *Theory and Research in Education*, 2, 283-308.

- Pellegrino, J. W. (1988). Mental models and mental tests. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 49-60). Hillsdale, NJ: Erlbaum.
- Pellegrino, J. W. (2002). Understanding how students learn and inferring what they know: Implications for the design of curriculum, instruction, and assessment. In M. J. Smith (Ed.), *NSF K-12 Mathematics and Science Curriculum and Implementation Centers Conference Proceedings* (pp. 76-92). Washington, DC: National Science Foundation and American Geological Institute.
- Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practices. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* (pp. 307-353). Washington, DC: American Educational Research Association.
- Poggio, A., Clayton, D. B., Glasnapp, D., Poggio, J., Haack, P., & Thomas, J. (April, 2005). *Revisiting the item format question: Can the multiple choice format meet the demand for monitoring higher-order skills?* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.
- Royer, J.M., Cisero, C.A., & Carlo, M. S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research*, 63, 201-243.
- Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3<sup>rd</sup> Edition., pp. 263-331). New York: American Council on Education, Macmillian.
- Taylor, K. L., & Dionne, J-P. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology*, 92, 413-425.

## APPENDIX A

*A brief summary of the qualifications for the three content specialists who worked on PSAT Mathematics.*

Specialist	Background	Teaching Experience	Test Development Experience
1	PhD, Mathematics	College Professor, 30+ years	College Board Math Consultant, 20+ years, including: <ul style="list-style-type: none"> <li>• Chair, CLEP Precalculus Test Development Committee</li> <li>• Chair, SAT Test Development Committee</li> <li>• Item Writer for CLEP Precalculus, AP Calculus, SAT Math</li> <li>• SAT scale anchoring study</li> </ul>
2	PhD, Mathematics	Mathematics Teacher K-12, 5 years; College Mathematics Professor, 15 years	College Board Math Consultant, 5 years, including: <ul style="list-style-type: none"> <li>• Member, SpringBoard (Web-based Diagnostic Mathematics)</li> <li>• Item Developer, SAT Test Development Committee</li> <li>• SAT scale anchoring study</li> </ul>
3	PhD, Mathematics	College Mathematics Professor, 22 years	College Board Math Consultant, 5 years, including: <ul style="list-style-type: none"> <li>• Member, SpringBoard (Web-based Diagnostic Mathematics)</li> <li>• Item Developer, SAT Test Development Committee (for both College Board and Educational Testing Service)</li> <li>• Item Developer, PSAT Test Development Committee</li> <li>• SAT scale anchoring study</li> <li>• Member, Item Alignment Committee, College Board Standards for College Success and State Math Standards</li> </ul>

**APPENDIX B**

*A brief summary of the qualifications for the three Alberta Education Diagnostic Mathematics staff.*

Specialist	Background	Teaching Experience	Test Development Experience
Manager	B.Ed. (Music, Mathematics); BA. (Music); P.D.A.D. (Education)	Teacher, 20 years, K-12; 15 years as Mathematics Instructor	<ul style="list-style-type: none"> <li>• Grade 3 Math and Language Arts marker, 3 years</li> <li>• Grade 3 Math and Language Arts Examiner, 2 years</li> <li>• Mathematics Exam Manager, 7 years</li> </ul>
Examiner Grade 3	B.Ed. (Secondary Mathematics)	Mathematics Teacher, 32 years, K-12	<ul style="list-style-type: none"> <li>• Grade 12 Diploma Exam Marker in Mathematics, 5 years</li> <li>• Item Writer, Mathematics 30, 5 years</li> <li>• Grade 3 Marker in Mathematics, 4 years</li> </ul>
Examiner Grade 6	B.Ed. (Music); BA. (Music)	Mathematics Teacher, 15 years, K-12	<ul style="list-style-type: none"> <li>• Mathematics Curriculum Specialists for textbook publisher, 1 year</li> <li>• Item Writer, Grade 6 Science, 1 year</li> </ul>

## APPENDIX C

*A brief summary of the qualifications for the seven Alberta Education Diagnostic Mathematics content specialists.*

