"Can you do addition?" the White Queen asked. "What's one and one and one and one and one and one and one and one and one and one?"
"I don't know," said Alice. "I lost count."

Through the Looking Glass

University of Alberta

Making Diagnostic Inferences about Student Performance on the Alberta Education Diagnostic Mathematics Project: An Application of the Attribute Hierarchy Method

by

Cecilia Brito Alves

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Measurement, Evaluation and Cognition

Department of Educational Psychology

Spring 2012
Edmonton, Alberta

To my father (in memoriam)

“Oh, pedaço de mim
Oh, metade afastada de mim
Leva o teu olhar
Que a saudade é o pior tormento
É pior do que o esquecimento
É pior do que se entrevar”
(Chico Buarque)

Abstract

Cognitive diagnostic assessments (CDA) is an approach where the psychology of learning is combined with methods and models in statistics for the purpose of making inferences about students' specific knowledge structures and processing skills. This study used a four-step principled approach to test design characterized by: (1) the development of cognitive models, (2) the construction of test items according to the knowledge and skills specified in the cognitive model, (3) the use of a diagnostic psychometric analysis to assess the plausibility of the underlying cognitive model and to providestudents' attribute probability estimates, and (4) the creation of detailed score reports that map examinees' mastery levels to provide more detailed information about students' problem-solving strengths and weaknesses.Being among the first applications of the AHM to non-retrofit data from an operational testing program, the findings of this study add substantially to our understanding of the necessity of a principled approach to assessment design, and also contribute to a growing body of literature on CDA. Results of this study revealed that cognitive models adequately fit the data for the total sample of students; however, the fit for the observed and expected response data differed for high and low ability students. The average attribute probability estimates were ordered, as expected, from least to most difficult. In addition, the ordering of the attributes did not differ as function of the performance level of the students and the correlational pattern of the probability estimates indicated both convergent and discriminant evidence supporting the hierarchical structure of attributes. Concerning the reliability of the models, all six attributes in Subtracting

2-digit numerals produced consistent interpretations about the mastery of attributes, whereas Comparing and ordering numbers, only the decisions made for Attribute 1 were found to be consistent. Limitations of the study and recommendations for future research were also discussed.

Acknowledgement

What shall I render to the Lord for all His benefits toward me? (Psalm116)

First and foremost, I would like to give thanks to the Almighty Lord who has guided me through this doctoral program. Jesus was my sustainer in moments of weakness, comfort in moments of sadness, friend when I was lonely. Thank you for saving my life, loving me, and guiding me along this life's journey.

I would also like to acknowledge the following people and groups who greatly contributed to the completion of my doctoral program:

I was blessed with a brilliant and supportive adviser, Dr. Mark Gierl. Since my first day at the University of Alberta, I knew I had made the right decision to cross the world to be supervised by him. I would therefore like to thank him for his constant support and guidance during these five years. I thank him for countless hours of discussion and patient manuscript revision. His valued feedback greatly enhanced this document. Mark, you are a model of competence, commitment, thoughtfulness, and kindness to me. Thank you for showing me, through your life, the kind of professor I want to be. There are no words that I can find to truly express my gratitude and admiration for you.

I would also like to thank my examining committee, Dr. Ying Cui, Dr. Jacqueline P. Leighton, Dr. George Buck, Dr. Rebecca Gokiert, and Dr. Eric Frenette for all their time and valuable comments.

I would like to thank Alberta Education for permitting me to use data from the Diagnostic Mathematics project. I would like to acknowledge the work and support of Renate Majeau, Bev Dekker, and Margo Cahn. Their help was instrumental for the successful completion of this research. Thanks Bev for many insightful conversations.

It is with deepest gratitude that I thank Education Ministry of the Brazilian Government (CAPES, proc. no. 343505-9) for the financial support provided throughout my Ph.D. Without their support, this doctorate would not have been possible.

I am very grateful to my colleagues at the Centre for Research in Applied Measurement and Evaluation (CRAME) for their advice and for their constant support throughout my doctorate. I would especially like to show my appreciation to Oksana, Mary, Andrea, Jiawen, Ken, Hollis, Xian, and Louise. Your friendship and support made these years a joy.

I would like to thank professor Neto who inspired me to pursue a doctoral degree, and whose friendship and enthusiasm helped me through it. Thank you very much, Adson, for calling Mark when I received the CAPES scholarship. That call changed the course of my life. Without you, I would not have been able to come to this university. I also want to thank all my friends at University of Brasilia and CESPE. I keep in my heart so many sweet memories from the time we spent together. Additionally, I would like to thank Maria Alice, Luciene, and Valeria for their constant love and friendship. I want also to thank Julie Ciancio for revising my proposal and dissertation. I want to thank my church family, especially sister Ruth & Pastor Pedro Pires, and Lisa & Steve Pollard, for their prayers, care and support. I would also like to thank all the friends I have made while at the U of A. Thank you Rose & Mauricio, Maira & Frederico, Carla & Leandro, Claudia & Malcolm, Ines & Alexandre, Luciane, Cibele, Ana& Byron, Renata & Lucca, Ana & Filipe, Fernanda, Susan, Denise, Anna, Carol & Andre, Na & Ed, Ranilce & Valentin, Bia, Rita & Alex, Raquel & Chad, Patricia & Fernanda, Anelise & Tiago, Leticia & Derek, Suellene & Andre, Suerda & Eraldo, and many others. You all are my family in Canada. Thanks to all for their joyful company and many treasured memories in the last five years.

I would like to thank my family, nieces and nephews, brothers- and sisters-in -law, and my siblings–Dora, Kennedy, Scheyla, Alfredo, Robson, and Shirley–for their tireless love and support. I thank God for my beautiful and lovely mom, to whom I have endless love, respect, and admiration. Her love and her prayers encouraged me to do my best in my Ph.D. I could never have done this without her constant support and unconditional love. My dad started the Ph.D. with me, but unfortunately he could not see my graduation. I miss him terribly, but his outspoken love and pride in me kept me going. I love you always and forever!

I also want to thank my parents-in-law, whom I came to love and care for. I am greatly indebted to my devoted mother-in-law who spent almost four months with us so I could finish writing my dissertation and to my father-in-law for being apart of his lovely wife for so long. You are such a wonderful example of unselfishness, love, and care. I will never forget this amazing act of love from you two.

Last but not least, I would like to greatly thank my beloved husband, Isac. He is certainly the best thing that ever happened to me. With you I have lived the best years of my life. I thank you for believing in me when I lacked confidence and for encouraging me to try even harder. Your love and support has enabled me to complete this degree. Thank you very much also for the most precious treasure you could ever give to me: our little angel, Alice. I owe my every achievement to both of you. I deeply love you both. “*E todo mundo diz que ele completa ela, e vice-versa, que nem feijão com arroz*”[1]

.

[1] From Legião Urbana’s song “Eduardo e Mônica”.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AE: Assessment Engineering

AHM: Attribute Hierarchy Method

CDA: Cognitive Diagnostic Assessment

ECD: Evidence Centered Design

HCI: Hierarchy Consistency Index

SAT: Scholastic Aptitude Test

TIMMS: Trends in International Mathematics and Science Study

CHAPTER I: INTRODUCTION

Cognitive assessment is now recognized as an important way to improve the quality and the validity of score interpretation (Cui & Leighton, 2009; Embretson & Gorin, 2001). It has also been recognized as one of the great challenges for the field of measurement and evaluation (Snow & Lohman, 1989; Pellegrino, Chudosky, & Glaser, 2001). By improving the quality of score interpretations, valuable information about students who may be at-risk for failure and the delivery of carefully designed remediation programs are made possible.

Traditionally, educational assessment has been used for the single purpose of measuring student success on tests and exams at the end of an instructional unit. Students receive feedback such as an overall mark or grade, which provides only a general indication of their overall mastery level. This type of feedback does not yield specific inferences about the examinees' strengths and weaknesses. Because traditional assessments are often used for summative purposes, this approach to assessment is seen as essentially passive and not having immediate impact on learning (Anderson, 1998; Sadler, 1989; Struyven, Dochy, & Janssens, 2003, 2008). Traditional assessments have also been related to testing only lower-order thinking skills (e.g., memorizing facts; Law & Eckes, 1995; Simonson, Smaldino, Albright, & Zvacek, 2000) and, consequently, providing limited feedback about how to improve a student's performance (Bailey, 1998).

As a result of the existing flaws in traditional assessment, this approach to testing has been met with reservation by some test users. In addition, the context of assessment, as of late, has been marked by the demand for new forms of assessments that can provide useful information to teachers and students. Huff and Goodman in a chapter of the book *Cognitive Diagnostic Assessment for Education* (2007) discuss the recent demand for cognitive diagnostic assessment in K-12 education. These authors claim that "much of the demand for CDA [cognitive diagnostic assessment] originates from discussion about the potential to inform teaching and learning by changing the way in which we design assessments" (p. 21). To investigate the extent that results from large-scale assessments were being used, Huff and Goodman conducted a national survey with mathematics and English language teachers in the United States. Results showed that a large proportion of teachers who received assessment results rarely or never use them to inform instruction.

Ideally, assessment should inform instruction, providing teachers and students with a clear appreciation about what learners understand and what gaps in their knowledge still exist. It is widely accepted that the way students are assessed has a major influence on their learning (Black & Wiliam, 2001; Garfield, 1994; Gibbs & Simpson, 2004; Stiggins, 2002). Most importantly, when adequately conducted, assessments have the potential to help motivate students, and to empower them to take control of their own learning (Tanner & Jones, 2003).

This context constitutes a rich ground for the discussion about the importance of cognitive diagnostic assessments and how they can be used to improve learning.

Educational Measurement and Cognitive Diagnostic Assessments

Cognitive diagnostic assessment (CDA) can be concisely described as an educational test for measuring learners' cognitive processes, learning, knowledge and skill development for diagnostic purposes (Ketterlin-Geller & Yovanoff, 2009). CDA is used to ascertain whether a student possesses specific knowledge and skills required to solve problems in a particular domain. By establishing a profile of students' cognitive strengths and weaknesses, the instructor has the means to remediate and to adjust program plans to meet each pupil's unique needs. CDA is also seen as an approach where the psychology of learning and the practices in assessment are gathered together for the purpose of making inferences about students' specific knowledge structures and processing skills (Nichols, 1994; see also Nichols, Chipman, & Brennan, 1995). Huff and Goodman (2007) define CDA as "the joint practice of using cognitive models of learning as the basis for principled assessment design and reporting assessment results with direct regard to informing learning and instruction" (p. 20).

These definitions highlight the importance of CDA for providing richer and more useful feedback to teachers and students as well as feedback that helps them to identify learning problems and remediate these problems (Gierl, Leighton, & Hunka, 2007). Also, the use of cognitive diagnostic assessments promises, in Nichols and Joldersma's words, "to provide teachers the kind of formative information with which to lever higher student achievement" (2008, p. 407). The term "formative" is used here to mean that the "results of the assessment are used to directly support teaching and learning, as contrasted with summative testing, which evaluates the student after the instruction is over" (DiBello, Roussos & Stout, 2007, p. 285).

As acknowledged by DiBello and colleagues (2007), "there is rapidly emerging a powerful need, and demand, for tests designed to formatively assess an appropriately chosen moderate number of relatively fine-grained chunks of knowledge in major subject or important cognitively defined areas" (p. 285). CDA results, when properly used, have the potential to guide the design of remedial instruction and the placement of students into supplemental intervention programs (Ketterlin-Geller & Yovanoff, 2009), as well as to empower students to reconsider their study strategies on the basis of CDA's feedback. Hence, outcomes from CDA may significantly impact the educational opportunities available to students, as these outcomes can be used to yield valuable information that can be used by practitioners to tailor remediation (de la Torre & Douglas, 2004).

Next, I will discuss how the combination of cognitive psychology and psychometric principles in the design of cognitive diagnostic tests may improve the inferences about students' strengths and weaknesses.

Cognitive Models and CDA

Although cognitive psychology is exerting its influence on testing practice, the investigation of the underlying learning processes has been neglected in most contemporary large-scale testing programs (Zhou, 2010). More attention has been paid to psychometric techniques for analyzing data than to the underlying knowledge and response processes students use to produce the data (Leighton, Gierl, & Hunka, 2004; Gorin, 2006). Therefore, information about why students perform poorly and how assessment outcomes can be utilized to improve teaching and learning has been limited.

Understanding students' knowledge acquisition and cognitive processes is essential for diagnosis since it may enhance test validity and reliability. Hence, valid measures of students' performance and learning processes help the enhancement of both instruction and learning. However, as claimed by DiBello et al. (2007), "much of the research and development work needed for skills-level formative assessment to truly flower remains to be carried out" (p. 286).

In order to gather diagnostic information, the use of a cognitive model is required. This model is intended to guide item development, allowing examinee's performance to be linked to specific cognitive inferences about their knowledge, processes, and strategies. In this way, the CDA approach—combining theories of cognition with models in statistics—is a valuable approach for making inferences about students' strengths and weaknesses on particular cognitive skills or attributes[2]. Tatsuoka and Tatsuoka (1997) define attributes as the cognitive processing and knowledge required for solving a problem in a target domain. Cognitive models are formed by different combinations of attributes. A cognitive model is defined by Leighton and Gierl (2007a) as "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" (p. 6). The expression

[2] The term "attribute" is used interchangeably with the term "skill" in this proposal.

"simplified description" is used to indicate that cognitive models serve as a simplified representation of a much more complex phenomenon. This representation reflects a small but key set of issues required to understand the phenomenon (Leighton & Gierl, 2011). Cognitive models have also been related to how people develop structures of knowledge, including the concepts associated with a domain and procedures for reasoning and solving problems (Pellegrino et al., 2001). Gierl, Roberts, Alves, and Gotzmann (2009) present four defining characteristics of cognitive models for CDA: granularity, measurability, hierarchy of ordered skills, and instructional relevancy. These four defining characteristics are explained in more detail in Chapter 2. Defining characteristics are critical for a better fit between the data and the cognitive model since they constitute the bridge between the cognitive model and the test items.

Different conceptual frameworks have been used to develop cognitive models. The use of a framework has the potential to enhance a deep understanding of the target construct by providing a structure for designing studies, interpreting data, and drawing conclusions (Eisenhart, 1991). In this dissertation, three frameworks–theoretical, content specialist, and combined–are described and discussed. The theoretical framework guides research activities by drawing on a formal theory, developed using an established, coherent explanation of certain types of phenomena and relationships. In the content specialist framework, the accumulated practical knowledge from experts' experience with students, curriculum, and learning environment serves as basis for cognitive model development. Finally, in the combined framework, research activities are based not only on previous research and literature, but also on an array of different sources, such as outcomes from theories, expert knowledge, empirical research, or other sources relevant to a research problem. Factors to consider when choosing a framework are, for example, the availability of human expertise as well as time and financial resources allocated for research and assessment implementation. Each of these three frameworks has benefits and limitations, as will be discussed in Chapter 2.

Applications of CDA, which entail the identification of cognitive attributes involved in learning, have been recognized as a way of bringing psychometrics and cognitive science together. This process is explained by Nichols and Joldersma (2008):

> The measurement model connects the knowledge, skills, and abilities of the construct to the observable performance on the test. The measurement model summarizes across tasks the evidence of the knowledge, skills, and abilities that is provided in the observable performance (p. 408).

Because the development of cognitive models is a laborious task, the number of practical applications of CDA has, so far, been relatively small. Consequently, most of the research conducted using this approach has been focused on retrofitted data. Retrofitting is a method based on the revision of existing test items with the goal of extracting information about the cognitive attributes measured by these items, even when the items were not initially developed from a cognitive perspective. In this way, retrofitting the cognitive model to existing test items is considered a *post hoc* procedure.

Although CDAs have great potential to facilitate instruction by identifying learners' strengths and weaknesses, the successful application of CDA may require new test development procedures and practices that will promote a closer match between test items and the cognitive model. Consequently, this match will foster more valid inferences about learners' knowledge and skills. By implication, retrofitting CDA to existing educational data is likely to yield unsatisfactory diagnostic classification results (Gierl, Alves, & Taylor-Majeau, 2010).

To prevent the problems associated with retrofitting models and items, an organizational structure required to implement a cognitive diagnostic assessment is discussed in this dissertation. This structure is based on a four-step procedure described as a *principled test design approach* where: (1) attributes are first identified and then structured to characterize the cognitive model underlying examinees' problem-solving skills, (2) test items are constructed to measure the attributes in the cognitive model, (3) a confirmatory psychometric procedure is used to analyze student response data from the model-based test items, and (4) a detailed score report is provided.

Although the importance of a principled test design approach is accentuated by the fact that the cognitive model provides the interpretative framework to guide both the development of items and the interpretation of examinees' scores, conducting these four steps "is not a standard approach to test design and it rarely, if ever, is used in operational testing situations" (Gierl, 2007, p. 337).

The Alberta Education Diagnostic Mathematics Project

The test created for the Diagnostic Mathematics project is an example of an assessment originally designed to follow the four-step process for principled test design. This assessment—conceived to be a diagnostic, cognitive-based, and non-retrofitted exam—started in January 2008 and it is funded by the Learner Assessment Branch at Alberta Education. The purpose of the assessment is to provide teachers with diagnostic information about students' cognitive knowledge and skills in Mathematics from Kindergarten to Grade 6. This project is based upon CDA principles, where the test score interpretations are linked to the cognitive skills required to solve the test.

This assessment is implemented with a computer-based, online administration system, designed to provide timely score reports to students and teachers that can support learning and instruction (Alberta Education, 2007). The cognitive model for the Diagnostic Mathematics assessment has its origin in the provincial curriculum adopted in Alberta (see *The Alberta K-9 Mathematics Program of Studies with Achievement Indicators*, 2007). Cognitive attributes and their hierarchical structure have been outlined for four content areas: Number, Patterns and Relations, Shape and Space, and Statistics and Probability at two grade levels, 3 and 6. This is an ongoing project and, to-date, only the content area Number has been tested. The Number strand, in Grade 3, has 13 cognitive models with 83 skills and in Grade 6 it has nine cognitive models with 72 skills.

Consistent with the goal of providing diagnostic information about students’ strengths and weaknesses in Mathematics, Alberta Education has created the Diagnostic Mathematics assessment with the intention of providing valid and accurate information on students’ mastery of cognitive skills in the Alberta curriculum. To meet this requirement, the skills measured by the test items must be well understood. To demonstrate the benefits of using the four-step to principled test design, my dissertation research was conducted with the purpose of analyzing this assessment using the AHM (Attribute Hierarchy Method), a cognitively-based psychometric approach.

Purpose of the Study

The purpose of this study is to investigate the accuracy and consistency of two cognitive models in the Diagnostic Mathematics project using the attribute hierarchy method. The accuracy and consistency by which cognitive diagnostic assessments classify students’ test responses are key components to providing diagnostic feedback about examinees’ knowledge and skills (Zhou, 2010). Specifically, the study is designed to answer the following research questions:

1. How does the observed student response data fit the expected response data produced by the cognitive models created by content specialists? It is also important to begin to evaluate how these results generalize across different subgroups of examinees that may differ in their response processes. Hence, I will also address the question: Does the fit for the observed and expected response data differ for high and low achieving students?

2. How reliable are decisions about the mastery of specific attributes for the students who wrote the diagnostic test? Do the reliability estimates differ for high and low achieving students?
3. Are the attribute probability estimates ordered from easy to difficulty across a student sample for each of the attribute hierarchies? Is the attribute order the same for high and low achieving students?
4. Does the correlations among attributes show convergent and discriminant evidence supporting the hierarchical structure of attributes? Do the correlational patterns different for high and low achieving students?

Organization of the Dissertation

This dissertation is organized into five chapters as follows. The first chapter (which is the current chapter) includes a short description of some limitations of traditional approaches to educational measurement, a summary of the relationship between cognitive models and CDA, an overview of the Diagnostic Mathematics assessment, and the statement about the purpose of the study. The second chapter contains the literature review of the study. This chapter describes the traditional context of test use, the demand for new assessment techniques from test users, the importance of CDA in the context of educational measurement, characteristics of cognitive models for CDA, methods for developing these models, steps for implementing a CDA, and a review of CDA studies in the domain of Mathematics. The third chapter describes the implementation of the four-step for principled test design: the cognitive model development, item development, confirmatory psychometric analyses, and score reporting. In this chapter, I also describe the sample and data collection design as well as the proposed procedures and statistical analyses. The fourth chapter presents the results from the AHM analysis. Chapter VI discusses the results, draws the conclusion and highlights the limitations of the study. Directions for future research are also presented.

CHAPTER II: LITERATURE REVIEW

Demand for New Assessment Techniques from Test Users

Increasingly, the demand for new ways to assess students is appearing in key policy statements. For example, the *No Child Left Behind* Act of 2001 calls for diagnostic information to be provided for each individual student, along with information for the parents, teachers, and principals to use in addressing individual student educational needs. The necessity to provide innovative assessments is also recognized by the Board on Testing and Assessment (BOTA, 2009). In a Letter Report to the U. S. Department of Education on the Race to the Top Fund (RTT), they state:

> Because of the extensive focus on large-scale, high-stakes, summative tests, policy makers and educators sometimes mistakenly believe that such tests are appropriate to use to provide rapid feedback to guide instruction. This is not the case. Tests that mimic the structure of large-scale, high-stakes, summative tests, which lightly sample broad domains of content taught over an extended period of time, are unlikely to provide the kind of fine-grained, diagnostic information that teachers need to guide their day-to-day instructional decisions. (p. 10-11).

The RTT fund of the American Recovery and Reinvestment Act of 2009 (74 Fed. Reg. 37804, proposed July 29, 2009) encouraged the inclusion of a rapid-time turnaround system, where they defined rapid-time turnaround as follows:

> Rapid-time, in reference to reporting and availability of school-and LEA [Local Education Agency]-level data, means that data is available quickly

> enough to inform current lessons, instruction, and related supports; in most cases, this will be within 72 hours of an assessment or data gathering in classrooms, schools, and LEAs (Section IV, p. 37811, in BOTA, 2009).

The Department of Education supports the use of statewide longitudinal data systems through the RTT fund so that the data will help inform and engage, as appropriate, key stakeholders (e.g., parents, students, teachers, principals, LEA leaders, community members, unions, researchers, and policymakers). Also, the data supports decision makers in the continuous improvement of instruction, operations, management, and resource allocation (Section III, p. 37809, in BOTA, 2009).

The National Association for the Education of Young Children (NAEYC) and the National Council of Teachers of Mathematics (NCTM) recognized that “well-conceived, well-implemented, continuous assessment is an indispensable tool in facilitating all children's engagement and success in mathematics” (NAEYC, 2002, p. 10). Several researchers have suggested that the assessments currently being implemented have not addressed the issues highlighted by the BOTA, RTT, NAEYC, and NCTM using the traditional approach to educational testing (e.g., Cui, Leighton, & Zheng, 2006; Dietal, Herman, & Knuth, 1991; Dikki, 2003; Frisby, 1999; Mislevy, 1994; Nichols, 1994). Mitchell (1992, as cited in Dikki, 2003) defines traditional assessment as a “single-occasion, unidimensional, timed exercise, usually in multiple-choice or short-answer form”. This type of assessment has dominated most classroom assessment activities (Earl, 2003). Students receive feedback such as overall mark or grade, which provides a general albeit coarse indication of their overall mastery level. However, this feedback has limitations when attempting to understand students' learning processes.

At least four important limitations of the traditional assessment approach can be identified. First, traditional assessments often measure a students' overall summative proficiency at the conclusion of an instructional module or unit in a particular content area. Further, according to Sadler (1989), this type of assessment is concerned with:

> summing up or summarizing the achievement status of a student, and geared towards reporting at the end of a course of study especially for purposes of certification. It is essentially passive and does not normally have immediate impact on learning, although it often influences decisions which may have profound educational and personal consequences for the student (p. 120).

Second, in the traditional assessment approach, one single score is provided to students: the total score. This score is usually represented as a mark or letter grade that summarizes the *average* performance across several topics in a specific content area. Total scores do not necessarily indicate which cognitive skills were mastered by the examinees and which attributes[3] were deemed as weaknesses. Furthermore, total score often "obscures important diagnostic information about more fine-grained attributes that students use to solve problems within some given domain" (Briggs & Alonzo, 2009, p. 9). Thus, if the objective is to know whether

[3] Attributes can include different procedures, skills, and/or processes that an examinee must possess to solve a test item.

a student masters a specific skill or strategy, a total score on a set of items—calculated using the performance across several different topics—will likely not help (Junker & Sijtsma, 2001). Consequently, specific inferences about the examinees' strengths and weaknesses cannot be made.

Third, the traditional assessment approach provides only a superficial understanding of the knowledge and skills examinees' use to solve items on the test. Some researchers have criticized the traditional approach by asserting that most tests assess only lower-order thinking skills (i.e., the learner's ability to memorize and recall discrete facts and knowledge; Law & Eckes, 1995; Simonson, Smaldino, Albright, & Zvacek, 2000). Several researchers claim that content-based tests provide only vague evidence of the process and mechanisms test takers use in responding to test items (Nichols, 1994), thus, students are held accountable only for *what* they know rather than considering *how* they know it (Briggs, 2007). Leighton and Gierl (2007a) claim that assessments derived from content specifications do not usually include a critical mass of items that allow the proper measurement of any particular skill. They argue that this deficit occurs because traditional assessments are designed to measure many different behaviors within a short time, but not with enough frequency or depth. As a result, it is difficult to make inferences about students' specific cognitive processes and problem-solving skills (see also Kato, 2009).

Fourth, score reports based on the traditional assessment approach often provide limited feedback about how to improve a student's performance (Bailey, 1998). The fact that only a single test score is provided to students, in addition to the fact that skills are only superficially measured, contributes to the limited nature of the assessment feedback. These score reports also fail to capture the breadth and richness of the knowledge and skills that could help students learn and teachers instruct more effectively (NRC, 2001). This limited feedback may also be due, in part, to the fact that items are written using general content specifications and test construction is driven solely by examining psychometric properties of those items (Embretson & Gorin, 2001). Given that the establishment of what the test measures is frequently conducted *after* the test administration, in an exploratory approach (as opposed to a confirmatory manner), the inferences supported by the score reports are generally too vague and too late to inform instruction (Kato, 2009).

Because of these four important limitations, some researchers deem traditional assessments as ineffective for promoting instruction and learning in a formative context (Bailey, 1998; Simonson, Smaldino, Albright, & Zvacek, 2000). Test users have also come to recognize that understanding student knowledge structures and processing skills–and how these knowledge and skills may change with instruction–requires continuous diagnostic feedback (Nichols, 1994). The demand for more useful information about students’ strengths and weaknesses provides an opportunity for a new area of psychometric research based on the use of diagnostic testing and feedback. One solution to overcome the limitations of the traditional assessment approach is a new form of assessment called cognitive diagnostic assessment (CDA).

Cognitive Diagnostic Assessment and Educational Measurement

Over the last two decades, many researchers have argued that advances in the cognitive and measurement sciences could provide a powerful basis for improving educational assessment (e.g., Baker, 1997; Cui & Leighton, 2009; Gierl, 2007; Messick, 1984; Mislevy, 1994; Nichols, 1994; Pellegrino, Baxter, & Glaser, 1999; Snow & Lohman, 1989). Pellegrino et al. (2001) assert that this merger could be mutually beneficial, with the potential to catalyze further advances in both fields. That is, reviewing advances in the sciences of how people learn and how such learning can be measured offer the potential for a much richer and more coherent set of assessment practices in the learning sciences (Pellegrino et al., 2001).

In addition to meeting psychometric standards, a cognitive diagnostic assessment would also yield specific information regarding the individual examinee's educational needs (McGlohen, 2004). Cognitive diagnosis, as defined by Ohlsson (1986), is the process of inferring a students' cognitive state from his or her test performance. More recently, cognitive diagnostic assessment has been defined as an educational test designed to measure students' cognitive processes, learning, knowledge, and skill development for diagnostic purposes (Ketterlin-Geller & Yovanoff, 2009). This approach uses the research from the learning sciences to structure psychometric models so inferences can be made regarding the structure and processes that underlie students' test performances (Nichols, 1994).

Researchers also indicate that using CDAs may help overcome some limitations of the traditional approach to educational testing (Nichols, 1994; Cui, Leighton, & Zheng, 2006). CDA has at least four important strengths. The first strength of this approach is that results obtained using CDAs, in addition to the summative purpose (CDA can indicate which skills have not been mastered as a result of instruction), can also promote formative inferences about student learning (revealing, for example, which skills should be taught in the subsequent instructional cycle). Formative inferences allow teachers to redesign instructional approaches, evaluate instructional resources, and remediate students' weaknesses (Jang, 2008). Nichols and Joldersma (2008) claimed that CDA "promises to provide teachers the kind of formative information with which to lever higher student achievement" (p. 407). The use of this formative approach is also viewed as a way to promote student engagement in learning by encouraging them to use assessment as a learning tool (Jang, 2008).

The second strength of CDA is the fine grain size of the diagnostic scores, as opposed to the coarse grain size of the total score provided by the traditional approach. Rather than assigning examinees a single score, CDA provides examinees with a profile of diagnostic scores that yield information about "whether or not they have mastered each of a group of specific, discretely defined skills, or attributes" (Huebner, 2010, p. 1). This approach provides opportunities for students' to enhance their learning by receiving detailed and precise information about their cognitive problem-solving strengths and weaknesses in a particular content area or area of study.

The third strength of CDA is that it provides richer and more detailed information about the knowledge and skills measured by an educational test. According to Nichols (1994), the substantive assumptions regarding the processes and structures a student uses in a testing situation, as well as how the knowledge structures develop and how students at various ability levels differ, are made explicit. As CDAs provide specific and detailed information about students' knowledge and skills, we may also gauge a student's readiness to move on to higher levels of understanding and skill acquisition in a given domain (Gott, 1990). This type of diagnostic feedback is possible because CDAs offer a more complex understanding of students' problem-solving skills and a more direct method for evaluating a students' strategic knowledge. Cui and Leighton (2009) claim that empirically analyzing test responses to confirm the cognitive model used in test design can also strengthen the validity argument about the construct being measured, which helps clarify the psychology that underlies test performance and provide more interpretable and meaningful test scores. By identifying and measuring complex cognition, knowledge structures, and processing skills, one can connect test performance to richer test score interpretations, which may help instructors and students design strategies to improve the teaching and learning processes.

The fourth strength of CDA is the potential to enhance the quality of diagnostic feedback provided to students. The CDA score report yields detailed information about what knowledge and skills were measured by the test and the degree to which the examinee has mastered these knowledge structures and processing skills. The reports also provide specific diagnostic feedback that has the potential to guide instructors, parents, and students in the teaching and learning process. In addition to the potential for identifying particular strengths and weaknesses, diagnostic outcomes can also help teachers highlight what a student has learned, detect incomplete and incorrect knowledge states, and decide how instruction needs to be adapted to meet the needs of the student (Ye, 2005). Moreover, diagnostic feedback may encourage students' involvement in the learning process by providing reporting tools that effectively identify their needs thereby allowing the students to identify strategies to foster learning and promote educational improvement.

However, in order to make detailed and useful inferences about an examinee's cognitive strengths and weaknesses, it is crucial that the assessment be based on a cognitive model, of some type. A cognitive model provides the interpretative framework to guide both the development of items and the interpretation of examinees' scores so test performance can be linked to specific inferences about examinees' knowledge and skills (Gierl, Alves, & Taylor-Majeau, 2010). A cognitive model is defined by Leighton and Gierl (2007a) as a simplified description of human problem-solving at some convenient grain size or level of detail that facilitates explanation and prediction of students' performance. Hence, cognitive models have been related to how people develop structures of knowledge, including the concepts associated with a domain, and procedures for reasoning and solving problems in that domain (Pellegrino et al.,2001).

Thus, in order to develop diagnostic assessment tasks and to generate diagnostic scores, a cognitive model must be identified and evaluated, items must be developed to measure the knowledge and skills in the cognitive model, and confirmatory model-based psychometric procedure must be used to score the data (Gierl & Cui, 2008). Leighton (2008) asserts that cognitive psychology, cognitive sciences, and/or the learning sciences are "supposed to guide the application of diagnostic cognitive models and to contextualize inferences drawn from their results" (p. 272-273). She also claims that "the underlying cognitive demands of the test, or the lack thereof, go unquestioned" (Leighton, 2008, p. 274). According to Gorin (2006) research and development of sophisticated statistical methods to model cognitive information has increased, but also claims that equal attention must be placed on the cognition behind the statistics. She states:

> The application of these cognitive-psychometric methods is fruitless if the tests to which they are applied lack a formal cognitive structure. If assessments are to provide meaningful information about student ability, then cognition must be incorporated into the test development process much earlier than in data analysis (p. 21).

A similar claim was made by Ye (2005) who declared that "to diagnose how well students learn, it is important to know beforehand how students learn" (p. 21). These statements indicate the need for continuous research to advance the understanding about the fundamental psychological processes underlying knowledge and skill acquisition. The terms *knowledge* and *skills* are used to represent, respectively, declarative knowledge (knowing what), and procedural knowledge (knowing how). Winterton, Delamare Le Deist, and Stringfellow (2006) claim that the acquisition of declarative knowledge (factual knowledge) must precede the development of procedural knowledge (skills).

Even though our understanding about how knowledge is organized, how children develop conceptual understanding, and how expertise is acquired in specific subjects has greatly advanced in the last 20 years (cf. Pellegrino et al., 2001), much work remains. The structures and processes required for learning need to be more clearly understood, so a sound theoretical foundation to diagnostic assessments can be developed. A cognitive model is necessary in order to generate specific diagnostic inferences underlying test performance. The knowledge and skills, also called *attributes*, are specified at a small grain size in order to generate specific diagnostic inferences underlying test performance. Attributes can include different procedures, skills, and/or processes that an examinee must possess to solve a test item[4]. Hence, cognitive models provide the interpretative framework for linking test score interpretations to cognitive attributes.

The use of a psychological perspective to think about cognition, learning, and cognitive models provides a framework for deciding “what to see as a problem, how to think about them, and how to solve them” (Mislevy, 2006, p. 261). According to Gierl and Leighton (2007), the information-processing perspective is required to model the psychology of test performance in CDA.

[4] This definition is fairly similar to the way Mislevy and Riconscente (2006) use KSAs (Knowledge, Skills, and Abilities), where they credited industrial psychologists to the use the acronym to refer to the targets of the inferences they draw about someone.

Mislevy (citing Rupp & Mislevy, 2007) purports a similar idea when he claims that cognitive diagnostic models often draw inferences about students under this perspective.

Cognitive Psychology and the Information-Processing Perspective

Information processing is a dominant psychological perspective used by cognitive psychologists (Casey & Moran, 1989; Palmer & Kimchi, 1986; Pashler, 1995; Solso, MacLin, & MacLin, 2005). According to Berg (2000), the information-processing perspective endeavors “to understand the processes by which intellectual products are formed by examining the processes, representations, and strategies individuals use to perform specific intellectual tasks” (p. 131). Hence, this perspective provides methods to tap into the internal processes of cognition so specific cognitive models can be developed (Yang & Embretson, 2007). Newell and Simon (1972) highlight that an analogy between information processing in computers and information processing in people has been used for understanding cognition. Even though the computer metaphor seems very simplistic to represent human cognition, this metaphor continues to have a positive impact on the development of cognitive psychology (Solso et al., 2005). This positive impact continues to occur because models of human cognition bear some similarities to the sequence of events involved in computer processing: they may be regarded as carrying out a task in a series of programmed steps, in which each step in the sequence changes its immediate predecessor (Casey & Moran, 1989). Thus, cognitive processes are assumed to occur as a “sequence of successively transformed states” (Hayes & Broadbent, 1988, p. 271).

According to Eysenck and Keane (2000), by using computational models—supported by experimental evidence—it is possible to have a sense about how the mind functions and about what takes place between observable input and output. Palmer and Kimchi (1986) explain this process as follows:

> Certain information from the environment (the "input") is available to the mind through sensory systems, much as input information is available to a computer program through peripheral devices such as terminal, card readers, and the like. Some of this information is then manipulated in more or less complex ways by mental operations, much as a computer program manipulates information according to the rules it embodies. Among these mental operations are ones that select, transform, store, and match information arising from the present situation, from memories of past situations, from plans for future situations, or (usually) some combination of these. As a result of such operations, the mind produces information in a different form (the "output") that is expressed as overt behavior, in much the same way that a computer program outputs information through the peripheral output devices (pp. 38-39).

Limitations of the information-processing perspective include (a) neglecting affect and conation[5], (b) lacking understanding about the contextual specificity of abilities, and (c) inherent limitations of the short-term memory. Lohman (2000) claims that the first limitation has an effect on the information-

[5] Definitions of conation include: 1) Aspects of mental processes or behavior directed toward action or change and including impulse, desire, volition and striving (American Heritage Dictionary of the English Language, 1971); and 2) …an aspect of man's psychic life having to dowith striving and will, traditionally distinguished from cognition and affection (Good, 1973, Dictionary of Education).

processing perspective as affect can constrict cognition (when represented by anxiety and frustration) or enhance or direct cognition (when represented by interest and surprise). Conation can affect cognition as people who adopt a constructive, motivational orientation will tend to exhibit better self-regulation when executing a task. The second limitation refers to the fact that cognitive abilities are situated, as Lohman (2000)—referring to Snow (1994)—claims that "abilities are reflected in the tuning of particular persons to the particular demands and opportunities of situations and thus reside in the union of person in the situation, not 'in the mind' alone" (p. 329). Twenty years earlier, Norman (1981) had already criticized the information-processing perspective for the neglect of social, cultural, motivational, and emotional factors in cognition. The third limitation refers to the number of units that can be processed at one time. Miller (1956) gave the number 7 ± 2 for the number of items one can hold in the short-term memory. More recent research suggests different spans of numbers, such as, three to five items as the population average (Cowan, 2001) and clusters of no more than three or four items (Halford, Wilson, & Phillips, 1998).

Despite these limitations, Palmer and Kimchi (1986) still believe that information-processing is a viable theoretical approach to cognition. This is because the underlying theory has to be clearly stated—including explicit information about its assumptions and necessary steps—in order to be implemented computationally. Another strength of the approach stems from the fact that the fundamental assumptions of information-processing theories are that "the human mind is complex and that changes in mental functioning occur through some combination of improvements in basic capacities, strategies, and content knowledge" (Chen & Siegler, 2000, p. 96, citing Klahr, 1992). These characteristics have contributed to the information-processing potential to identify strengths and weaknesses of students as they build an understanding of problems and search for solutions to problems (Zoanetti, 2010). Therefore, this approach is also deemed important for developing cognitive models for cognitive diagnostic assessments.

Characteristics of Cognitive Models for CDA

Cognitive models possess at least four defining characteristics when used in CDA (Gierl, Roberts, Alves, & Gotzmann, 2009). 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). This specific information can be generated because the grain size of these models is fine (as opposed to coarse, which occurs when a single test score is used), thereby increasing the depth to which both knowledge and skills are measured with the test items. Gierl, Wang, and Zhou (2008) claimed that the attribute grain size is a "constant concern when developing a diagnostic test because the attribute must characterize the knowledge and skills used by all examinees as they solve items" (p. 6). If the attribute grain size is too coarse, then the score reports will allow only broad and, potentially, uninformative inferences about the examinees' cognitive skills, they affirm. The specificity of the cognitive inference desired to be reported should help one to choose the appropriate attribute grain size. Depending on the assessment characteristics, a finer grain size cognitive model may be used. The frequency and amount of time allocated for assessing students, and the amount of time and money assigned for the development of cognitive models and test items are examples of characteristics that may affect the depth (or the attribute grain size) of the assessment. For example, a finer grain size model may be used if an assessment does not impose constraints about the frequency that students can be assessed (where frequently means having their classes interrupted or requesting them to answer the test during their free time), and there are an adequate number of content specialists to develop a very detailed cognitive model and the associated test items needed to measure the skills in the model.

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, Leighton, Wang, Zhou, Gokiert, & Tan, 2009; see also Anderson, 1996; Dawson, 1998; Kuhn, 2001; Mislevy, Steinberg, & Almond, 2003). In the cognitive model development process, the items that measure each attribute must maintain the cognitive structure outlined in the hierarchy and must directly measure specific cognitive processes of increasing complexity (Gierl, Wang, & Zhou, 2008). As explained by Gierl et al. (2008), if test performance is to be linked to information about examinees' cognitive skills, then items in a cognitive diagnostic assessment should be designed systematically using this hierarchical order. Although this is a controversial topic, as many of the current CDA applications do not entail a hierarchical cognitive model, the benefits of using a structured fashion to organize the attributes are numerous. Ordering the attributes is a desirable characteristic because it allows a more structured way to understand examinees' cognitive skills. It also helps content specialists to comprehend the interconnections between items, which facilitates the item development process. This is significant for test development because the items that measure the attributes must maintain the cognitive structure outlined in the hierarchy and must directly measure specific cognitive processes of increasing complexity. In other words, the items in a cognitive diagnostic assessment could be designed systematically using this hierarchical order, facilitating the link between test performance and examinees' cognitive skills.

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. For example, to "employ mathematical reasoning throughout a complex problem to reach a solution" is not a measurable skill as worded because the terms "mathematical reasoning" and "complex problem" are vague expressions and do not provide enough information for accurate item development. Supposing that three content specialists were asked to develop an item for this skill, it is possible that three completely different tasks would be designed. That is, the skill is not stated clearly, it does not entail an observable outcome, and it is difficult to operationalize using a test item. Usually, a measurable skill contains an action verb that requires the examinee to demonstrate a knowledge state or enact a specific problem-solving skill while answering a test item. In short, the measurability of the skill is an important feature that facilitates the link between the cognitive model, test design, and students' performance. Thus, in order for a test to accurately *measure* the application of the knowledge, skills, and abilities possessed by the examinee, it is vital that the skills are well specified, allowing students to demonstrate their skills mastery or domain using some type of permissible test item.

The fourth characteristic is the instructional relevancy of the skill. The structure of an assessment must be developed around the concept that the outcomes of a diagnostic assessment are both instructionally relevant and meaningful to students, their parents, teachers, and other school officials. Because outcomes from a diagnostic assessment can provide information that might be used for planning classes, modeling instruction, measuring students' progress, and offering feedback to students and parents, it is important that the measured skill be consistent with instructional objectives and learning outcomes from the educational system. The extent to which a diagnostic assessment measures relevant skills reflects how meaningful the assessment results will be to the stakeholders. For example, instructionally relevant outcomes—based on, for example, age-appropriate tasks and real-world situations—may be used by teachers to form a picture of student performance across skills, to identify struggling students, to better isolate and analyze their deficiencies, and to develop interventions with a higher likelihood for success. Hence, by measuring instructionally relevant skills, student performance can be easier linked to instructional actions.

Conceptual Frameworks for Developing Cognitive Models

Different conceptual frameworks have been used to develop cognitive models. In this dissertation, three frameworks–theoretical, content specialist, and combined–are described and discussed.

The first framework, *theoretical,* guides research activities by its reliance on a formal theory developed using an established, coherent explanation of certain types of phenomena and relationships (Eisenhart, 1991). The social psychologist, Kurt Lewin, once noted that "there is nothing as practical as a good theory" (1952, p. 169). Ideally, a theory of task performance would direct the development of the cognitive model as theories provide a valuable method for developing cognitive models as they are based on knowledge and skills used by students as they respond to test items (Gierl, Leighton, Wang, et al., 2009). Although there are few theories readily available in the educational or psychological literature, the role of cognitive psychology theory cannot be underestimated, as it may provide unique information on how to conceptualize complex problem-solving skills (Leighton, 2008).