Specialist	Background	Teaching Experience	Test Development Experience
1—Grade 3	B.Ed. (French)	Mathematics Teacher, 16 years, K-6	<ul style="list-style-type: none"> <li>No test development experience</li> </ul>
2—Grade 3	B.Ed. (Early Childhood, Fine Arts)	Mathematics Teacher, 30 years, K-Grade 12	<ul style="list-style-type: none"> <li>Item Writer, Grade 3 Language Arts, 9 years</li> <li>Item Writer, Grade 3 Mathematics, 9 years</li> <li>Grade 3 Marker in Language Arts, 7 years</li> </ul>
3—Grade 3	B.Ed. (Early Childhood, Language Arts)	Mathematics Teacher, 23 years, K-3	<ul style="list-style-type: none"> <li>Grade 3 Language Arts Assessment, Standard Setting panel member, 3 years</li> <li>Reviewer, Grade 3 Math Items, 3 years</li> <li>Item Writer, Grade 3 Mathematics, 2 years</li> <li>Grade 3 Marker in Language Arts, 9 years</li> </ul>
1—Grade 6	B.Ed. (French), Diploma (Music)	Teacher, 13 years, K-6	<ul style="list-style-type: none"> <li>Reviewer, Grade 6 Math and Science Items, 10 years</li> <li>Item Writer, Grade 6 Math and Science Items, 10 years</li> <li>Grade 6 Marker in Language Arts, 5 years</li> </ul>
2—Grade 6	B.Ed. (Social Studies, Mathematics)	Mathematics Teacher, 13 years, K-6	<ul style="list-style-type: none"> <li>Item Writer, Grade 6 Science, 1 year</li> <li>Item Writer, Grade 6 Mathematics, 1 year</li> <li>Grade 6 Marker in Language Arts, 4 years</li> </ul>
3—Grade 6	B.Ed. (Elementary)	Mathematics Teacher, 16 years, 3-7	<ul style="list-style-type: none"> <li>Curriculum Specialists for textbook publisher, 5 years</li> <li>Item Writer, Grade 6 Language Arts, 1 year</li> <li>Grade 6 Marker in Language Arts, 2 years</li> </ul>
4—Grade 6	B.Ed. (Secondary, Social Studies, French)	Teacher, 23 years, K-6	<ul style="list-style-type: none"> <li>Item Writer, Grade 6 Mathematics, 5 years</li> <li>Item Writer, Grade 6 Social Studies, 1 year</li> <li>Item Writer, Grade 6 Science, 1 year</li> <li>Grade 6 Marker in Language Arts, 1 year</li> </ul>

## APPENDIX D

*A sample of the final five cognitive models for the PSAT Mathematics content area Numbers and Operations produced in Stage #2.*

### Numbers and Operations

The following skills are hierarchically organized, from the most complex (top of the list) to the most simple (bottom of the list).

#### Hierarchy 1

- Applies properties of real number operations (associative, commutative, identity, inverse, distributive, closure, properties of equality and inequality)
- Reads, writes, and evaluates expressions involving natural-number powers of positive integers. \*  
[Note: Also used in an A&F hierarchy]
- Recall basic mathematical facts/definitions about exponential notation such as properties of exponents, scientific notation, place value

#### Hierarchy 2

- Solve problems involving counting techniques.
- Solves problems involving the addition of integers using number lines, patterns, or models.

#### Hierarchy 3

- Create and use ratio, proportions or percents including algebraic expressions in solving problems
- Solve multi-step application problems involving fractions, ratios, proportions, decimals, or percents
- Identify, use, and represent fractions, decimals, or percents in arithmetic and algebraic settings \*
- Solve one step problems involving proportions or rates \*

#### Hierarchy 4

- Solve multi-step word problems involving rates, ratios, proportions, or percents \*
- Translates a verbal statement into a numerical statement.
- Solve one step problems involving proportions or rates \*