An example of CDA research guided by a theoretical framework can be found in Briggs and Alonzo (2009). They developed a diagnostic assessment based on the science content domain of The Earth in the Solar System (ESS). The cognitive model, called *learning progression* by the authors, was developed based on national science education standards (American Association for the

Advancement of Science [AAAS], 1993; NRC, 1996) and research literature on students' understanding of the targeted concepts [see Briggs and Alonzo (2009) for a complete list of the literature]. According to Briggs and Alonzo, each learning progression outlines one possible pathway that students might take in moving from their initial ideas to fully understanding the construct. In this investigation, the attribute hierarchy method (Leighton, et al., 2004) was applied to Ordered Multiple-Choice (OMC) item responses to classify a student at a distinct location of a learning progression. The authors concluded that the reviewed literature provided important information about the prevalence of particular ideas at different ages; however, little documentation of students actually progressing through these ideas was found in this literature material.

Although the theoretical framework can be used to help develop cognitive models for CDAs, the critical limitation is that there is a scarcity of theories applicable to cognitive model development for educational testing in many domains. That is, few models are available because cognitive theories specifying the underlying processes, strategies, and knowledge structures that are required to solve tasks on educational tests are simply not available. Another common problem associated with the use of a theoretical framework is the lack of solid evidence to support scholars' claims and their tendency to address and explain research problems by *theoretical decree* (Einsenhart, 1991). This decree creates a disconnection between the conclusions produced by the theoretical description and the usefulness of the descriptions in addressing problems in day-to-day practice (Lester, 2010; Einsenhart, 1991). Ferrara and DeMauro (2006) highlighted this problem, claiming that empirical outcomes from the psychological sciences have, so far, provided little guidance in the development of most educational tests. Reasons for this disconnection may include the use of technical language, different grade levels or content areas, coarse research outcomes that are uninformative for the development of cognitive models, outcomes that are too specific thereby limiting construct representation, and limited content coverage as measured by a diagnostic assessment.

The second framework, *content specialist*, is the process where accumulated knowledge and experience from experts serve as a basis for cognitive model development. Terms that have been used to refer to the content specialist method include: *expert judgment* (Leighton, Heffernan, Cor, Gokiert, & Cui, 2008), *subject-matter expertise* (Mislevy, Behrens, Bennett, et al., 2010), *expert practitioner* (Shute, Graf, & Hansen, 2006), *practical* (Eisenhart, 1991), and *expert elicitation* (Knol, Slottje, Sluijs, & Lebret, 2010). These terms refer to a structured approach where subject-matter experts organize the knowledge in an explicit way (Knol et al., 2010). The content specialist method synthesizes available knowledge, when conclusive scientific evidence is still not available. The importance of these professionals in the educational research field is also described by Leighton, Cui, and Cor (2009), who stated that “content experts are often in a good position to anticipate the knowledge and skills students use to respond correctly to test items since many have worked as teachers and therefore have insights into student thinking and performance” (p. 234). Being directly involved with teaching and learning helps experts to identify the knowledge and skills that students may use to solve items. Also, the accuracy of the results is increased by aggregating the opinions of multiple experts compared to using the opinion of a single expert (NRC, 2009).

An example of research conducted using the content specialist method is provided by Gierl, Wang, and Zhou (2008). The purpose of this study was to apply the attribute hierarchy method to a subset of SAT (Scholastic Aptitude Test) algebra items to promote cognitive diagnostic inferences about examinees. They evaluated examinees' cognitive skills in algebra using a cognitive model developed by content specialists. Their cognitive model was developed by asking content specialists to review the SAT algebra item, identify their salient attributes, and order the item-based attributes into a hierarchy. Gierl et al. (2008) concluded that examinees’ response patterns provided a good approximation to the skills in the cognitive model identified by the content specialists.

Content specialists are used extensively in the design and development of educational assessments, from the initial stage of cognitive model development to the final stage of item validation (e.g., Gierl, Roberts, et al., 2009; Jang, 2008). However, relying solely on expert judgments may have some disadvantages. Because knowledge and skills are not directly observable, content specialists may provide only a superficial understanding of the knowledge and skills examinees’ use to solve items on the test. Also, as pointed out by Leighton et al. (2009), expert and novice problem solvers differ in many ways. Experts not only possess more knowledge and skills than novices but these knowledge and skills but also organize this knowledge differently, usually more efficiently (Leighton et al., 2009). Besides commanding more facts and concepts than novices, experts generally make richer interconnections among them, and view, represent, and approach problems differently than novices (Mislevy, 1994). Therefore, by having experts develop cognitive models for students, it is possible that the generated models (and expected response processes) do not generalize to all ability levels in the student sample. As content specialists have no direct knowledge about the cognitive path invoked by the item (Schmeiser & Welch, 2006), cognitive models developed by these specialists may or may not be an accurate representation and description of the knowledge and skills that examinees actually use when solving items. Another limitation of this method is that of unintentional bias (NRC, 2009). Unintentional bias means that the source of the bias is not purposeful, but rather unconscious. For example, bias can occur when the experts place too much importance on one aspect of an event, do (or believe) things because other experts do (or believe) the same, and interpret information in a way that confirms one's preconceptions. These limitations are accentuated when a single content specialist or small number of them are used for the cognitive model development.

The third framework, *combined,* is also based on previous research similar to the theoretical method, but this method can encompass an array of current and possibly far-ranging resources (Eisenhart, 1991). It may combine outcomes from theories, expert knowledge, empirical research, or other sources relevant to a research problem. Eisenhart (1991) claims that the researcher is less likely to explain empirical evidence using *decree, convention,* or *accident* using the combined framework because it helps to triangulate theory, expertise, and empirical investigation. The combined method has also been used by educational measurement researchers to develop cognitive models. In fact, there are similarities between this method and what Nichols (1994) calls “substantive” research. Substantive research is based not only on assumptions, but also on research reviews and accumulation of scientific evidence. Nichols (1994) claims that two elements form the foundation of substantive research in education measurement research. The first element is a description of the cognitive mechanisms a student uses in a testing situation. The second element is the specification of the item characteristics that are hypothesized to influence the cognitive mechanisms used by the students when answering a test. By recognizing the different processes and knowledge structures that a student brings to a test situation, the test developer may construct items that require these structures (Nichols, 1994).

An example of the combined method is provided by Ketterlin-Geller, Jung, Geller, and Yovanoff (2008). They evaluated a cognitive diagnostic test that measured division of fractions. The cognitive attributes were identified using task analysis, mathematics textbooks review, expert review, and verbal protocols. During the task analysis of the skills and knowledge needed to divide fractions, the fraction problems were classified into three categories: a proper fraction is divided by a proper fraction, a fraction is divided by a whole number, and problems involving dividing mixed numbers. Mathematics textbooks were reviewed to determine the precise steps involved in dividing fractions. According to the authors, the attributes were defined by examining the mathematical rationale and isolating the specific steps needed for students to understand and execute the problem, as outlined in the texts. Expert review was used to refine the attribute list by removing irrelevant attributes or adding necessary skills that were missed during the previous steps. Verbal protocols with students were conducted to collect additional evidence about the cognitive model. The hierarchical relationship among the attributes was also investigated by the research team by carefully analyzing the sequence of skills and knowledge needed to divide fractions. Members of the research team and a mathematics content specialist then wrote 252 items. The data were analyzed using a one-parameter IRT model. Of the 252 items, 229 adequately fit the model, as evaluated using the mean square residual fit statistics. The authors concluded that the cognitive model and the test items were appropriate for the purposes of the project. Hence, the combination of task analysis, mathematics textbook review, expert review, and verbal protocols was useful for the cognitive model development.

The combined method plays an important role in the development of cognitive models. Nonetheless, as with the theoretical and the content specialist methods, the combined method also has limitations. In addition to the limitations discussed for the theoretical and content specialist methods, the implementation of the combined method is laborious and costly, as it often requires the development of cognitive models based on evidence of examinee response processes. Gierl, Leighton, Wang, et al. (2009) claimed that cognitive models are “expensive to develop initially and to refine over time because they require extensive—typically experimental—studies of problem solving on specific tasks; and they require cognitive measurement expertise, which is uncommon" (p. 2). For example, research that focuses on the metacognitive techniques can be very time consuming as it often involves "recruiting and interviewing students, audiotaping or videotaping their responses, and then transcribing, categorizing, coding, and finally interpreting the contents of the reports so as to design plausible cognitive models of task or test performance" (Leighton et al., 2009, p. 230; see also Leighton, 2004). These constraints together with the paucity of information currently available on the knowledge, processes, and strategies that characterize student performance in many domains accentuates the challenges inherent to developing cognitive models using the combined framework.

In the previous section, three frameworks for cognitive model development in educational measurement were discussed. Each framework has strengths, weaknesses, and trade-offs. Leighton and Gierl (2007a) suggested that in an ideal world, different approaches would be "blended" to reap the benefits of one another. However, as added by these authors, blending approaches are not always possible in practice. In most practical situations, one must consider the desirable breadth of the assessment on the one hand and the depth of the model on the other hand (Leighton & Gierl, 2007a), being aware that these two are not always compatible. Other factors to consider when choosing a framework concern the availability of human expertise as well as time and financial resources for research and assessment implementation.

Steps for Implementing a Cognitive Diagnostic Assessment

Next, I outline an organizational structure for delineating the major components required to implement a cognitive diagnostic assessment. This structure, which can be described as a principled test design approach, is based on a four-step procedure: (1) attributes are first identified to characterize the cognitive model underlying examinees' problem-solving skills, (2) test items are constructed according to the cognitive model, (3) a confirmatory psychometric procedure is used to analyze the data, and (4) a detailed score report is provided. The four steps help gather evidence systematically to support the assessments purposes, i.e., to be cognitive and diagnostic, providing useful information about students' strengths and weaknesses. Gierl (2007) states:

> [A] principled approach to test design and analysis is required where the cognitive model is first identified and evaluated, then the test items are developed to measure the attributes in the model, and finally model-based statistics are used to analyze the data, generate the scores, and guide the interpretations of examinees' performance (p. 337).

The principle approach proposed by Gierl draws on the evidence-centered design (ECD) framework introduced by Mislevy, Steinberg, and Almond (2003) and assessment engineering (AE) introduced by Luecht (2007). Briefly, ECD is understood as a set of activities that facilitate explicit thinking about the purpose of an assessment, what is the evidence in student performance required to support the assessment's purpose, and how a test can be developed to offer students an optimal opportunity to provide the observable evidence that they achieved the assessment's purpose (Huff, Steinberg, & Matts, 2009). As a conceptual framework for designing, producing, and delivering educational assessments using evidentiary arguments, ECD helps to ensure that evidence supports the underlying knowledge the assessment is intended to measure (Mislevy, Steinberg, & Almond, 2002). This standpoint is valuable for principled test design because it clarifies the conceptualization of assessment as a structured, coherent, and purposeful process that should lead to more valid inferences about student performance on exams. AE also represents a relatively new area of research in educational measurement that provides an integrated framework for assessment design, item writing, test assembly, and scoring (Luecht, 2008).

Masters (2010) claimed that AE consists of defining the progression of ordered claims about proficiencies and skills, documenting of the universe of observable actions, responses, and/or products that would qualify as evidence for a particular proficiency claim, constructing task models, designing templates, writing items, and calibrating and scoring data. By emphasizing the importance of content expertise and technology, AE is a valuable tool in the principled test development approach adopted in this dissertation. The role of content experts is critical for the creative task of designing and developing meaningful models (Gierl, Zhou, & Alves, 2008). At the same time, the role of technology is emphasized in tasks such as executing the algorithms necessary to combine a large number of elements and allow the operationalization of many procedures such as automatic item generation, test assembly, and psychometric analyses.

Step 1: Cognitive Model Development

The first step in principled test design requires the identification of the attributes that students are expected to master in a specific domain. When the attributes are structured, they form a hierarchy that approximates a cognitive model, which is used to guide item development and to provide the interpretative framework needed for linking test score interpretations to cognitive attributes. Accordingly to Ketterlin-Geller et al. (2008), “the cognitive model is composed of attributes that are domain specific prerequisite skills and knowledge needed to demonstrate mastery in the targeted task” (p. 4). Attributes can be identified by studying the knowledge, processes, and strategies used by examinees that are believed to underlie conceptual understanding in a particular domain. Inquiry methods for analyzing student thinking processes such as expert review, task analyses, and verbal protocols are useful tools for the identification and/or validation of the attributes required to develop a cognitive model (Ketterlin-Geller et al., 2008; Gorin, 2006). For example, when using the judgement from content specialists for developing cognitive models for diagnostic assessments, these specialists draw on their experience in the content domain to anticipate the relevant knowledge and skills with which to describe a construct (Leighton et al., 2009). They may also review textbooks, academic journals, and curricular documents to inform the model development process. To facilitate the explanation and prediction of students’ performance, it is desirable to consider some aspects during the development of a cognitive model. Namely, the attributes should be written at a fine grain size, ordered by complexity, measurable, and instructionally relevant.

Step 2: Item Development

The second step consists of using the cognitive model generated in the first step as a guide for developing test items. Items should be developed from empirically-based cognitive models of learning to support diagnostic inferences about examinees' thinking processes (Leighton & Gierl, 2007b; Nichols, 1994). Item development is often based on an iterative process of generation-revision-modification of the items in which teachers and/or content specialists participate. An example of item development for a diagnostic assessment using a principled test design approach is presented in Briggs, Alonzo, Schwab, and Wilson (2006), where ordered multiple-choice items (OMC) were written according to cognitive models (which they call *construct maps*). OMC items combine the efficiency of traditional multiple-choice items with the qualitative richness of the responses to open-ended questions. The items were developed based on the underlying construct map−a central element to the design and interpretation of the OMC items. As described by the authors, each item response option is linked to a specific developmental level of student understanding. For example, the distractors were written to represent different levels of the construct map (based on the description of understandings and common errors expected of a student at a given level). Responses obtained from open-ended version of the items were also examined. Common or expected students' responses to those items were incorporated into the development of the distracters. The items were revised extensively by their research team, regional directors of a statewide science reform program, and developers of an elementary school science curriculum. A pilot test was conducted with the items. Results from this pilot test revealed that the estimated reliability (Spearman-Brown) for these diagnostic items was comparable to traditional multiple-choice items. The study also presents some preliminary evidence suggesting that the test scores based on ordered items have the potential for greater diagnostic value than test scores based on traditional items. In short, this study presents a systematic way of linking item development and cognitive modeling. The authors successfully demonstrated that by using design principles it is possible to establish a connection between students' item response and the developmental progression of student understanding, embodied in the cognitive model. In other words, using attributes as a foundation for item development has the potential to facilitate a direct connection between cognitive theory and assessment practice. Evidence for response processes can be gathered empirically through pilot and field tests, which constitute an essential aspect of the item development process (Briggs et al., 2006). Chapter 3 of this dissertation describes another example of item development for diagnostic assessments using a principled test design.

Step 3: Confirmatory Psychometric Analysis

The third step consists of using a confirmatory psychometric analysis to evaluate the cognitive model. The term confirmatory indicates a manner to assess the plausibility of the model-data fit relative to the intended underlying cognitive model the assessment is designed to measure (Gierl, Zhou, & Alves, 2008). For example, confirmatory analysis may be applied to verify a cognitive model developed by expert analysis (which functions like an *a priori* model) against the data collected from students test responses. By using a confirmatory procedure to interpret test performance, the test developer gains control over the scores and the inferences about processes and skills associated with test performance (Gierl, Leighton, Wang, et al., 2009). The number of diagnostic psychometric methods created to analyze cognitive data structures has dramatically increased in recent years (Gorin, 2006). Interpretation of assessment data is facilitated by using a powerful psychometric procedure, such as the attribute hierarchy method. These procedures are also necessary to verify whether a given student possesses the attributes required to complete a task, such as solving a problem in a certain domain. Rather than assigning examinees a general mark or letter grade, some diagnostic psychometric tools provide examinees not only with information concerning whether or not they have mastered each of a group of specific skills, but also with individual skill probabilities. The choice of the psychometric procedure affects how reliable and useful the outcomes will be for making inferences about students' strengths and weaknesses. Complexities of cognitive psychological theories (represented by attribute hierarchies, for example) require sophisticated psychometric procedures that can handle complex cognitive structures and the dependencies among skills, and yield specific cognitive inference. Cognitive diagnostic models are deemed to hold great promise for enhancing the quality of diagnostic feedback provided to students because they contain complex structures including attribute dependencies and they can guide specific cognitive inferences (Huebner, 2010).

Step 4: Score Reporting

The forth step specifies how diagnostic scores can be transformed into valuable score reports for test users. Score reports must present information that is useful to the students, teachers, parents, and relevant school administrators. Constructing meaningful reports is not an easy task, however. The difficult role of translating scores in informative reports is highlighted when we consider what scores represent, or should represent. According to Snow and Lohman (1989) "...a score reflects a complex combination of processing skills, strategies, and knowledge components, both procedural and declarative and both controlled and automatic, some of which are variant and some invariant across persons, or tasks, or stages of practice" (p. 268). To construct score reports that better reflect the construct being measured, researchers and practitioners need to understand not only the relations between cognition and task performance, but also the features that affect the usefulness of a score report. In a recent paper published by Roberts and Gierl (2010), they claim that establishing a structured approach to the test score reporting process is an important way to ensure that relevant features are identified and presented in the report. These authors conducted an extensive review of current test score reporting practices in education. They also presented a sample diagnostic score report illustrating a framework that can be applied to CDA using the attribute hierarchy method. This sample report was designed in three sections with related but different functional purposes:

> The top section of the report contains orienting information in the form of an overview of contents for the reader. Student identification information and a summary score is brought to the attention of the reader by placing it in a colored, boxed area in the top-left hand corner of the page. [...] The middle section of the report, “Review Your Answers” contains diagnostic information regarding attribute mastery along with item-level performance. [...] The bottom section of the report is structurally and visually separated from the middle section by the use of a box. This section contains mostly text-based information using bullets with left alignment for clarity in presentation and ease of reading (Roberts & Gierl, 2010, p. 33).

The back page of the report provides a description of the skill category as defined by the cognitive model, an explanation of how diagnostic profiles are produced based on a student’s response pattern, and contextual information for interpreting the contents on the front page of the report (Roberts & Gierl, 2010). The goal of score reporting, as defended by Roberts and Gierl, should be focused towards the *clarity of communication* between the test developers and test users. Unfortunately, few studies have focused on implementing these guidelines on cognitive diagnostic reports (Roberts & Gierl, 2010).

CDAs in the Context of Mathematics

Many systematic reviews have been conducted in an attempt to determine “how, where, and why people learn or do not learn mathematics” (Begle & Gibb, 1980, p. 8) and, ultimately, to better understand how to offer high-quality early mathematics instruction to students. Siegler (2003) claims that analyzing mathematical procedures and concepts is crucial to promote effective learning, provide students with instruction and examples that help them learn the component skills of the task, help teachers to anticipate types of misunderstandings that most often arise in the learning process, and prepare teachers with the means for helping students move beyond these misunderstandings.

In order to gather information about mathematical procedures and concepts, an increasing volume of research on cognitive diagnostic assessments, especially in the domain of mathematics, is accumulating. Examples of areas in mathematics that have been the focus of research include pre-algebra patterns (Ye, 2005), algebra (Gierl, Cui, & Zhou, 2009; Gierl, Wang, & Zhou, 2008; Russell, O'Dwyer, & Miranda, 2009), mixed-number subtraction (Henson, Templin, & Willse, 2009; Sinharay & Almond, 2007; Tatsuoka, 1990), proportional reasoning (Baxter & Junker, 2001; Béland & Mislevy, 1996), fractions (de la Torre & Douglas, 2004, 2008), and multiplication and division with exponents (Birenbaum & Tatsuoka, 1993). The application of cognitive diagnostic models to *mathematics*, although not abundant, is more prevalent compared to other domain areas such as scientific reasoning and reading comprehension.

Table 1 presents a list of CDA studies in the domain of mathematics that were reviewed for this dissertation. The table contains six columns. The first column presents the authors and the publication year of the manuscript (the manuscripts are listed from most to least recently published). The second column presents the specific mathematics subject area being studied. Columns three to six present information about the four necessary steps for principled test design in CDA (i.e., the presence of: cognitive model, item development, confirmatory analysis, and detailed score report). Check marks in each these four columns indicate that the step was conducted in the study.

Table 1. *Summary of the reviewed cognitive diagnostic studies in Mathematics.*

	Authors	Subject	Cognitive Model (CM)			Item development	Confirmatory Analysis	Detailed score report
			Presents CM	Describes CM development	Non-retrofit CM			
1	Daniel & Embretson (2010)	Problem-solving in mathematics	√	√	√	√	√	
2	DeCarlo (2010)	Fraction Subtraction	√				√	
3	Gierl, Alves, & Taylor-Majeau (2010)	Problem-solving in mathematics	√	√	√	√	√	√
4	Roberts & Gierl (2010)	Algebra and Functions	√					√
5	de la Torre & Song (2009)	Math, Math Computation, Spelling, Social Studies					√	
6	Gierl, Roberts, Alves, & Gotzmann (2009)	Mathematics	√	√			√	
7	Gierl, Cui, & Zhou (2009)	Algebra	√	√			√	
8	Gierl, Leighton, Wang, Zhou, Gokiert, & Tan (2009)	Algebra	√	√			√	
9	Gotzmann, Roberts, Alves, & Gierl (2009)	Mathematics	√	√	√		√	
10	Henson, Templin, & Willse (2009)	Mixed-Number Subtraction	√				√	
11	Kunina-Habenicht, Rupp, & Wilhelm (2009)	Arithmetic and modeling skills	√	√	√	√	√	
12	Leighton, Cui, & Cor (2009)	Algebra	√	√	√		√	
13	Roberts, Alves, Gotzmann, Gierl (2009)	Mathematics	√	√	√	√		
14	Castro (2008)	Fraction multiplication and division	√	√	√			
15	de la Torre & Douglas (2008)	Fraction subtraction	√				√	
16	Dogan & Tatsuoka (2008)	Mathematics	√	√	√		√	

	Authors	Subject	Cognitive Model (CM)			Item development	Confirmatory Analysis	Detailed score report
			Presents CM	Describes CM development	Non-retrofit CM			
17	Gierl, Wang, & Zhou (2008)	Algebra	√	√			√	√
18	Ketterlin-Geller, Jung, Geller, & Yovanoff (2008)	Division of fractions	√	√	√	√		
19	Dimitrov (2007)	Calculus	√				√	
20	Sinharay & Almond (2007)	Mixed-number subtraction	√	√			√	
21	Sinharay (2006)	Mixed-Number Subtraction	√	√			√	
22	Ye (2005)	Pre-Algebra	√	√	√	√	√	
23	Birenbaum, Tatsuoka, & Yamada (2004)	Measurement, probability, geometry; and algebra	√				√	
24	de la Torre & Douglas (2004)	Fraction subtraction	√				√	
25	Tatsuoka, Corter, & Tatsuoka (2004)	Mathematics	√	√			√	
26	Dimitrov & Raybkov (2003)	Algebra	√				√	
27	Baxter & Junker (2001)	Proportional Reasoning	√	√	√			
28	Tatsuoka & Tatsuoka (1997)	Fraction addition	√	√		√	√	
29	Embretson (1995)	Mathematical problem solving	√	√	√		√	
30	Birenbaum & Tatsuoka (1993)	Multiplication and division with exponents	√	√			√	√
31	Birenbaum, Kelly, & Tatsuoka (1993)	Linear equations with one unknown	√				√	
32	Tatsuoka & Tatsuoka (1992)	Fraction	√	√		√	√	

Thirty-two studies, including journal articles, conference papers, and dissertations, were considered in this section. All studies are related to cognitive diagnosis in mathematics. Specifically, the subject areas covered by the studies are: fractions (25.0%), algebra (21.9%), general mathematics (21.9%), mixed number subtraction (9.4%), problem solving (6.3%), and other domains (15.6%). This review comprises studies from 1992 to 2010. About 84 percent of the studies were conducted from 2001 to 2010. From these, 78 percent were published in the past 5 years.

The first step in principled test design for CDA is divided into three categories in Table 1. The categories include “presents cognitive model”, “describes cognitive model development”, and “non-retrofit cognitive model”. Overall, 31 out 32 studies presented a cognitive model, of some type. Frequently, the attributes in the cognitive model were organized in a Q-matrix form[6]. However, a description about how the cognitive model and/or Q-matrix were developed was missing in some of the studies (31 percent did not describe the cognitive model development). Another frequent limitation in the studies reviewed is that in many cases the cognitive model is developed by having experts code pre-existing test items for attributes, as opposed to using the attributes for its original purpose in the test development process, that is, guiding the development of items. Only 38 percent of the studies contained information about developing new items (i.e., not using pre-existing test items) for the CDA application. This practice of using the existing item for a new purpose on the diagnostic test can be described as a retrofitting approach because the original items are used to generate a cognitive model. The retrofitting approach to cognitive model development is less than optimal because the generated model will

[6] A Q-matrix is often used as representation of the underlying knowledge and skills that are required for answering test items. This matrix presents a pool of possible items that lists all combinations of attributes, assuming that the attributes are independent.

be constrained by the attribute specifications that happen to occur among the existing items. Consequently, this model might not accurately represent students' knowledge and skills required for mastering a certain domain, as the model is not based on cognitive theory, expertise[7], or empirical studies.

While retrofitting is convenient because test items and examinee response data are already available for analysis, this approach is also limited because the model is generated *post hoc*, and only existing items are used to operationalize the attributes. Accordingly to Gierl, Leighton, Wang, and colleagues (2009), an important consequence when using retrofitting approach to cognitive model development is that the fit between the cognitive model and the item-based attributes is tenuous. Gierl, Roberts, Alves, and Gotzmann's study (2009) used content specialists to rate the quality of their item-by-skill alignment. Results from this study indicated that even when cognitive skills align with some of the existing items, the fit was precarious. In 2005's assessment data, for instance, of the 55% of the items that aligned to the skills, only 23% were judged to be a good fit. The tenuous fit between cognitive model and items happens because the items that support the model development were not designed from an explicit cognitive framework or a principled approach to test design.

For the second step, the practice of developing items using a cognitive model as a guide was not prevalent in the studies reviewed. Across the 31 studies, item development was conducted for seven studies only. However, in two of these seven studies (Tatsuoka & Tatsuoka, 1997; Tatsuoka & Tatsuoka, 1992), there is no explicit statement concerning the use of the cognitive model for guiding item development. For example, in Tatsuoka and Tatsuoka (1997), the authors state: "three fraction *diagnostic* tests–a pretest, a posttest, and a retention test–were

[7] In this case, experts are requested to create the a cognitive model based on their experience, curriculum, and/or other material sources, as opposed to simply categorizing pre-existing items into attributes.

given in 1988 to students" [emphasis added] (p. 14). From this vague description of the test (as *diagnostic*), it is not possible to discern if the items are developed from a cognitive model. The 24 remaining studies either did not provide information concerning the item development, did not use a cognitive model as a guideline for item development, or used existing items from prior test administrations. In fact, a frequent practice among researchers is to use items and/or students' response data sets from existing large-scale assessments for CDA applications. For example, the item responses from TIMSS (Trends in International Mathematics and Science Study) were analyzed in Dogan and Tatsuoka (2008), Birenbaum, Tatsuoka, and Yamada (2004), and Tatsuoka, Corter, and Tatsuoka (2004). Gierl, Cui, and Zhou (2009), Gierl, Wang, and Zhou (2009), and Leighton, Cui, and Cor (2009) used existing items from the SAT. Gierl (2007) warns readers that although using existing items is convenient, given their availability, the distribution of the attributes is likely to be uneven across tasks because the test design of the original assessments is not guided by an explicit cognitive model. Hence, precarious inferences for some problem-solving skills in the model can be expected as the fit between the retrofit cognitive model and the existing test data is tenuous. In addition, attempting to retrofit items and data from assessments that were originally developed for unidimensional scaling purposes may result in estimation problems (Kunina-Habenicht, Rupp, & Wilhelm, 2009). Luecht, Gierl, Tan, and Huff (2006) warn readers about the gravity of applying unidimensional data structure to a multidimensional scoring model:

> Inherently unidimensional item and test information cannot be decomposed to produce useful multidimensional score profiles—no matter how well intentioned or which psychometric model is used to extract the information. Our obvious recommendation is not to try to extract something that is not there (p. 6).

In summary, retrofitting approaches are not recommended and their use may affect the validity of the diagnostic inferences, undermining the value of an assessment (Gierl, 2007).

For the third step, the majority of the studies (84%) used sophisticated statistical and psychometric models to analyze the data. In four studies, however, no statistical analysis was used because this step was beyond their objective. One study (Castro, 2008), while presenting an elaborate cognitive model and developing diagnostic items according to this model, presented a less than ideal statistical analysis (t-test for independent samples). Castro compared pretest and posttest scores for two groups, control and experimental. However, an important goal of a cognitive diagnostic tool—estimate students' mastery of knowledge and skills, providing useful diagnostic information about student strengths and weaknesses (Cui & Leighton, 2009)—is not attended with the use of t-tests. Hence, the use of diagnostic psychometric procedures yields finer information about the mastery of each skill and, consequently, more diagnostic information about student's learning. In conclusion, 84 percent of the studies used sophisticated psychometric models to analyze the data. These results suggest that the third step is receiving most of researchers' attention and interest in CDA, possibly at the expense of the other three steps.

The fourth step was implemented in only four of the 32 studies reviewed. This finding confirms what Roberts and Gierl (2010) asserted when stated that the number of studies on score reporting has been limited and that few studies focus on the features of student-level score reporting. Despite the important role of CDA in providing students and teachers with diagnostic score information intended to help remediate students' weaknesses and inform instructional practices, few CDA studies address the issue of reporting. Given that "the success of CDA in accomplishing its goal of providing more formative feedback to educational stakeholders rests, in part, on the test developer's ability to effectively communicate this information through score reports" (Roberts & Gierl, 2010. p. 25), more research is required on reporting strategies.

To conclude, this CDA literature review highlights that few researchers are conducting their studies using the four-step approach in principled test design and analysis, where a cognitive model is first created, then items are developed accordingly to the cognitive model, a confirmatory analysis is conducted, and finally, score reports are produced. Gierl (2007) claimed that conducting these four steps "is not a standard approach to test design and it rarely, if ever, is used in operational testing situations" (p. 337). Yet the importance of a principled test design approach is accentuated by the fact that if the first step is missed, arguably the entire analysis is compromised because the final diagnostic scores may not be interpretable.

In an attempt to contribute to the literature on CDA and to overcome some of the inherent limitations associated with CDA research as previously presented, one of the purposes of my study is to investigate a cognitive diagnostic assessment that used the four-step process specified in principled test design to guide assessment design and implementation. By using this approach for implementing CDAs, I expect to achieve better control over the specific attributes measured by the test, which, in turn, will lead to more specific inferences about the examinees' cognitive skills. Hence, by executing a more structured approach to a) cognitive model development, b) item development, c) confirmatory diagnostic analysis, and d) detailed score reporting it is hoped that the benefits associated with principled test design can be realized and demonstrated in a practical testing context.

CHAPTER III: METHODOLOGY

This chapter outlines the substantive and statistical procedures that were used in this study. A description of the participants, a framework for the data collection, and subsequent statistical analyses is also presented.

An important step in CDA is to create a clear specification of the construct to be measured. In order to provide information about students' cognitive strengths and weaknesses, a test should be developed using a cognitive model that characterizes the hypothesized knowledge structures and processing skills required to solve test items.

Four Steps for a Principled Approach to Test Design and Analysis

Before describing the details of the cognitive model development, a brief summary of the conceptual framework (namely, *theoretical*, *content specialist*, and *combined*) adopted for the Diagnostic Mathematics project will be provided. Each of these frameworks has strengths and weaknesses. Even though researchers and practitioners generally have the best intentions, the choice of a framework is, often, affected by practical constraints (such as money and time availability). Hence, among the three frameworks for cognitive model development, the *content specialist* approach was considered the most appropriate and feasible framework for the Diagnostic Mathematics project. The limited availability of financial resources, time, and human expertise (such as cognitive psychologists specialized in mathematical reasoning and problem solving) were just some of the factors that prevented the development of cognitive models based on learning sciences and the validation of these models with human studies (e.g., collecting evidence of examinee response processes through verbal reports).

Fortunately, important benefits are evident when the content specialist framework is used. For example, the use of judgments from content specialists serves as a popular and commonlyused methodology for identifying the cognitive skills employed by examinees to solve test items, and its continued use has contributed to the consolidation of this framework and its practices (such as how to facilitate group discussion, keep the group focused on the task, and manage the time). Also, the extensive experience with teaching and learning helps content specialists to identify the knowledge and skills that students may use to solve test items. This experience contributes to the development of the cognitive models. Additionally, this framework yields relatively fast and timely results when compared to studies that require verbal analysis of students' responses. Finally, costs associated with this framework are often more affordable than methodologies that require human studies (such as verbal and protocol analysis) or sophisticated technology (such as eye tracking research).

Next, the procedures used to implement the four-steps for principled test design on the Diagnostic Mathematics project are discussed. The specific procedures for developing the cognitive model (Step 1) are discussed first. Next, the process of item development (Step 2) is described, followed by a discussion of the confirmatory analytical approach used to investigate the data collected for the Diagnostic Mathematics project (Step 3). Finally, score reporting (Step 4) procedures used in this study are considered.

Step 1: Cognitive Model Development

One important aspect of developing an assessment is to have a sound conceptual foundation for outlining what the test is designed to measure and how the test scores are intended to be interpreted. A sound conceptual foundation includes a clear specification of the construct measured by an assessment. A cognitive model constitutes the foundation for CDAs. That is, a test should be developed using a model that characterizes the knowledge structures and processing skills required to solve test items, thus, providing information about students' cognitive strengths and weaknesses in a specific domain. Hence, cognitive model development is the first step in the principled approach to test design. This step requires the identification of the attributes that students are expected to master in a specific domain. These attributes, once structured in a cognitive model, are used to guide item development and to provide the interpretative framework needed to link test score interpretations with the cognitive attributes that are believed to underlie test performance. Attributes can be identified by studying the knowledge, processes, and strategies used by examinees as they solve items in a specific problem-solving domain. The cognitive models used in the Diagnostic Mathematics project were based on judgement from content specialists who drew on their own insights and experiences in the content domain to anticipate the relevant knowledge and skills students use to solve items on the test. The procedures implemented by these specialists when developing the cognitive models are described for Step 1.

The development of the cognitive models used in the current study was guided by three mathematics specialists from Alberta Education: one assessment manager and two test examiners in mathematics. The professionals on this team were experienced classroom teachers who had a broad range of practical experiences in large-scale test development. The assessment manager has a B.Ed. (Music, Mathematics), BA. (Music), and P.D.A.D.[8] (Education), as well as 20 years of teaching experience from Kindergarten to Grade 12 (15 years as a Mathematics instructor), and over 10 years of experience in test development at Alberta Education. The first test examiner has a B.Ed. (Secondary Mathematics), 32 years as a Mathematics Teacher from Kindergarten to Grade 12, and five years of experience in test development. The second test examiner has a

[8] P.D.A.D means Professional Diploma After Degree.

B.Ed. (Music), BA. (Music), 15 years as a Mathematics Teacher from Kindergarten to Grade 12, and 2 years of experience in test development at Alberta Education.

The assessment manager developed the preliminary cognitive models for the content area of Numbers at Grade 3. The two test examiners, together with seven content specialists, used this preliminary work as a starting point for their discussions. At this stage of model development, the role of the seven content specialists was to review the preliminary cognitive models and the development of cognitive models for other strands and grade levels. The seven content specialists were all active teachers with a range of backgrounds and experience in teaching (17 years, on average) and test development (5 years of experience as item writer, on average). Following these discussions, the cognitive models developed by the exam manager were scrutinized by this group. The first draft of the cognitive models for the remaining areas and grade levels was developed by the two test examiners from Alberta Education.

Preliminary cognitive models were created using the content, knowledge, and skills specified in the provincial curriculum 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 Specific Outcomes (SOs) outlined by this document were used by the content specialists to create a list of skills and knowledge the students needed to achieve those outcomes. According to this program of studies, specific outcomes are “statements that identify the specific skills, understanding and knowledge that students are required to attain by the end of a given grade" (Alberta Education, 2007, p. 13). To provide teachers with examples of evidence that may be used to determine whether or not students have achieved a specific outcome, a list of achievement indicators is presented in Alberta's program of studies. These indicators can be used as evidence for the desired learning to be achieved and to form a clear picture of the scope of each specific outcome (Alberta Education, 2007).

It is important to note that a specific student achievement perspective was adopted for the development of the cognitive models. The attributes in the cognitive models were developed to represent the skills that would be mastered by moderately high achievers[9]. Hence, this achievement group serves as the reference point for the cognitive model design used in the current study. Students with moderately high ability are capable of applying the mathematical concepts with a relatively high degree of consistency to solve problems. This group of students are confident learners; they create their own connections and apply their previous knowledge and skills to solve novel problems. They demonstrate perseverance when working through challenging problems. Students in this reference group are capable of understanding and grasping the mathematical concepts at their grade level at an achievement level of approximately 70-80%. In a practical sense, moderately high achievers can be thought as students that would probably achieve around 80% in math classroom assessments and around 75% in large scale achievement tests. Moderately high achievers, after receiving instruction, quickly understand new concepts, can work independently, and can transfer the skill from one context to another. It is also expected that these students will answer the majority of a test's items correctly. The assumption behind the choice of developing cognitive models based on moderately high ability students is that low and moderate ability students will, through instruction, maturation, and

[9] This information was provided by two test examiners from the Diagnostic Mathematics project in response to an enquiry about the reference group for whom the cognitive models were designed.

effort, reach the level of performance characteristic of the moderately high achievers. A complete description of the characteristics of the reference group can be found in the third column of the document *Diagnostic Mathematics: Performance Level Descriptors*(see appendix). This document was created by the assessment manager and the two test examiners when asked to characterize three levels of test performance for the Diagnostic Mathematics project.

The development of the necessary skills and knowledge for each specific outcome was directed by the assessment manager. After splitting the teachers into groups—according to grade—teachers were asked to work through the specific outcomes. The skills in each SO were placed in a hierarchical form, increasing in difficulty, according to the expert judgment of the team. According to one test examiner responsible for developing the cognitive models, when writing the attributes, she thought about the step-by-step skills she would teach to students in order for them to understand an outcome. Therefore, skills were ordered according to what she thought should be taught first, second, and so forth.This hierarchy[10] of ordered skills, or attributes, within a domain constituted the cognitive model for the Diagnostic Mathematics project. Using linear hierarchies was considered a good starting point in this project because of their simplicity and easy interpretability. Given the young age of the student participants, the linear format was deemed more appropriate as it yields information that is easier to understand, which increases the usefulness of the score reports. Notwithstanding the fact that the linearity of the models can be supported by the information-processing perspective—as this approach endorses the ordered sequence of stages occurring in the cognitive processes (Mulder, 1983)—I acknowledge that linear models (and cognitive models, in general) constitute a relatively simplistic representation

[10] Each hierarchy here represents a cognitive model. The two terms are used interchangeably in my study.

of a much more complex phenomenon of cognitive information processing. For the current study, this representation serves as the starting point.

The development of cognitive models was accomplished by the team in approximately six days. Throughout the development process, the specialists 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. Guiding the cognitive models development using these four characteristics is one aspect that differentiated the Diagnostic Mathematics project from more traditional approaches to achievement test development.

Once the initial draft of the cognitive models was written, the assessment manager and two test examiners evaluated and revised the skills within each category to clarify the skills. That is, the content specialists ensured that the necessary knowledge and skills were specified in each hierarchy. In addition, they added any important skills that were deemed to be missing in order to fill in the gaps in the models using outcomes specified in the Mathematics Program of Studies as well as by using their own judgments and experiences.

After this initial development work, a 1-day round table meeting with another group of specialists was conducted. This group included eight experienced teachers from Alberta schools. The meeting was led by the assessment manager and the two test examiners. The objective of this meeting was to revise and validate the cognitive models based on teachers' experience and judgment. The teachers were asked to carefully read and evaluate the models. After a 15-minute discussion, they were then asked to answer a short questionnaire evaluating the hierarchies. Results from this questionnaire showed that the layout of the information was easy to follow; the necessary skills and knowledge were included in each hierarchy; the attributes in each hierarchy were ordered from easiest to hardest; and the wording of each attribute was clear.

In total, seven iterations of discussions and revisions were necessary for model development in Grade 3. Thirteen models were completed through this process, and two of these are presented next. These two models were chosen for analysis in my dissertation because these models contained an adequate sample size during student field testing and yielded satisfactory preliminary results (e.g., increasing level of difficulty of the items).

Description of the Cognitive Models. The two outcomes of the Diagnostic Mathematics project are in the content strand *Develop Number Sense*. The first cognitive model is based on *Comparing and ordering numbers* (also known as specific outcome 3 or SO3) and the second cognitive model on *Applying mental strategies for subtracting two 2-digit numerals* (also called specific outcome 7 or SO7). These models are described next in more detail.

Model 1: Comparing and ordering numbers. The first cognitive model measured students' ability to compare and order numbers from 100 to 1000. The specific skills or attributes necessary to master this outcome are presented in Figure 1.

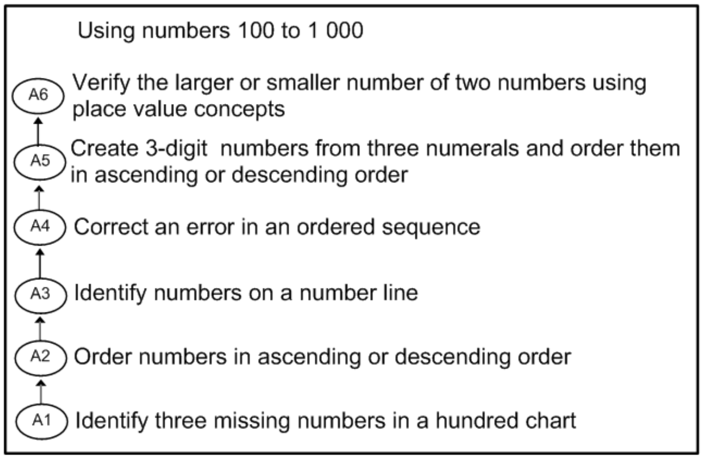

Using numbers 100 to 1 000
A6
Verify the larger or smaller number of two numbers using place value concepts
A5
Create 3-digit numbers from three numerals and order them in ascending or descending order
A4
Correct an error in an ordered sequence
A3
Identify numbers on a number line
A2
Order numbers in ascending or descending order
A1
Identify three missing numbers in a hundred chart

Figure 1. Cognitive attributes required on *Comparing and ordering numbers.*

For *Comparing and ordering numbers*, six attributes were identified following a linear hierarchical form. This means that A1 is the first cognitive skill, and it serves as the prerequisite skill to all other skills in this model. This linear hierarchy also implies that A1 is prerequisite to A2; A1 and A2 are prerequisite to A3; A1, A2, and A3 are prerequisite to A4, etc. As a prerequisite attribute, A1 reflects the most basic skill, "identify three missing numbers in a hundred chart". That is, in order to master the second attribute, "order numbers in ascending or descending order", the examinee needs to be knowledgeable about identifying missing numbers in a hundred chart (i.e., A1), as well as on ordering numbers (i.e., A2). By implication, an examinee is not expected to possess A2 unless A1 has been mastered. This characteristic requires that the attributes must represent different levels of the construct, from the lower level to the higher. In this case, one level of mastering is built upon the other. The right side of Figure 2 depicts this feature.