#### Hierarchy 5

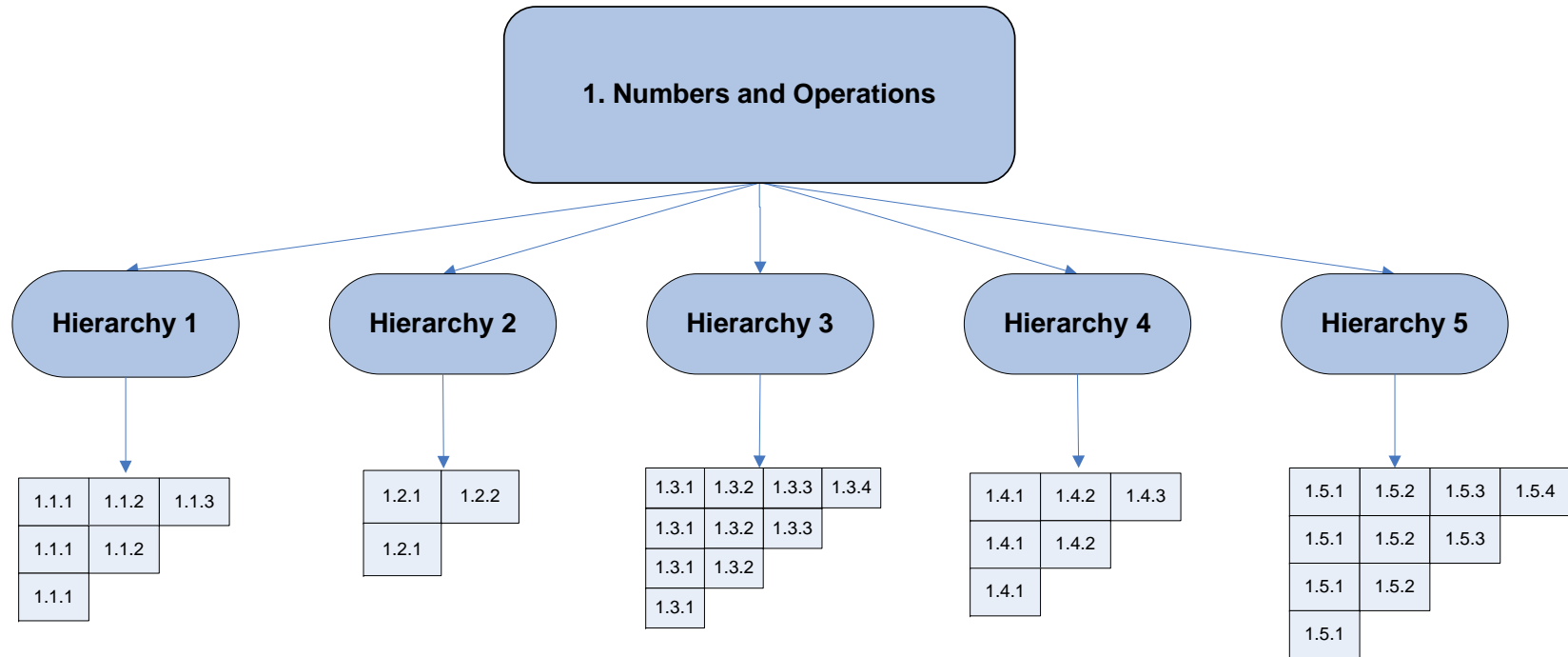
- Applies properties of real number operations (associative, commutative, identity, inverse, distributive, closure, properties of equality and inequality)
- Use properties of real number ordering and the zero-product property \*
- Use properties of inequalities to compare and order integers or rationals \*
- Apply the use of parentheses and order of operations to organize a numerical expression with a mixture of arithmetic operations \*

\*Denotes skill was not placed in the Stage #1 cognitive model

Red denotes skill that is repeated more than once across hierarchies within one page

Blue denotes skill that is repeated more than once across hierarchies across pages

Purple denotes skill that belongs in more than one content category (e.g., Numbers and Operations and Algebra and Functions)



1.1.3 =Applies properties of real number operations (associative, commutative, identity, inverse, distributive, closure, properties of equality and inequality)

1.1.2 =Reads, writes, and evaluates expressions involving natural-number powers of positive integers. [Note: Also used in an A&F hierarchy]

1.1.1 = Recall basic mathematical facts/definitions about exponential notation such as properties of exponents, scientific notation, place value

1.2.2 =Solve problems involving counting techniques

1.2.1 =Solves problems involving the addition of integers using number lines, patterns, or models.

1.3.4 =Create and use ratio, proportions, or percents including algebraic expressions in solving problems

1.3.3 =Solve multi-step application problems involving fractions, ratios, proportions, decimals, or percents

1.3.2 =Identify, use, and represent fractions, decimals, or percents in arithmetic and algebraic settings

1.3.1 =Solve one step problems involving proportions or rates

1.4.3 =Solve multi-step word problems involving rates, ratios, proportions, or percents

1.4.2 =Translates a verbal statement into a numerical statement

1.4.1 =Solve one step problems involving proportions or rates

1.5.4 =Applies properties of real number operations (associative, commutative, identity, inverse, distributive, closure, properties of equality and inequality)

1.5.3 =Use properties of real number ordering and the zero-product property

1.5.2 =Use properties of inequalities to compare and order integers and rationals

1.5.1 =Apply the use of parentheses and order of operations to organize a numerical expression with a mixture of arithmetic operations

## APPENDIX E

A sample of the final three cognitive models for the Alberta Education Diagnostic Mathematics Project in the content area of Number produced in Stage #2.