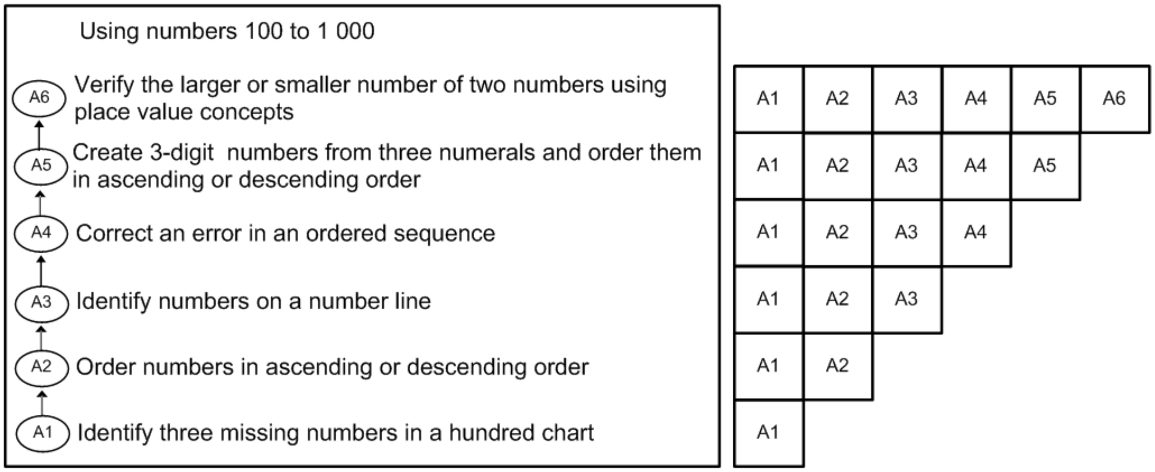

Using numbers 100 to 1 000
A6
Verify the larger or smaller number of two numbers using place value concepts
A5
Create 3-digit numbers from three numerals and order them in ascending or descending order
A4
Correct an error in an ordered sequence
A3
Identify numbers on a number line
A2
Order numbers in ascending or descending order
A1
Identify three missing numbers in a hundred chart
A1 A2 A3 A4 A5 A6
A1 A2 A3 A4 A5
A1 A2 A3 A4
A1 A2 A3
A1 A2
A1

Figure 2. Linear hierarchy for *Comparing and ordering numbers.*

This dependent relationship, as specified in the cognitive model and operationalized by the linear hierarchy, is then implemented through the development of items specifically created to measure the attributes and their dependencies (see the right-hand side of Figure 3).

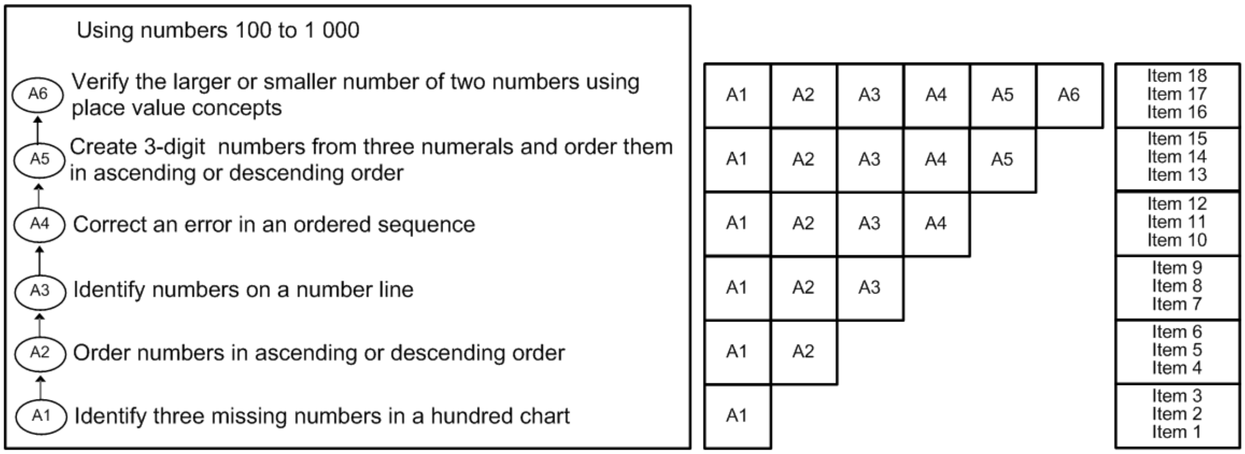

Using numbers 100 to 1 000
A6 Verify the larger or smaller number of two numbers using place value concepts
A5 Create 3-digit numbers from three numerals and order them in ascending or descending order
A4 Correct an error in an ordered sequence
A3 Identify numbers on a number line
A2 Order numbers in ascending or descending order
A1 Identify three missing numbers in a hundred chart
A1 A2 A3 A4 A5 A6
A1 A2 A3 A4 A5
A1 A2 A3 A4
A1 A2 A3
A1 A2
A1
Item 18 Item 17 Item 16
Item 15 Item 14 Item 13
Item 12 Item 11 Item 10
Item 9 Item 8 Item 7
Item 6 Item 5 Item 4
Item 3 Item 2 Item 1

Figure 3. Linear hierarchy implemented by test items for *Comparing and ordering numbers.*

Model 2: Subtracting two 2-digit numerals. The second cognitive model measures students' ability to apply mental mathematics strategies for subtracting two 2-digit numerals, such as taking the subtrahend to the nearest multiple of ten and then compensating, thinking of addition, and using doubles. The specific skills necessary to master this outcome are presented in Figure 4.

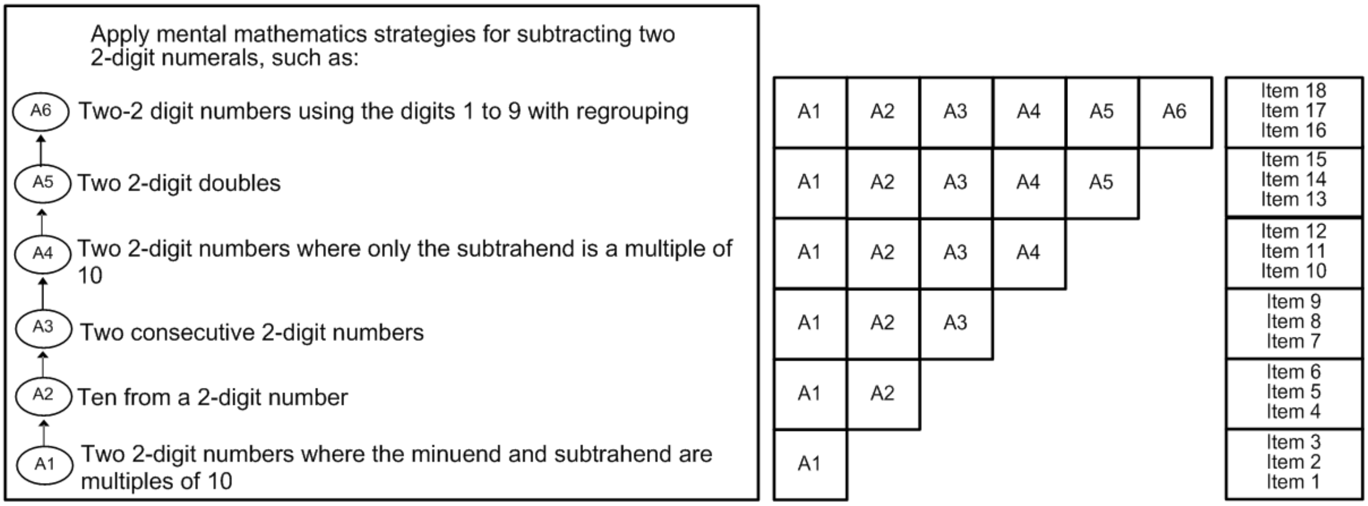

Apply mental mathematics strategies for subtracting two 2-digit numerals, such as:
A6 Two-2 digit numbers using the digits 1 to 9 with regrouping
A5 Two 2-digit doubles
A4 Two 2-digit numbers where only the subtrahend is a multiple of 10
A3 Two consecutive 2-digit numbers
A2 Ten from a 2-digit number
A1 Two 2-digit numbers where the minuend and subtrahend are multiples of 10
A1 A2 A3 A4 A5 A6
A1 A2 A3 A4 A5
A1 A2 A3 A4
A1 A2 A3
A1 A2
A1
Item 18 Item 17 Item 16
Item 15 Item 14 Item 13
Item 12 Item 11 Item 10
Item 9 Item 8 Item 7
Item 6 Item 5 Item 4
Item 3 Item 2 Item 1

Figure 4. Linear hierarchy implemented by test items for *Subtracting two 2-digit numerals.*

For *Subtracting two 2-digit numerals,* six attributes were identified, as presented on the left side of Figure 4. Similarly to *Comparing and ordering numbers*, this cognitive model specifies a linear hierarchical structure, with A1 being the simplest cognitive skill that also serves as a prerequisite to all other attributes. As a prerequisite attribute, A1 reflects the most basic skill: "apply mental mathematics strategies for subtracting two 2-digit numbers where the minuend and subtrahend are multiples of 10". That is, an examinee is not expected to possess A2—"apply mental mathematics strategies for subtracting ten from a 2-digit number"—unless A1 has been mastered. The middle part of Figure 4 depicts the hierarchical feature of the cognitive model, from the lower level to the higher, where one attribute is built upon the other. The right side of the figure depicts the attributes and their dependencies after items are developed to operationalize the cognitive model.

Given that a cognitive model is a vital part of the diagnostic assessment design, adhering to a principled approach to test design and analysis for constructing cognitive models played an important role in the Diagnostic Mathematics project.

Step 2: Item Development

The second step in the principled test design approach consists of using the cognitive model generated in the first step as a guide for developing test items. When items are developed from cognitive models, assessment results may yield more accurate diagnostic inferences about examinees' knowledge and skills.

The process of developing items for the Diagnostic Mathematics project was based on the knowledge and skills specified in the cognitive models created prior to item writing. Employing the cognitive model as the foundation is a requirement to implement a principled test design approach. In this section I describe the specific procedures for developing items for the Diagnostic Mathematics project.

Test development was conducted through the coordinated efforts of qualified item writers. The importance of the item-writers' qualification cannot be underestimated, as it ensures that high-quality test items are produced, and consequently, used in the assessment. According to Haladyna (2004), "the quality of items depends directly on the skill and expertise of the item writers. No amount of editing or the various reviews […] will improve poorly written items" (p. 16). The fact that the item writers for the Diagnostic Mathematics project were certified teachers in the province who had extensive teaching experience in mathematics (17 years, on average) and had experience (5 years, on average) in writing test items for Alberta Education suggests that these professionals were qualified for the task of writing mathematics items for this assessment.

Item writers also received training, including information about the purposes of the Diagnostic Mathematics project. Knowing the purposes of the project was important because item writers gained a clear idea of how the performance on the test would be translated into information that has the potential to identify students' cognitive problem-solving strengths and weaknesses, as well as direct teaching and instructional strategies. Six item writers participated on this 1-day training conducted by the assessment manager. Among these six people, two were permanent members of the Diagnostic Mathematics project (test examiners). The other four members were staff from Alberta Education, but not permanent members of the project. This training provided item writers with information about the project framework, the purpose of the assessment, the cognitive models, and the guidelines for item writing.

In addition, item writers followed a manual (Alberta Education, 2008, 2009) specifically developed to guide the item writing process. This manual comprises a set of established principles, such as items should reflect the cognitive attribute (skill), items should be clear and contain simple language, items should be free of cultural, gender, or other biases. Careful attention to item-writing guidelines is critical for many reasons, including the fact that it helps ensure that items perform properly from content and psychometric item-analysis perspectives, thus, adding to the precision of scores. Overall, the item writing guidelines adopted by Alberta Education were consistent with what is recommended in the item development and test construction literature (see Downing, 2006; Haladyna, 2004; Haladyna& Downing, 1989; Haladyna, Downing, & Rodriguez, 2002; Schmeiser & Welch, 2006).

As part of the item development process, the team was asked to construct items at different difficulty levels for each attribute. In this way, items for the most basic attribute (A1) of a cognitive model should be developed to be very easy and items for the most complex attribute (A6) should be more difficult. This requirement helped ensure the items would meet the ordering feature and, thus, measure the skills in hierarchy from simple to complex. They were also asked to create three parallel items for each cognitive attribute. The reason for creating three items was to enhance the reliability of the measures per attribute. Two types of items were constructed: multiple-choice (MC) and numerical response (NR).

Item development was based on an iterative process of revision-modification. A 6-step process was used to develop the items. First, two test examiners were responsible for writing the assessment items. Items from half of the cognitive models were written by each examiner. The number of items written was equal to the number of skills in the cognitive model times 3, since three parallel items were required for each skill. For example, a model that contains six skills required the development of 18 items. Second, the first examiner revised all items created by the second examiner, and vice-versa. Third, the project manager revised all the items. Fourth, three to five[11] other Alberta Education staff members were invited to revise the items. Specifically, they verified the key, checked if the items in each attribute were structured from easiest to hardest, and ensured the accuracy of wording in each item. Fifth, the items were sent to the graphic designer who created the art for each item, such as pictures, motion, and audio. And sixth, the items were again reviewed in order to check if (a) all the requirements for the graphics and art work were met, (b) the item matched the intended skill, (c) the item matched the expected difficulty level, and (d) the source, stem, and alternatives were correct. For each cognitive model, the first three steps took approximately two weeks, if the examiners worked exclusively with that hierarchy. Overall, the 6-step process required two months.

The first cognitive model evaluated in this study, *Comparing and ordering numbers*, was measured by 18 multiple-choice items with four options each. The second cognitive model, *Subtracting two 2-digit numerals*, was measured by 18 numerical-response items. In both models, all attributes were assessed by three items.

[11] Depending on the SO, a different number of specialists participated.

According to Schmeiser & Welch (2006), item review is a crucial step for the quality of assessments and, although it takes time in the developmental schedule, it should not be ignored. Haladyna (2004) suggests eight interrelated item-review activities that may provide a substantial body of evidence supporting the validity of test score interpretations. Table 2 provides a summary of these activities.

Table 2. *Item review activities implemented during the item development process.*

1. Item-writing review: Checks items against guidelines for violations.
2. Cognitive demand review: Checks item to see if it elicits the cognitive process intended.
3. Content review: Checks for accuracy of content classification.
4. Editorial review: Checks items for clarity and any grammar, spelling, punctuation, or capitalization errors.
5. Sensitivity and fairness review: Checks items for stereotyping of persons or insensitive use of language.
6. Key check: Checks items for accuracy of correct answer. Ensures that there is only one right answer.
7. Answer justification: Listens to test takers' alternative explanations for their choices and gives them credit when justified.
8. Think-aloud: During the field test, this procedure subjects each item to a round-table discussion by test takers.

*Table adapted from Haladyna, 2004, p.201.

The following information was provided by one test examiner from the Diagnostic Mathematics project in response to an enquiry by me about the item-review process used for this assessment. According to her, among the eight steps described in Table 2, Alberta Education implemented steps 1 to 6 on a consistent basis. For the Diagnostic Mathematics project, the first six activities were completed by the three permanent members of the project (the assessment manager, and two test examiners), then by a team of reviewers from Alberta Education. After reviewing the items, this team of reviewers submitted the suggested changes on the graphics, questions, and alternatives to the permanent members of the project. After consideration, changes were made and the items were then sent to the editorial team. Final changes were made and the items were sent to field testing. Step 7 was implemented for one specific outcome (SO 4, from

Grade 3, *Numbers* strand). The two test examiners together with a content specialist visited three different schools and interviewed individual students as they completed the field test. They asked students to explain why they chose a specific alternative and recorded students' answers. Based on their findings, the test examiners made some changes to the items and alternatives prior to field testing them again. Another undertaking also related to Step 7 consisted of sending some of the questions from specific outcomes 6 and 7 (from Grade 3, *Numbers* strand) in written format to a few teachers. These teachers were asked to have their students complete the items (numerical response) on paper and send them back to Alberta Education. The received responses were useful for creating different formats of the questions or different multiple-choice alternatives. Due to time and money constraints, interviewing examinees was not feasible for all specific outcomes. This scenario is not unique to Alberta Education. As pointed out by Haladyna (2004), very few assessments conduct think-aloud methods or report this kind of validity evidence.

In summary, the item-review process conducted by Alberta Education for the Diagnostic Mathematics project together with using cognitive models as a foundation for item development, and having appropriately qualified item writers, who received proper training, and who used a well-established set of principles for item-writing, constitute important actions toward implementing a principled test design approach.

Step 3. Confirmatory Psychometric Analyses

The third step in the principled approach to test design consists of using a confirmatory psychometric procedure to analyse students' item response. In order to verify how well the cognitive models developed by the mathematics specialists function in an operational testing situation, a confirmatory analysis is conducted using data collected from students' test item responses. Hence, in the Diagnostic Mathematics project, the examinee item response data were analysed using the attribute hierarchy method (AHM).

The AHM is a sophisticated psychometric procedure that can handle complex cognitive structures and the dependencies among skills, yielding not only information concerning whether or not examinees have mastered each of a group of specific skills, but also individual skill probabilities. Another strength of the AHM lies in the fact that this method is considered a *confirmatory* approach where the plausibility of the model-data fit relative to the intended underlying cognitive model is assessed. In addition to estimating the attribute probability, the AHM is informed by two other methods for evaluating the cognitive models: the hierarchical consistency index (HCI), and the attribute reliability estimates. Using the HCI, it is possible to investigate the degree to which an observed examinee response pattern is consistent with the specified attribute hierarchy. Attribute reliability assesses the consistency of the decisions made with respect to examinees' mastery of specific attributes. Additionally, the AHM also has a convenient way of providing feedback to examinees, in which score reports map observed examinee item response patterns onto expected examinee item response patterns derived from the cognitive model (Wang, 2007). By mapping examinee's mastery level and providing detailed and precise information about their cognitive problem-solving strengths and weaknesses, this diagnostic information may help teachers and students to design strategies to improve the teaching and learning processes. Another advantage of the AHM approach lies in its facility to guide test development, as test developers can create items according to the hierarchical organization of the attributes (Wang, 2007). As a result, the test developer achieves control over the specific skill measured by the items. In sum, the AHM holds great promise for enhancing the quality of diagnostic feedback provided to students because of its capability of modeling complex cognitive structures including attribute dependencies that can indicate specific cognitive inferences about students' strengths and weaknesses.

The AHM is a consolidated and reliable method for analyzing cognitive assessments and, beginning with its introduction in 2004, many research studies have been conducted using this method. Studies conducted using the AHM have been published in journals, book chapters, technical reports, conference papers, and dissertations. Applications of the AHM include, but are not limited to identifying and interpreting cognitive skills that produce group differences (Gierl, Zheng, & Cui, 2008; Gotzmann, Roberts, Alves, & Gierl, 2009), evaluating the technical aspects of the AHM and its application to the domain of syllogistic reasoning (Leighton, et al., 2004), evaluating the performance of different classification methods (Cui, Leighton, & Zheng, 2006), making diagnostic inferences about examinees' cognitive skills (Gierl, Leighton, & Hunka, 2007), comparing expert-based and student-based cognitive models (Leighton, Cui, & Cor, 2009), evaluating person fit for cognitive diagnostic assessment (Cui, 2007; Cui, Leighton, Gierl, & Hunka, 2006; Cui & Leighton, 2009), investigating learning progression assessments with ordered multiple-choice items (Briggs & Alonzo, 2009), investigating cognitive models of task performance in algebra (Gierl, Leighton, Wang, & Tan, 2005; Gierl, Leighton, Wang, Zhou, et al., 2009; Gierl, Tan, & Wang, 2005; Gierl, Wang, & Zhou, 2008), evaluating a diagnostic assessment using principled test design (Gierl, Alves, & Taylor-Majeau, 2010), investigating the reliability and attribute-based scoring in cognitive diagnostic assessment (Gierl, Cui, & Zhou, 2009), comparing rule-space model and AHM (Gierl, 2007), developing score reports for cognitive diagnostic assessments (Roberts & Gierl, 2009; Roberts & Gierl, 2010), developing and validating reading profiles (VanderVeen, Huff, Gierl, McNamara, Louwerse, & Graesser, 2007), investigating the cognitive processes on critical reading (Wang, 2007; Wang & Gierl, 2007), discussing IRT-based cognitive diagnostic models and related methods (Bolt, 2007), and estimating attribute-based reliability in cognitive diagnostic assessment (Zhou, 2010).

The AHM is a cognitively-based psychometric approach used to assess examinees' performance on a specific domain with the use of a cognitive model of attributes as its foundation. For this method, the cognitive model is essential in its development and application. The cognitive model is operationalized by the attribute hierarchy, which also serves as a framework to guide item development and score interpretation.

The AHM evolved from Tatsuoka's rule-space model (Tatsuoka, 1983) and it is based on the assumption that test performance depends on a set of hierarchically ordered competencies or attributes. The attributes' ordering is based upon their logical and/or psychological properties (Leighton & Gierl, 2007b) and requires specific procedures, skills, and/or processes for an examinee to solve an item. The attribute hierarchy is representative of a cognitive model that allows for the prediction and explanation of students' performance.

In this study, a linear model structure is used to specify the relationships among the attributes. Figure 5 depicts a hypothetical linear model containing six attributes aligned in a single branch.

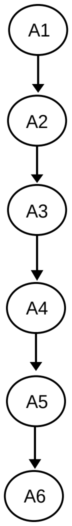


Figure 5. A linear cognitive model containing six attributes.

This model specifies a linear hierarchical form, where A1 is the first prerequisite cognitive skill. From a linear hierarchy perspective, if attribute 1 is not present, then all attributes that follow are not expected to be mastered by the examinee (Leighton et al., 2004).

After defining the attribute hierarchy, four different sequential matrices must be developed using the information in the attribute hierarchy: the adjacency, reachability, incidence, and reduced incidence matrices. These represent the attribute hierarchy in terms of attribute and expected response patterns. The adjacency and reachability matrices represent the relationships among the attributes in the hierarchy in a binary form. The incidence and reduced incidence matrices relate the relationship of the attributes in the hierarchy to the items available in the diagnostic test item bank. For descriptive purposes, the linear hierarchy of Figure 5 is used to illustrate these matrices. A detailed description of each matrix is presented next.

The direct relationship among attributes is specified by a binary adjacency matrix (A) of order (*k*, *k*), where *k* is the number of attributes. In this matrix, the diagonal elements are denoted as 0s while the off-diagonal elements are 1s or 0s depending on the relationship between the two attributes. If attribute j is a direct prerequisite to attribute k, the position (j, k), where $j \neq k$, is filled with a 1; if there is no direct prerequisite between them, a 0 is placed in this off-diagonal. The adjacency matrix for the hypothetical linear example shown above is presented below:

$$\begin{bmatrix} 010000 \\ 001000 \\ 000100 \\ 000010 \\ 000001 \\ 000000 \end{bmatrix} \qquad \text{(Matrix 1)}$$

Because A1 (row 1) is prerequisite to A2 (column 2), the element in position a_{12} is equal to 1. Since A1 is a direct prerequisite to no other attribute, all other elements in this row are filled with 0s. Row 6, which is a row of 0s, indicates that attribute A6 is not a prerequisite to any other attributes.

The reachability matrix (R), of order (k, k), is used to specify both the direct and indirect relationships among attributes. The R matrix can be derived from the adjacency matrix by performing Boolean addition and multiplication. As explained by Tatsuoka (2009), Boolean addition is defined by 1+1=1, 1+0=1, 0+1=1, and 0+0=0. Boolean multiplication is defined by 0*0= 0, 1*0=0, 0*1=0, and 1*1=1. The R matrix is calculated using $R = (A + I)^n$, where n = 1, 2, …, k, is the integer required for R to reach invariance; A is the adjacency matrix; I is the identity matrix; and k is the number of attributes. For the hypothetical example, A + I is equal to

$$\begin{bmatrix} 111111 \\ 011111 \\ 001111 \\ 000111 \\ 000011 \\ 000001 \end{bmatrix} \qquad \text{(Matrix 2)}$$

In this hypothetical reachability matrix, row 1 indicates that attribute A1 has a relationship with A1 and all other attributes through direct or indirect connections as all elements in row 1 are 1s. Row 6 indicates that attribute A6 is a direct prerequisite of A6 and is not a direct or indirect prerequisite of any of the other attributes (i.e., only $r_{6,6}$=1).

The underlying knowledge and skills that are required for answering test items are represented in a Q-matrix, where the rows display attributes and the columns display the items, or vice-versa (Tatsuoka, 2009). This matrix presents a pool of possible items that list all combinations of attributes, assuming that the attributes are independent. Leighton et al. (2004) call this pool a *set of potential items*, in which the total number of items can be calculated by the expression 2^k - 1, where *k* is the number of attributes. Since the Q matrix is a Boolean matrix, all the entries are 1s or 0s, where q_{ki}=1 denotes that attribute *k* is required for answering the item *i* and q_{ki}=0 denotes that attribute *k* is not a requisite for answering a specific item *i*.

$$\begin{bmatrix} 101010101010101010101010101010101010101010101010101010101010101 \\ 011001100110011001100110011001100110011001100110011001100110011 \\ 000111100001111000011110000111100001111000011110000111100001111 \\ 000000011111111000000001111111100000000111111110000000011111111 \\ 000000000000000111111111111111100000000000000001111111111111111 \\ 000000000000000000000000000000011111111111111111111111111111111 \end{bmatrix} \text{(Matrix 3)}$$

There are 63 columns and six rows in Matrix 3, where each column represents one item and each row represents an attribute. The first column of this matrix specifies that only attribute A1 is required to correctly answer item 1. Conversely, in order to correctly answer the last item (column 63), all six attributes must be mastered. Although 63 patterns are possible when attribute and item are matched, as shown in Matrix 3, only seven of these are items that can be used in measuring the cognitive model.

The set of potential items can be reduced, producing the *reduced Q matrix (Q_r).*The second column serves as a good example of an item that must be removed. Item pattern (010000) results from attribute pattern (010000)—a student mastered A2 without mastering A1—which does not fit the attribute hierarchy as the hierarchical structure states that A1 is prerequisite to A2. The Qr matrix is of order (*k*, *i*), where *k* is the number of attributes and *i* is the reduced number of diagnostic items resulting from the constraints in the attribute hierarchy.

$$\begin{bmatrix} 111111 \\ 011111 \\ 001111 \\ 000111 \\ 000011 \\ 000001 \end{bmatrix}. \tag{Matrix 4}$$

The importance of the Qr matrix is made clear by Gierl et al. (2010) when they claim that because this matrix describes all attribute-by-item combinations hierarchy and identifies each item that must be developed to measure the attribute, it is significant for principled test design.

After items are developed and administered, the AHM psychometric procedures can be applied to promote specific diagnostic inferences about the students' mastery of attributes. But, before the examinees' observed response patterns are used to produce the attribute-based scores, it is necessary to evaluate the fit between the examinees' expected response patterns outlined in the hierarchy and the observed pattern produced by examinees responding to the test items. This evaluation of model-data fit involves analyzing student item responses produced from their responses to the test items with the expected item response patterns predicted by the cognitive model.

Fit Analyses. Few model-data fit statistics that have been found in the literature are explicitly designed for cognitive diagnostic assessments (Cui & Leighton, 2009). To address this limitation, Cui and Leighton (2009) developed an index called the Hierarchy Consistency Index (HCI; see also, Cui, 2007; Cui, Leighton, Gierl, & Hunka, 2006). They describe the way the HCI functions as follows:

> For an educational test that is designed to measure a set of hierarchically ordered attributes, students are expected to answer correctly items that measure simple attributes if they have also produced correct answers to items requiring complex attributes. The logic of the person-fit statistic HCI is to examine whether students' actual item response patterns match the expected response patterns based on the hierarchical relationship among attributes measured by test items (p. 433).

The degree to which the observed response patterns are consistent with the attribute hierarchy and the reduced Q matrix is given by:

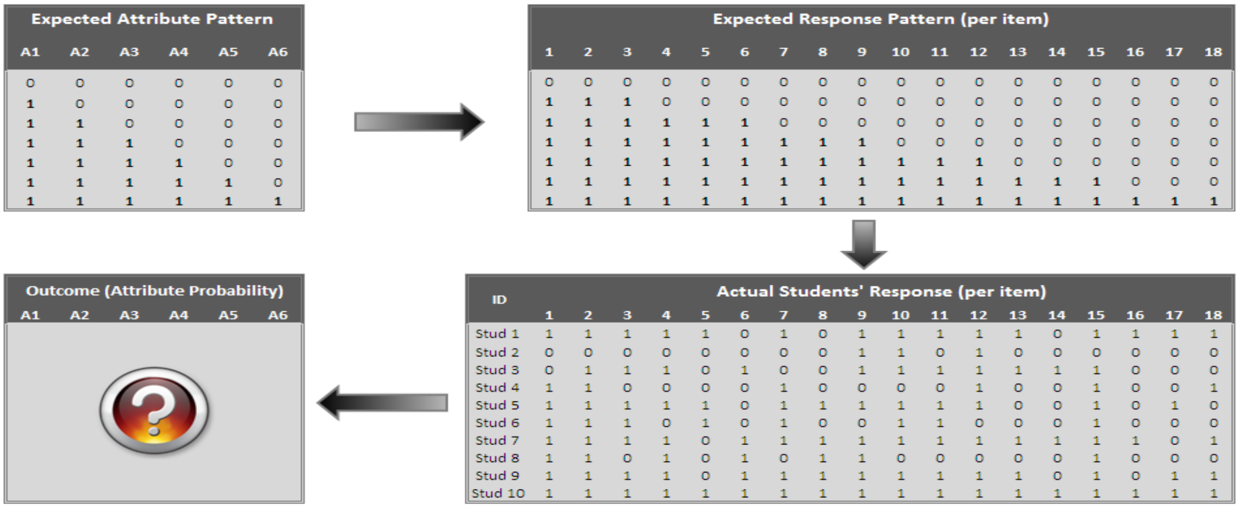
$$HCI_i = 1 - \frac{2\sum_{j\epsilon Scorrect\ _i}\sum_{g\epsilon S_j} X_{ij}\,(1 - X_{ig})}{N_{ci}}$$

where $S_{correct_i}$ includes items that are answered correctly by examinee i, $X_{i_j}$is student *i*'s score (1 or 0) to item *j*, S_j 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 answered correctly by examinee *i*.

A misfit between examinee *i*'s response and the reduced Q-matrix happens when $X_{i_j}\left(1 - X_{i_g}\right) = 1$. When examinee *i* correctly answers item *j*, X_{ij}=1, the examinee is expected to also answer item *g* that belongs to S_j correctly, X_{ig}= 1 ($g \in S_j$). If the examinee fails to answer item *g* correctly, then X_{ig}= 0, $X_{i_j}\left(1 - X_{i_g}\right) = 1$, and it is considered a misfit between examinee i's observed response pattern and the expected response patterns specified by the attribute hierarchy. Thus, $\sum_{j \epsilon Scorrect\ _i} \sum_{g \epsilon S_j} X_{ij} (1 - X_{ig})$is equal to the total number of misfits. N_{Ci} contains the total number of comparisons for items that are answered correctly by examinee i. When the numerator of the HCI is multiplied by 2, the HCI has the property of ranging from –1 to +1, which promotes interpretation (Leighton et al., 2009). Values of the HCI range from -1 to +1 were a value of 1 indicates an observed response pattern fitting an expected response pattern perfectly. Conversely, the HCI value is -1 when the response pattern maximally misfits the hierarchy. Cui (2007) claimed that median HCI values above 0.60 suggest moderate model-data fit whereas values above 0.80 indicate excellent fit. Once fit between the cognitive model and the observed data is established, the attribute probabilities for each examinee can be estimated. Next, I present a method to estimate the probability that examinees master the attributes using the artificial neural network (ANN) within the AHM approach.

Estimating Examinee Attribute Probabilities. ANNs serve as efficient models for statistical pattern recognition (Bishop, 2006) that are characterized by a set of nodes and connections between nodes. Artificial neurons are computational models inspired by biological neurons, so that ANN mimics the way biological neurons function, where a signal is sent to neurons through synapses. If the signal is large enough to surpass a particular threshold, then the neuron is activated and emits a signal through the axon. This signal might be sent to another synapse and might activate other neurons. The ANN functions in a similar way, where three components must be in place: inputs (resembling the synapses), weights (strength of the respective signals), and the output (whose activation depends on the strength of the signal). The neural network model is a nonlinear function from a set of input variables to a set of output variables controlled by a vector of adjustable parameters or weights (Bishop, 2006). Neural networks can be used to extract patterns and detect trends that are too complex to be noticed by humans or other computer techniques (Sengur, Turkoglu, & Ince, 2007). The process of adjusting the weights is called learning or training. Though the training begins with random weights, the program's goal is to find the set of weight values that will minimize the error.

The ANN is employed to estimate the probability associated with attribute mastery, given the observed item response pattern. In addition to the item responses, a matrix of expected attribute patterns and a matrix of expected responses are required to estimate the probabilities for attributes specified in the cognitive model. Figure 6 depicts these necessary components for computing the attribute probability with the ANN, using the hypothesized linear model containing six attributes.



Expected Attribute Pattern

A1	A2	A3	A4	A5	A6
0	0	0	0	0	0
1	0	0	0	0	0
1	1	0	0	0	0
1	1	1	0	0	0
1	1	1	1	0	0
1	1	1	1	1	0
1	1	1	1	1	1

Expected Response Pattern (per item)

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

Actual Students' Response (per item)

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

Outcome (Attribute Probability)

A1	A2	A3	A4	A5	A6

*Figure 6.*Necessary matrices for scoring under the ANN.
(Figure adapted from Lai & Gierl, 2010)

The expected attribute pattern is the transpose of a Qr matrix with one additional row containing 0s for all attributes. The first row of this matrix indicates a possible pattern in which none of the attributes in the hierarchy is mastered. The second row of Qr transpose indicates a pattern in which attribute 1 is mastered but not attributes 2 through 6; and in the last row, the pattern indicates that all six attributes are mastered. Corresponding to each of the attribute patterns is an expected response pattern (notice that the rows correspond). Row 1 of the expected response matrix should be interpreted as follows: Supposing that each attribute is measured by three items, an examinee who have only mastered attribute A1 (row 2 of the attribute pattern matrix) is expected to answer only the first three items correctly, producing the expected examinee response pattern (111000000000000).

From an AHM perspective, the goal of the ANN is to generate an outcome that expresses the probability of the student's attribute mastery, which is calculated by minimizing the differences between any expected response patterns with the student vector to find a pattern that is most fitting for the student (Lai & Gierl, 2010). Thus, in order to compute the students' attribute probability, two steps are necessary: the ANN training and scoring. The training process for the ANN is completed using the expected response pattern as inputs and expected attribute patterns as outputs. Using these two matrices, the weights are estimated. After the ANN has been trained, the network of the trained weights is used to calculate attribute-based probabilities from the actual student responses. The items function as input nodes, and the attribute probabilities for each student represent the output node. The output ranges from 0 to 1, where a larger value indicates that the examinee has a higher probability of possessing a specific attribute. These probabilities serve as diagnostic scores that may be reported to students, teachers, and parents.

Step 4. Score Reporting

The fourth step for a principled approach to test design specifies how diagnostic scores can be transformed into score reports for test users. Score reports must present information that is useful to the students, teachers, parents, and relevant school administrators. Because one of the purposes of CDA is to make diagnostic inferences about learners' cognitive strengths and weaknesses on the attributes required for mastering a specific domain, score reporting has been regarded as the "most critical step" (Gierl, Leighton, & Hunka, 2007) in the diagnostic testing process. Thus, evaluating score reports used in the Diagnostic Mathematics project was an important procedure in the process of implementing an assessment according to principled test design. The framework presented by Roberts and Gierl (2010) is used here to evaluate the score reports.

Roberts and Gierl (2010) presented an adaptation of the score reporting framework initially developed by Jaeger (1998) and Ryan (2003). In this framework, the following set of characteristics is suggested to be important when providing cognitive diagnostic feedback (see Table 3).

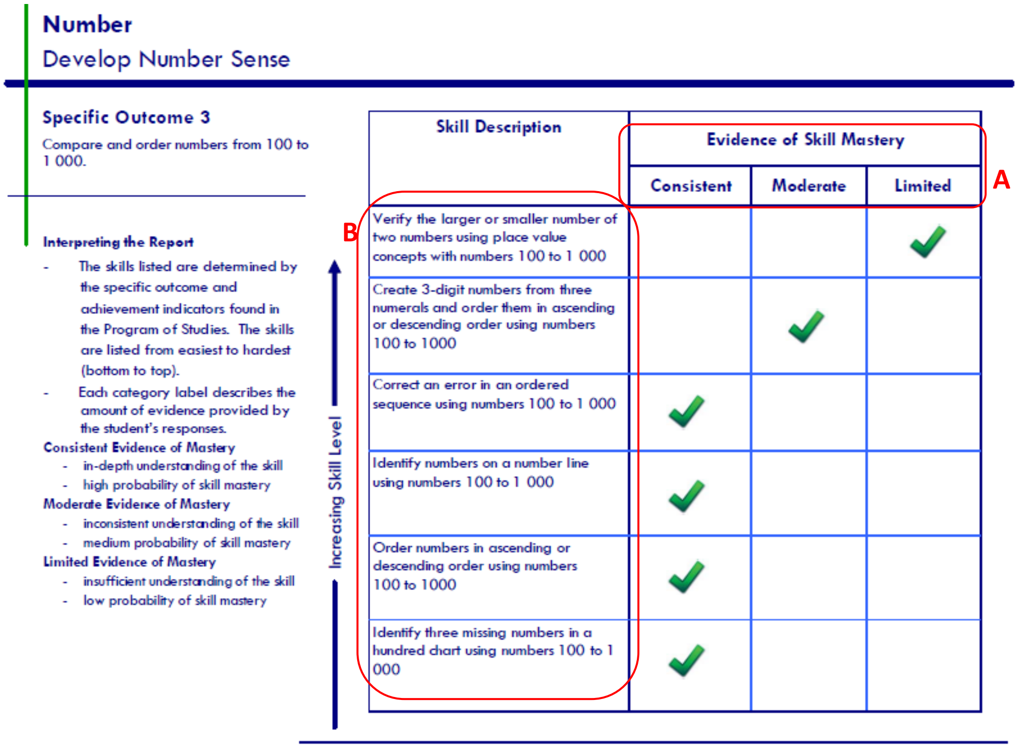
Table 3. *Framework for diagnostic reporting.*

Reporting Characteristics
Form of reporting results
Scale
Assessment unit
Reference for interpretation
Reporting unit
Error of measurement
Mode of reporting results
Numerical
Graphical
Narrative
Medium for dissemination of results
Application of design principles

*Table adapted from Roberts and Gierl (2010), p.30.

Form of Reporting Results. The form of reporting results is discussed by addressing its five defining features. Concerning the first feature, scale, the Diagnostic Mathematics project used the computed attribute probabilities as diagnostic scores for each examinee (see Figure 7, highlighted area A). Hence, the assessment unit of analysis is at the attribute level (see Figure 7, highlighted area B). In this report, the attribute probabilities were classified into three categories: “Consistent”, “Moderate”, and “Limited” evidence of mastery. The probability for the consistent evidence of mastery ranged from 0.80 to 1.00 and suggested in-depth understanding of the skill. The attribute probabilities for moderate evidence of mastery are higher than 0.50 and less than 0.80 and suggested an inconsistent understanding of the skill. Examinees with limited evidence of mastery received probabilities lower than 0.50, suggesting insufficient understanding of the skill. These cut offs are the current reporting standards used by Alberta Education for the Acceptable Standard and the Standard of Excellent. Other standards could be used; however, for convenience, this study adopted the existing Alberta Education standards. A description of the students characteristics (or traits) concerning the expected performance on *Subtracting 2-digit numbers* and *Comparing and ordering numbers*is presented on Appendix. For example, a student who performs at the *Consistent Evidence of Mastery* is expected to independently and confidently solve both familiar and unfamiliar problems while a student at the *Limited Evidence of Mastery*, independently, only can solve familiar problems. As discussed in the first section of this chapter, the attributes in thecognitive models were developed to represent the skills that would be mastered by moderately high achievers. Using the three levels of performance described in the Appendix, *moderately high achievers* correspond to students performing at the *Consistent Evidence of Mastery* and to some of the students performing at the *Moderate Evidence of Mastery* (i.e., the top performers of this category).

The cognitive model, together with these standards, served as reference for interpretation on the Diagnostic Mathematics project. Students in this reference group are capable of understanding and grasping the mathematical concepts at their grade level at an achievement level of approximately 70-80%. In a practical sense, moderately high achievers can be thought as a student that would probably achieve around 80% in math classroom assessments and around 75% on large scale achievement tests. In addition, diagnostic scores were interpreted within a criterion-referenced framework and they indicated how well examinees had learned a specific body of knowledge and skills and not how their performance compared to a group of students, as in norm-referenced tests. In general, score reports were developed to be helpful to teachers, students, and parents, as they constitute a major reporting unit. Even though the error of measurement was not shown on the Diagnostic Mathematics score reports, the attribute reliabilities—which serve as a measure of the precision of diagnostic scores—were computed using the AHM for this assessment. The decision to not present the reliability of the attributes was framed with a view to make the score reports as simple as possible, due to the young age of the examinees.


A
B
Increasing Skill Level

Number
Develop Number Sense

Specific Outcome 3
Compare and order numbers from 100 to 1 000.

Interpreting the Report

- The skills listed are determined by the specific outcome and achievement indicators found in the Program of Studies. The skills are listed from easiest to hardest (bottom to top).
- Each category label describes the amount of evidence provided by the student's responses.

Consistent Evidence of Mastery
- in-depth understanding of the skill
- high probability of skill mastery

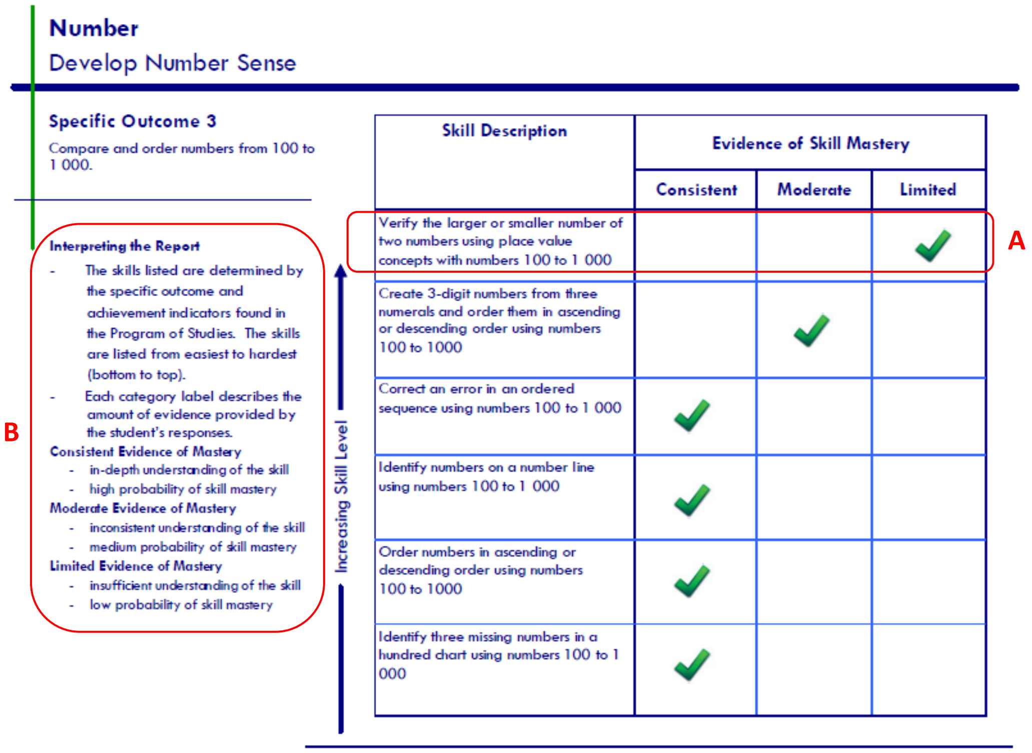
Moderate Evidence of Mastery
- inconsistent understanding of the skill
- medium probability of skill mastery

Limited Evidence of Mastery
- insufficient understanding of the skill
- low probability of skill mastery

Skill Description	Evidence of Skill Mastery		
	Consistent	Moderate	Limited
Verify the larger or smaller number of two numbers using place value concepts with numbers 100 to 1 000			✓
Create 3-digit numbers from three numerals and order them in ascending or descending order using numbers 100 to 1000		✓	
Correct an error in an ordered sequence using numbers 100 to 1 000	✓		
Identify numbers on a number line using numbers 100 to 1 000	✓		
Order numbers in ascending or descending order using numbers 100 to 1000	✓		
Identify three missing numbers in a hundred chart using numbers 100 to 1 000	✓		

Figure 7. Form of reporting results on the Diagnostic Mathematics score report.