<b>General Outcome:</b> Develop number sense		
<b>Specific Outcome</b>	<b>Achievement Indicators</b>	<b>Hierarchy</b>
1. Demonstrate an understanding of place value, including numbers that are: <ul style="list-style-type: none"> <li>• greater than one million</li> <li>• Less than one thousandth.</li> </ul>	<ul style="list-style-type: none"> <li>➤ Explain how the pattern of the place value system, i.e., the repetition of ones, tens and hundreds within each period, makes it possible to read and write numerals for numbers of any magnitude.</li> <li>➤ Provide examples of where large and small numbers are used; e.g., media, science, medicine, technology.</li> </ul>	1.1.7 Represent the value of a decimal number to ten thousandths by converting from standard to expanded forms or vice versa 1.1.6 Represent the value of a whole number up to billions by converting from standard to expanded forms or vice versa 1.1.5 Apply understanding of zero as a placeholder for a column to the right or the left of the decimal 1.1.4 Order decimal numbers to ten thousandths from least to greatest or greatest to least 1.1.3 Order whole numbers greater than one million from least to greatest or greatest to least 1.1.2 Apply understanding of place value by representing a given number in a place value chart (numbers from ten thousandths to billions) 1.1.1 Identify the place value of a digit to the left or the right of the decimal (numbers from ten thousandths to billions)
2. Solve problems involving whole numbers and decimal numbers.	<ul style="list-style-type: none"> <li>➤ Identify which operation is necessary to solve a given problem, and solve it.</li> <li>➤ Determine the reasonableness of an answer.</li> <li>➤ Estimate the solution to, and solve, a given problem.</li> <li>➤ Determine whether the use of technology is appropriate to solve a given problem, and explain why.</li> <li>➤ Use technology when appropriate to solve a given problem.</li> </ul>	1.2.7 Justify the solution of a whole number or decimal real-life context problem by applying the appropriate representation 1.2.6 Verify the reasonableness of a response by applying another appropriate strategy for a real-life context problem 1.2.5 Solve multiple step whole number or decimal real-life context problems using appropriate strategies (apply choice of operation) 1.2.4 Solve one-step whole number or decimal real-life context problems using appropriate strategies (apply choice of operation) 1.2.3 Apply estimation to predict possible solutions to a real-life context problem involving whole numbers and decimals

		<p>1.2.2 Identify the operation (s) required (addition, subtraction, multiplication or division) to solve a real-life context problem involving whole numbers up to billions and decimals to ten thousandths</p> <p>1.2.1 Identify relevant information in a real-life context problem that uses numbers ranging from ten thousandths to billions</p>
<p>3. Demonstrate an understanding of factors and multiples by:</p> <ul style="list-style-type: none"> <li>• determining multiples and factors of numbers less than 100</li> <li>• identifying prime and composite numbers</li> <li>• Solving problems using multiples and factors.</li> </ul>	<ul style="list-style-type: none"> <li>➤ Identify multiples for a given number, and explain the strategy used to identify them.</li> <li>➤ Determine all the whole number factors of a given number, using arrays.</li> <li>➤ Identify the factors for a given number, and explain the strategy used; e.g., concrete or visual representations, repeated division by prime numbers, factor trees.</li> <li>➤ Provide an example of a prime number, and explain why it is a prime number.</li> <li>➤ Provide an example of a composite number, and explain why it is a composite number.</li> <li>➤ Sort a given set of numbers as prime and composite.</li> <li>➤ Solve a given problem involving factors or multiples.</li> <li>➤ Explain why 0 and 1 are neither prime nor composite.</li> </ul>	<p>1.3.1 Solve a given problem from a real life context using multiples or factors less than one hundred</p> <p>1.3.1 Identify the factors for a given number less than one hundred</p> <p>1.3.1 Represent with arrays all whole number factors of a given number less than one hundred</p> <p>1.3.1 Apply the process of skip counting to 100 to determine the multiples of a given number</p>