Mode of Reporting Results. The mode of reporting results concerns the use of numerical, graphical, or text-based information on a score report. The score reports for the Diagnostic Mathematics project relied on visual/graphical and textual sources to inform students. Diagnostic information was provided through three categories of mastery. Check marks in the category label indicated the level of skill mastery based on the student's responses. For example, a checkmark in the *Limited* column for the top skill suggested that this examinee has limited skills in verifying the larger or smaller number of two numbers using place value concepts with numbers 100 to 1,000 (see Figure 8, highlighted area A). Textual information about how to interpret results is provided on the left side of the score report (see Figure 8, highlighted area B). Numerical outcomes of the attribute probability (a number ranging from 0.00 to 1.00) were not provided, since this information could be confusing and, potentially, misunderstood by the students due to their young age. Similarly, the Alberta Education team opted to remove information containing the total score from the prototype score report. This decision was based on the notion that the total score could be misleading. For example, a teacher could think that a student who answered more than half of the items correctly was doing alright, not noticing that most of the correctly answered items were of the lower level attributes and the student needed to work on the higher level skills. For this reason, Alberta Education decided to eliminate the total scores and report only check marks for each attribute, so that teachers and students would have a clear understanding of where to focus their efforts.



Number

Develop Number Sense

Specific Outcome 3

Compare and order numbers from 100 to 1 000.

Interpreting the Report

- The skills listed are determined by the specific outcome and achievement indicators found in the Program of Studies. The skills are listed from easiest to hardest (bottom to top).
- Each category label describes the amount of evidence provided by the student's responses.

Consistent Evidence of Mastery

- in-depth understanding of the skill
- high probability of skill mastery

Moderate Evidence of Mastery

- inconsistent understanding of the skill
- medium probability of skill mastery

Limited Evidence of Mastery

- insufficient understanding of the skill
- low probability of skill mastery

Skill Description	Evidence of Skill Mastery		
	Consistent	Moderate	Limited
Verify the larger or smaller number of two numbers using place value concepts with numbers 100 to 1 000			✓
Create 3-digit numbers from three numerals and order them in ascending or descending order using numbers 100 to 1000		✓	
Correct an error in an ordered sequence using numbers 100 to 1 000	✓		
Identify numbers on a number line using numbers 100 to 1 000	✓		
Order numbers in ascending or descending order using numbers 100 to 1000	✓		
Identify three missing numbers in a hundred chart using numbers 100 to 1 000	✓		

Figure 8. Mode of reporting results on the Diagnostic Mathematics score report.

Medium for Dissemination of Results. The electronic files of the score reports for the Diagnostic Mathematics were generated at the Centre for Research in Applied Measurement and Evaluation (CRAME) at the University of Alberta. These files were then sent to Alberta Education, who distributed them to teachers by e-mail. Teachers printed the files and distributed them to their students.

Application of Design Principles. Clarity of communication was emphasized in the process of designing score reports for the Diagnostic Mathematics project. Given students' young age, score reports were intended to be short and to provide only essential information on students' skills.

Sample and Data Collection Design

This study is part of a larger project undertaken by Alberta Education, focusing on diagnostic mathematics from Kindergarten to Grade 6. Grade 3 was selected as the target audience for this study due to its relatively advanced stage of item development, data collection, and student participation.

All schools in the province were given an opportunity to voluntarily participate in the study. In September 2009, a letter was sent from Alberta Education to the superintendents of all school divisions in the province, which were then forwarded to approximately 1,330 schools. Principals, in turn, were asked to convey the information to the teachers whom they supervise and teachers from these schools were asked to indicate their interest in participating in a research study involving diagnostic mathematics. Eighty teachers who expressed interest in participating in the study were contacted by email in February of 2010. Of the 80 teachers, 57 (71%) agreed to take part in the study. Forty-six teachers have had their students complete at least one of the diagnostic tests. This group of students formed the convenience sample for this study. These students were presented with both models; first with the multiple-choice test items (i.e., 18 items from *Comparing and ordering numbers*) and next with the numerical-response test items (i.e., 18 items from *Subtracting 2-digit numerals*). Only students who answered more than 50 percent of the items were included in the psychometric analyses. From the total of 385 students who answered the tests, 12 students left more than 50 percent of the questions blank on *Comparing and ordering numbers* and 61 students on *Subtracting 2-digit numerals*. Therefore, data from 373 students were considered for *Comparing and ordering numbers* and from 324 students for *Subtracting 2-digit numerals*.

Students were given 50 minutes to answer the 36 test items. The diagnostic test containing both cognitive models was ordered randomly except the first item, which was always related to the most basic skill of the model.

The test was administered at the end of the 2010 school year, after most of the instructional units were completed. The delivery method was computer-based, online, supervised by the teacher. The online data collection system was developed by Alberta Education and is called Quest A+. Students who engaged in problem-solving activities were allowed to use only the resources available in the test delivery system and the usage of a calculator, paper, or pencil was not permitted after the test was started.

Psychometric Analyses

Three psychometric analyses—hierarchy consistency index, students' attribute probability, and attribute reliability—were used to evaluate two cognitive models from the Diagnostic Mathematics: *Comparing and ordering numbers* and *Subtracting 2-digit numerals*.

Hierarchy Consistency Index

The HCI was used to evaluate whether the attribute hierarchies accurately reflected the cognitive attributes employed by the examinees on the Diagnostic Mathematics project for *Comparing and Ordering Numbers* and *Subtracting 2-digit Numerals*. The median HCI across all the students was used as the indicator of the overall model-data fit. The HCI ranges from -1.00 to 1.00, with values close to -1.00 indicating that students respond unexpectedly or differently from the responses expected under a given cognitive model (Cui & Leighton, 2009). Syntax developed by Cui in the Mathematica programming language was used to calculate the median and standard deviations of the HCI values.

Attribute Probability Estimates

This stage includes the generation of the Qr, attribute-by-item pattern, expected item response pattern matrices, and calculation of attribute probability estimates. After specifying the attribute hierarchy, the observed response data was compared to the expected response patterns derived from the cognitive model. Through this procedure the identification of the attribute combinations students are likely to possess, or not, is made possible.

Neural network analysis was used to investigate the relationship between the expected response patterns and their associated attribute patterns by presenting each pattern to the network repeatedly until each association is learned (Wang & Gierl, 2007). Once the network learns the associations successfully, a set of weight matrices are produced and used to obtain the probabilities of the individual attributes for any observed response pattern. A probability close to 1 indicates that the corresponding attribute is likely to be mastered by the examinee. Conversely, a probability close to 0 indicates that the corresponding attribute is not likely to be mastered by the examinee. This analysis was conducted using SPSS 16.0 software program (SPSS Inc., Chicago, IL).

Attribute Reliability

Attribute reliability refers to the precision of score decisions about examinees' mastery of specific attributes. Attribute reliability is calculated as a variation of Cronbach's α , as follows:

$$\alpha_k = \frac{n_k}{n_k - 1}\left[1 - \frac{\sum_{i \in S_k} W_{ik}^2 \sigma_{X_i}^2}{\sigma^2_{\sum_{i \in S_k} W_{ik} X_i}}\right]$$

where α_k is the reliability for attribute *k*, n_k is the number of items that are probing attribute *k*in the reduced Q-matrix (i.e., the number of elements in S_k), $\sigma_{X_i}^2$ is the variance of the observed scores on item *i*, $\sum_{i \in S_k} W_{ik}^2 \sigma_{X_i}^2$ is the sum of the weighted variance of the observed scores on the items that are measuring attribute *k*, and $\sigma^2_{\sum_{i \in S_k} W_{ik} X_i}$ is the variance of the weighted observed total scores.

High values of reliability are always preferred. Because CDA items are usually designed to measure a combination of attributes, the attribute dependency implies that prerequisite attributes of the hierarchy, such as attribute 1, are expected to have higher reliability estimates compared to attributes in the final nodes of the hierarchy, such as attribute 6 (refer to Figure 3 for an example). This occurs because the number of items that measure each attribute, either directly or indirectly, is influenced by the dependencies among the attributes.

CHAPTER IV: RESULTS

In this chapter I present the psychometric results for the two cognitive models in the Diagnostic Mathematics project—*Comparing and ordering numbers* and *Subtracting 2-digit numerals* (see Figures 1 and 4)—for the Grade 3 student sample that completed the tests. The chapter is organized into four main sections. Each of these sections addresses one type of psychometric analysis. In the first section I present the characteristics of the items, where difficulty and discrimination are discussed. In the second section I discuss the fit of the models relative to the actual student response data using the Hierarchy Consistency Index (HCI). In the third section I present results from the attribute probability estimates. These estimates provide examinees with specific information about their attribute mastery. In the fourth section I present results for the attribute reliability estimates, which are used to evaluate the consistency of decisions made with respect to the examinees' attribute mastery.

Group and Item Characteristics

The mean student performance (with standard deviation in parentheses) on the 18 items for the content area *Comparing and ordering numbers* was 10.55 (SD = 3.46); the mean item difficulty was 0.59 (SD = 0.18), and the mean item discrimination was 0.41 (SD = 0.17). For the content area *Subtracting 2-digit numerals*, the mean student performance was 13.03 (SD = 4.63), the mean item difficulty was 0.72 (SD = 0.18), and the mean item discrimination was 0.83 (SD = 0.18).

Table 4 summarizes the difficulty and discrimination level of each item measured as the percentage of correct answers (hereafter called p-values) and biserial correlation, respectively.

Table 4. *Difficulty and discrimination of test items in Comparing and ordering numbers and Subtracting 2-digit numerals.*

Comparing and ordering numbers			Subtracting 2-digit numerals		
	Diffic.[a]	Discr.[b]		Diffic.[a]	Discr.[b]
A1.1	0.83	0.55	A1.1	0.76	0.58
A1.2	0.88	0.61	A1.2	0.78	0.87
A1.3	0.76	0.53	A1.3	0.80	0.96
A2.1	0.78	0.55	A2.1	0.84	0.89
A2.2	0.66	0.53	A2.2	0.87	1.11*
A2.3	0.77	0.52	A2.3	0.85	0.94
A3.1	0.70	0.36	A3.1	0.86	1.06*
A3.2	0.60	0.45	A3.2	0.80	0.68
A3.3	0.63	0.53	A3.3	0.84	1.01*
A4.1	0.61	0.40	A4.1	0.77	0.79
A4.2	0.38	0.15	A4.2	0.72	0.78
A4.3	0.65	0.57	A4.3	0.75	0.82
A5.1	0.36	0.37	A5.1	0.74	0.82
A5.2	0.39	0.39	A5.2	0.77	0.92
A5.3	0.45	0.49	A5.3	0.79	0.98
A6.1	0.45	0.11	A6.1	0.35	0.56
A6.2	0.38	0.23	A6.2	0.34	0.57
A6.3	0.29	0.07	A6.3	0.33	0.53
Average	0.59	0.41	Average	0.72	0.83
Stdev	0.18	0.17	Stdev	0.18	0.18

Notes: [a]Percent of correct responses (p-values); [b]Biserial correlation; *Biserial correlation may be larger than 1.0 as the distribution of the data assumes a non-normal shape.

As shown on the left side of Table 4 (*Comparing and ordering numbers*), the lowest p-value (0.29) and discrimination (0.07) values were for item A6.3 (that is, the third item measuring Attribute 6). Interestingly, one item (A1.2) produced both the highest p-value (0.88) and discrimination (0.61) values. This outcome is interesting because items with p-levels in the mid-range usually have the best discrimination power and both very easy and very hard items are not likely to be strongly discriminating (PTI, 2006). In other words, an item with a very high p-value may be interpreted as being easy for almost the entire set of examinees and, for that reason, does not usually provide much discrimination or differentiation between high ability and low ability examinees (Slatt, Steiner, Hollar, et al., 2011). Dawber, Rogers, and Carbonaro (2004, citing Alberta Education, 1999) claim that 0.20 is a minimum level for the point-biserial correlation that can be deemed acceptable for differentiating examinees. Given that this value corresponds to a biserial correlation between 0.25 and 0.30 (Dawber et al., 2004), items with discrimination values lower than 0.20 were removed. Using this decision rule, items A4.2, A6.1, and A6.3 were removed from *Comparing and ordering numbers*, and these items were not included in the subsequent psychometric analysis. The decision to delete these items was also influenced by the fact that these items did not contribute to the pattern of decreasing p-values in the hierarchy (especially item A4.2). This pattern serves as one indicator of the alignment between test items and the cognitive model. If these two (items and model) are not aligned, the consistency of the hierarchy will decrease and adversely affect the HCI estimate. For these reasons, the three items were deleted.

Figure 9 shows the item characteristics after the three items (i.e., A4.2, A6.1, and A6.3) were removed from *Comparing and ordering numbers*. The grey line in this graph represents the discrimination values of the items. The biserial correlations did not vary substantially among items. This graph shows a slight decrease in the item p-values, as the attributes increase in complexity. Because the cognitive attributes that are believed to underlie test performance were placed in a hierarchical form of increasing difficulty, according to the conceptualized cognitive model, this pattern of decreasing p-values (as attributes increase in complexity) is expected.

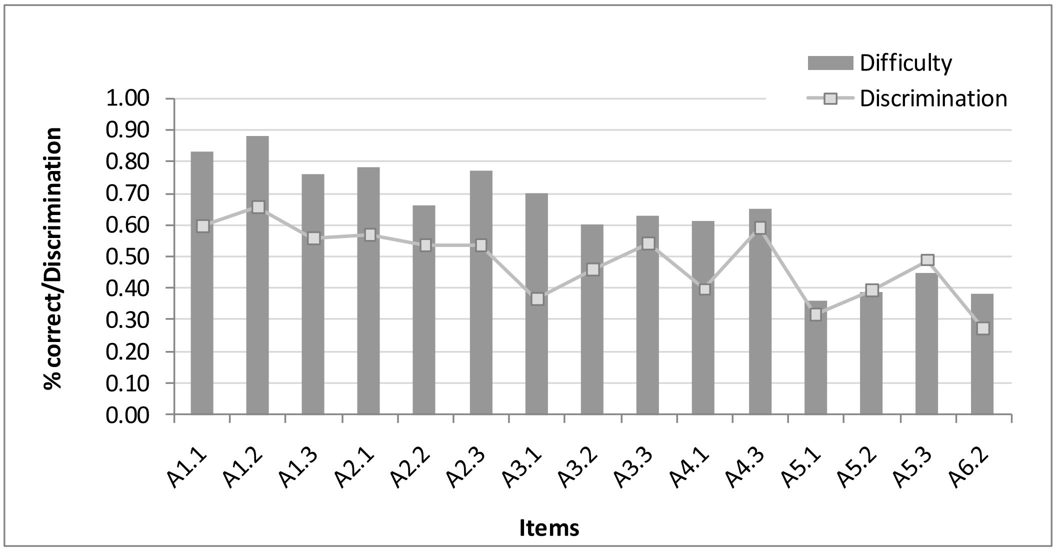


Figure 9. Difficulty and discrimination of the test items in *Comparing and ordering numbers.*

The overall trend for the second model, *Subtracting 2-digit numerals,* does not follow the same pattern of decreasing p-values as in the first model (see Figure 10). Except for attribute 6, whose item difficulty level abruptly decreased, the p-values of the remaining items were fairly similar (i.e., ranging from 0.72 to 0.86).

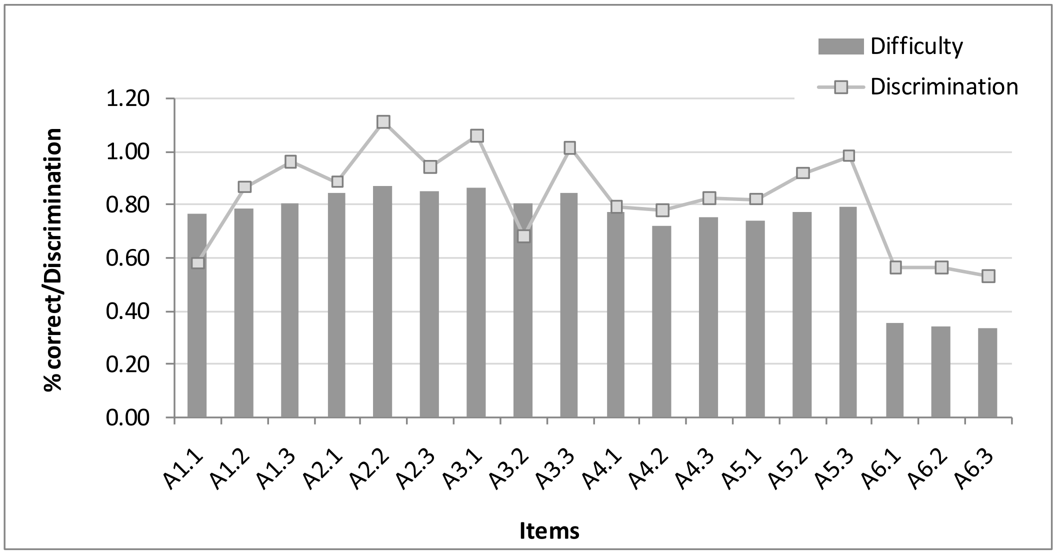


Figure 10. Difficulty and discrimination of the items in *Subtracting 2-digit numerals.*

To further evaluate the psychometric properties of the two cognitive models, examinees were divided into two groups according to their ability level as indicated by the total scores. The division of the examinees into groups according to their ability level was made as an attempt to better understand the difference in their response patterns and to investigate how well the cognitive models represented students with different ability levels. The cognitive attributes in the cognitive models were developed to represent the skills in which the moderately high achievers would be proficient[12]. It is assumed that moderately high ability students serve as a target for the low and moderate ability students and that the latter two groups will eventually develop the same type of skills displayed by the former group with proper instruction and adequate learning strategies.

In order to establish an empirical foundation for evaluating the data across these two groups, the K-means clustering method was used. The goal of the K-means algorithm is to find the best division of *n* entities (that is, the total number of students according to their total score) in *k* groups (i.e., two 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). Results from the K-means cluster analysis indicated that, in the content area *Comparing and ordering numbers*, the first group consisted of students with the minimum total score of 9.0 and the second group consisted of students with a total score of 8.0 or below. For *Subtracting 2-digit numerals*, the first group is formed by students who achieved total score values of 10.0 or higher. For the sake of simplicity, the first group is referred to as the “high” ability group and the second, the “low” ability group, although both groups would also include those students with “moderate” ability.

[12] In a practical sense, a moderately high achiever can be thought of as a student that would probably achieve around 80% on math classroom assessments and around 75% on large scale achievement tests.

The total score distributions for *Comparing and ordering numbers* and *Subtracting 2-digit numerals* are presented in Figures 11 and 12, respectively.

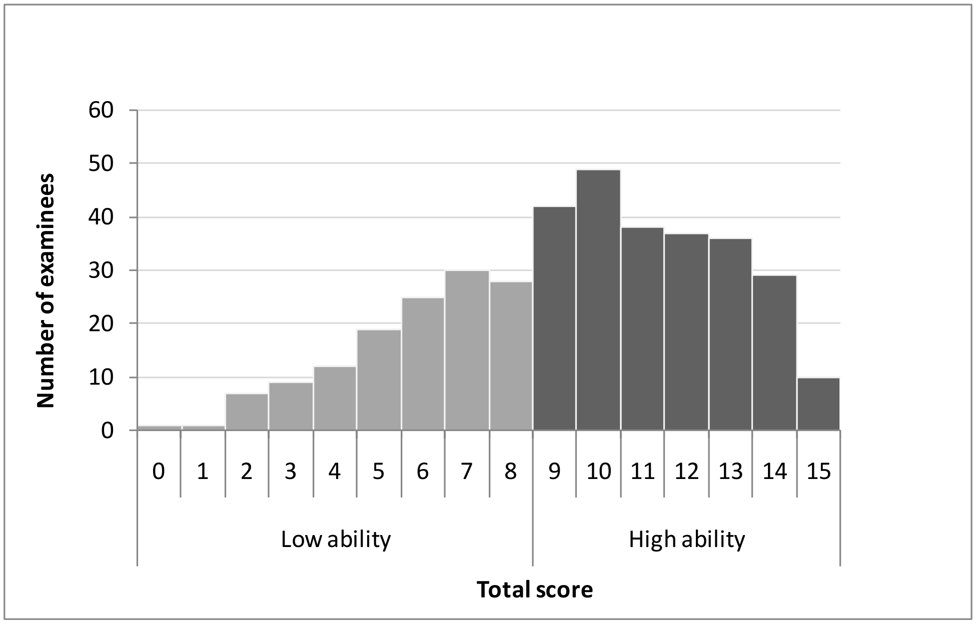


Figure 11. Total score distribution for *Comparing and ordering numbers.*

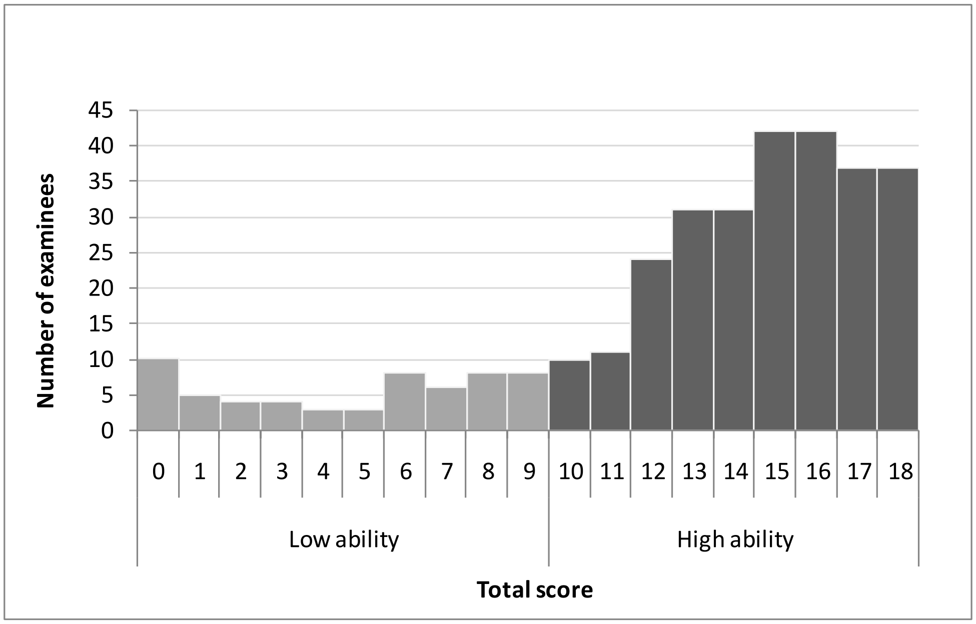


Figure 12. Total score distribution for *Subtracting 2-digit numerals.*

Figures 13 and 14 depict the difficulty level of the items, for each model, according to the two ability levels (low and high). Dark grey columns represent the percentage of high ability students who correctly answered each item. Light grey columns represent the percentage of low ability students who correctly answered each item. For example, item A1.1 (the first item measuring A1) was correctly answered by approximately 94 percent of the high ability students and by63 percent of low ability students. The trend line represents the attribute difficulty, which is the average of the items that directly measure the corresponding attribute. For example, A2 was measured by items A2.1, A2.2, and A2.3; thus, the attribute difficulty is the average of these three items. As a result, the average difficulty of A2 is 0.88 for the high ability group and 0.47 for the low ability group.

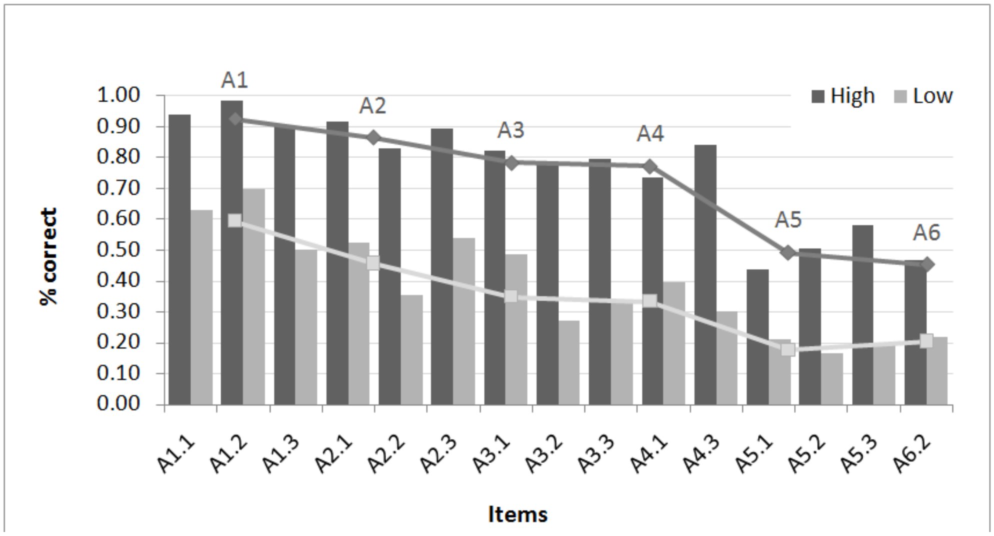


Figure 13. Item difficulty for *Comparing and ordering numbers* as a function of students' ability level.

As shown in Figure 13, changes in the performance of students on attributes 1 to 4 are more noticeable for low ability students. In both groups, the decrease in the p-values for attributes 5 and 6 is more pronounced.

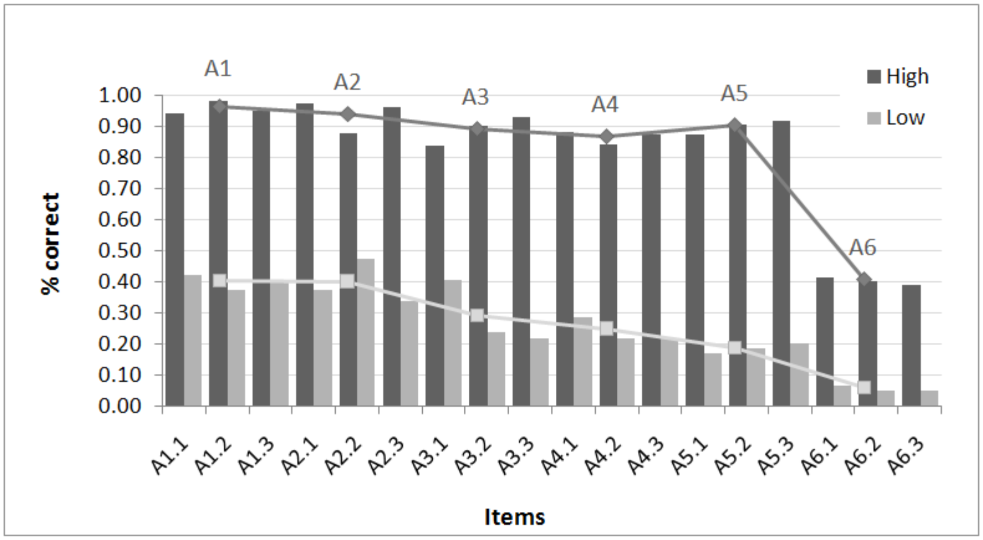


Figure 14. Item difficulty for *Subtracting 2-digit numerals* as a function of students' ability level.

In Figure 14, except for attribute 6, the difficulty level of the items only slightly changed for the high ability group. In addition, the average difficulty of the least complex attribute, A1, does not substantially differ from a more complex attribute, A5, where the former has a p-value of 0.96 and the latter 0.90. For the low ability students, the average difficulty consistently increased from one attribute to the other.

Summary

Overall, the two tests developed to measure the two cognitive models were moderately easy and, for the most part, highly discriminative (especially *Subtracting 2-digit numerals*). By comparing the results between two ability levels, the percent of correct responses is noticeably and consistently larger for the high ability students compared to the low ability students. A pattern of decreasing p-values was more evident for *Comparing and ordering numbers,* especially for the high ability group. This is a desirable characteristic because it confirms the structure of the cognitive models, where the skills are placed in a hierarchical form of increasing difficulty. The first five attributes of *Subtracting 2-digit numerals* were correctly answered by almost the entire set of high ability students, which reveals that the majority of the items on the test were very easy for this group. Implications of these results will be further examined as the three psychometric analyses within the AHM—hierarchy consistency index, students' attribute probability, and attribute reliability —are conducted and described in the next sections.

Hierarchy Consistency Index

In this section, the results from the HCI analysis are reported. The HCI results for the total sample as well as for the high and low ability groups are reported and compared. The median, mean, and standard deviations of the HCI values according to these three groups are summarized in Table 5.

Table 5. *HCI statistics for Comparing and ordering numbers and Subtracting 2-digit numerals as a function of students' ability level.*

	Comparing and ordering numbers			Subtracting 2-digit numerals		
	Total	High	Low	Total	High	Low
Median	0.57	0.76	-0.03	0.86	0.90	0.00
Mean	0.44	0.71	-0.05	0.68	0.83	0.00
Std. Dev.	0.48	0.22	0.43	0.45	0.19	0.64
N. of examinees	373	241	132	324	265	59

As shown in Table 5, for *Comparing and ordering numbers*, a median HCI value of 0.57 was obtained when the data for the total sample of students were considered. This median HCI value suggests that the fit between the cognitive model and the response data for the total sample was moderate. The highest median HCI value was obtained when the data for only the high ability students were used. The median HCI value of 0.76 indicates 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. Conversely, the cognitive model poorly predicted the observed response vectors of the low ability students (median HCI = -0.03).

For *Subtracting 2-digit numerals*, Table 5 indicates a strong fit between the cognitive model and the response data obtained from the total sample of students (median HCI = 0.86). The highest median HCI (0.90) was obtained when high ability students' response data were used, which suggests an excellent fit between the cognitive model and students' response data. However, the cognitive model did not seem to predict response patterns for the low ability students well, with the median HCI being 0.00.

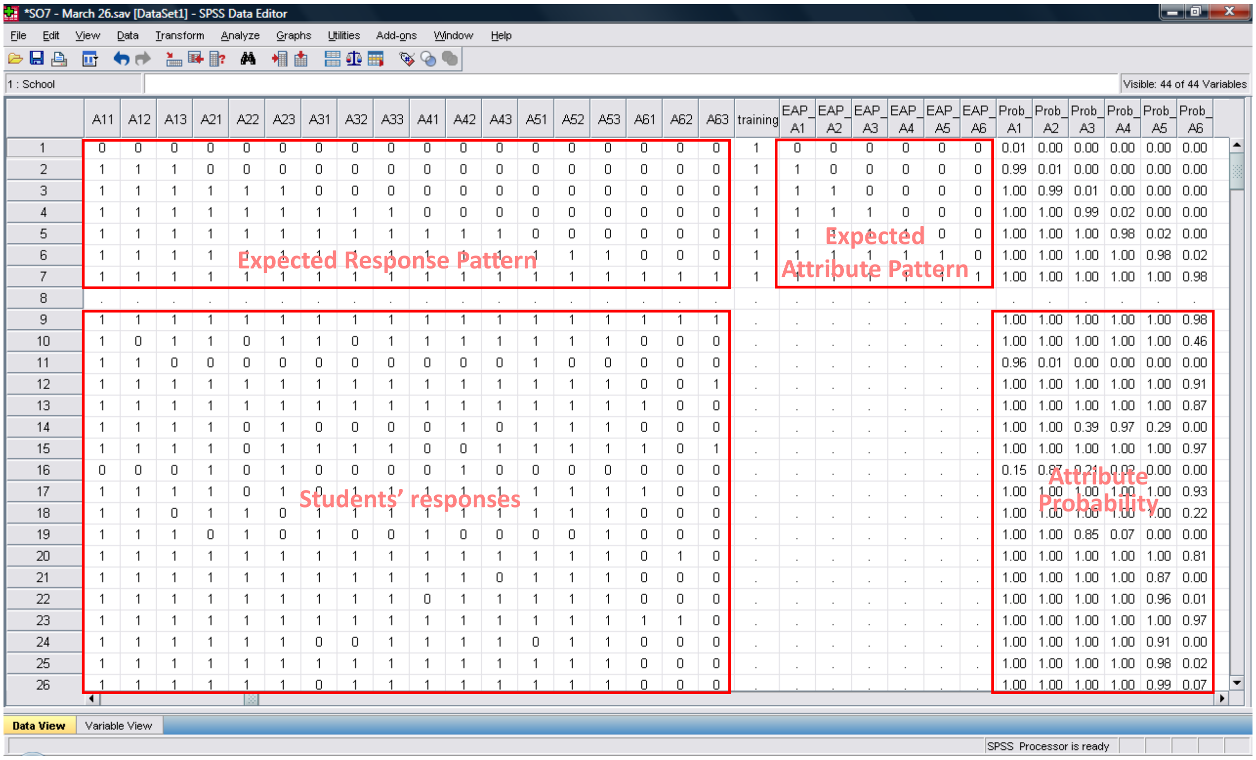
Summary

Cui (2007) 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 was considered moderate for *Comparing and ordering numbers* and strong for *Subtracting 2-digit numerals*, when data from the total sample of students are examined. Taking the two ability levels into consideration, the median HCI values suggest that the model-data fit was satisfactory for the high, but not for the low ability students. The poor fit between the cognitive model and the responses of the low ability examinees is not surprising as cognitive models were generated using moderately high achieving mathematics students as the point of reference by focusing on how these students would solve problems in mathematics (see Chapter III, Step 1). This assumption may substantially affect how well the cognitive models predict responses for low and high ability groups, as research has shown that expert and novice problem solvers differ in many ways (Chi, Glaser, & Farr, 1988; Leighton, et al., 2009; Mislevy, 1994).The HCI results indicate that the cognitive models fit not only the data from the total sample well but also for the intended population (i.e., high ability students).

Attribute Probability

The artificial neural network (ANN) was employed to estimate the probability associated with attribute mastery, given the observed item response pattern. In addition to the item responses, a matrix of expected attribute patterns and a matrix of expected responses were required to estimate the probabilities for attributes specified in the cognitive model. The process for developing the expected attribute pattern and the expected response pattern was described in the section *Step 3. Confirmatory Psychometric Analyses* of Chapter III.

The input to train the neural network was the matrix of expected responses, which was derived from the hierarchy as specified by the cognitive model. This matrix can be justifiably used for training the network because the model fit the data, as suggested by the HCI values. The relationship between the expected response vectors and their associated expected attribute was established by presenting each pattern to the network repeatedly until it learned each association. After this training, the network of the trained weights was used to calculate attribute probabilities from the actual student responses. The item responses functioned as input nodes, whereas the attribute probabilities for each student represented output nodes. Figure 15 displays the position of these matrices in SPSS.



	A11	A12	A13	A21	A22	A23	A31	A32	A33	A41	A42	A43	A51	A52	A53	A61	A62	A63	training	EAP_A1	EAP_A2	EAP_A3	EAP_A4	EAP_A5	EAP_A6	Prob_A1	Prob_A2	Prob_A3	Prob_A4	Prob_A5	Prob_A6
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0.01	0.00	0.00	0.00	0.00	0.00
2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0.99	0.01	0.00	0.00	0.00	0.00
3	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	1.00	0.99	0.01	0.00	0.00	0.00
4	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	1.00	1.00	0.99	0.02	0.00	0.00
5	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	0	0	1.00	1.00	1.00	0.98	0.02	0.00
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	0	1.00	1.00	1.00	1.00	0.98	0.02
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.00	1.00	1.00	1.00	1.00	0.98
8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	1.00	0.98
10	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	0	0	0	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	1.00	0.46
11	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	.	.	.	.	.	.	.	0.96	0.01	0.00	0.00	0.00	0.00
12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	1.00	0.91
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	1.00	0.87
14	1	1	1	1	0	1	0	0	0	0	1	0	1	1	1	0	0	0	.	.	.	.	.	.	.	1.00	1.00	0.39	0.97	0.29	0.00
15	1	1	1	1	0	1	1	1	1	0	0	1	1	1	1	1	0	1	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	1.00	0.97
16	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	.	.	.	.	.	.	.	0.15	0.87	0.21	0.02	0.00	0.00
17	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	0	0	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	1.00	0.93
18	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	1.00	0.22
19	1	1	1	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	.	.	.	.	.	.	.	1.00	1.00	0.85	0.07	0.00	0.00
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	1.00	0.81
21	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	0.87	0.00
22	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	0	0	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	0.96	0.01
23	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	1.00	0.97
24	1	1	1	1	1	1	0	0	1	1	1	1	0	1	1	0	0	0	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	0.91	0.00
25	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	0.98	0.02
26	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	0	.	.	.	.	.	.	.	1.00	1.00	1.00	1.00	0.99	0.07

Figure 15. Required matrices for computing students' attribute probability estimates in SPSS.

The descriptive statistics for the attribute probability estimates are presented in Tables 6 and 7 for *Comparing and ordering numbers* and *Subtracting 2-digit numerals*, respectively.

Table 6. *Mean and standard deviation for the attribute probability estimates as a function of the students' ability level on Comparing and ordering numbers.*

		A1	A2	A3	A4	A5	A6
Total	Mean	0.98	0.91	0.88	0.67	0.54	0.33
	Std. Deviation	0.12	0.24	0.29	0.43	0.45	0.42
High	Mean	1.00	0.99	1.00	0.87	0.66	0.39
	Std. Deviation	0.00	0.04	0.01	0.29	0.42	0.45
Low	Mean	0.94	0.75	0.67	0.31	0.31	0.22
	Std. Deviation	0.20	0.35	0.41	0.40	0.41	0.34

As shown in Table 6, the average attribute probability level ranged from 0.98 to 0.33 for the total sample. For instance, using the total sample of students, the value of 0.98 suggests that, on average, students have a high probability of mastering A1. Conversely, A6 received the lowest average probability, indicating that this attribute was not mastered by many students (on average, students had only a 33 percent chance of mastering it). The average attribute probability estimates from A1 to A6 are, ordered from least to most difficult. This pattern of decreasing probabilities is expected given the dependent relationship among the attributes—as specified in the cognitive model and operationalized by the linear hierarchy—where A1 is the simplest cognitive skill that also serves as a prerequisite to all other attributes. As a prerequisite attribute, A1 reflects the most basic skill of the cognitive model and is expected to be mastered by the majority of students. As the complexity of the skills increases, it is expected that the probability of mastering those skills decreases. Comparing high and low ability groups, the average probability estimates reveal that high ability students demonstrated higher probability of mastering the attributes. For instance, high ability students are much more likely to master the attribute A4 (average probability estimate = 0.87) than low ability students (average probability estimate = 0.31). The nearly *unanimous* mastery of the first three attributes (A1, A2, and A3) by the high ability students is demonstrated not only by the high average probability estimates (1.00, 0.99, and 1.00, respectively), but also by the small standard deviation associated with these three attributes (0.00, 0.04, and 0.01, respectively). Higher variability of the probability estimates was obtained when using data from the low ability students, ranging from 0.20 (A1) to 0.41 (A3 & A5).This increase in the variability of response may be related to a higher incidence of random or inconsistent responses by the low ability group. In addition the average attribute probability estimates from A1 to A6 are, again, ordered from least to most difficult. This pattern of decreasing probabilities for high and low ability students supports the hierarchical structure specified in the cognitive models and operationalized by the linear hierarchy.

Table 7. *Mean and standard deviation for the attribute probability as a function of the students' ability level for Subtracting 2-digit numerals.*

		A1	A2	A3	A4	A5	A6
Total	Mean	0.93	0.93	0.90	0.87	0.77	0.44
	Std. Deviation	0.25	0.24	0.28	0.33	0.38	0.46
High	Mean	1.00	1.00	1.00	0.99	0.91	0.53
	Std. Deviation	0.00	0.00	0.01	0.08	0.23	0.46
Low	Mean	0.61	0.63	0.46	0.32	0.16	0.03
	Std. Deviation	0.47	0.46	0.46	0.43	0.32	0.14

Likewise, the average attribute probabilities for *Subtracting 2-digit numerals*, in general, decreased as the complexity of the attribute increased, as shown in Table 7. For the total sample of students, the higher average probability values were observed for attributes A1 and A2, with the averages being the same (0.93). As expected, attribute A6 had the lower average attribute probability for the three comparison groups (0.44 for the total group, 0.53 for the high ability group, and 0.03 for the low ability group). As mentioned in Chapter III, the attribute probability estimates for the Diagnostic Mathematics project were classified into three categories: "Consistent", "Moderate", and "Limited" evidence of mastery. The probability for the consistent evidence of mastery ranged from 0.80 to 1.00 and suggested in-depth understanding of the skill. The attribute probability for moderate evidence of mastery is higher than 0.50 and less than 0.80 and suggested an inconsistent understanding of the skill. Examinees with limited evidence of mastery received a probability lower than 0.50, suggesting insufficient understanding of the skill. Taking these ranges into consideration, the performance of the total sample of students on the first four attributes (A1 to A4) denoted consistent evidence of mastery. The average probability estimates for this group suggest a moderate evidence of mastery on attribute A5, and limited evidence on attribute A6.Overall, a trend of increasing variability can be seen for the total group as the average attribute probability decreases. The high ability group demonstrated consistent evidence of mastery in five of the six attributes; the average probability estimates revealed limited evidence of mastery in only one attribute (A6). The variability of the estimates for this group was very small for the first four attributes (standard deviations varying from 0.00 to 0.08) and high for A6 (standard deviation = 0.46). Conversely, low ability students, on average, did not produce consistent evidence of mastery for any of the attributes. For four attributes (A3 to A6), the average probability estimates were lower than 0.50, suggesting insufficient understanding of these skills by many of the low ability students. Except for A6, the standard deviations for the low ability group were consistently higher than the standard deviations for the high ability group. As the probability estimate moves away from the extremes (either 0.00 or 1.00) toward 0.50, there is more opportunity for the estimates to vary (Fisher, 2009).Therefore, higher values of standard deviation were encountered for the low ability group, as their attribute probability estimates assumed more intermediate values than the high ability group.

Tables 8 and 9 show correlations among the attribute probability values for *Comparing and ordering numbers* and *Subtracting 2-digit numerals* for the three groups (total, high, and low), respectively.

Table 8. *Correlations among attribute probability estimates as a function of the students' ability level for Comparing and ordering numbers.*

Subgroup		A1	A2	A3	A4	A5	A6
Total	A1	1.00					
	A2	.49**	1.00				
	A3	.48**	.69**	1.00			
	A4	.28**	.57**	.64**	1.00		
	A5	.20**	.32**	.47**	.69**	1.00	
	A6	.11*	.11*	.29**	.37**	.70**	1.00
High	A1	1.00					
	A2	.84**	1.00				
	A3	.48**	.59**	1.00			
	A4	.60**	.40**	.37**	1.00		
	A5	.07	.033	.18**	.55**	1.00	
	A6	-.18**	-.13*	.09	.26**	.65**	1.00
Low	A1	1.00					
	A2	.45**	1.00				
	A3	.43**	.59**	1.00			
	A4	.24**	.49**	.62**	1.00		
	A5	.21*	.30**	.57**	.75**	1.00	
	A6	.13	.09	.44**	.48**	.81**	1.00

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

As shown in Table 8, all the correlations were significant for the total sample of students. For instance, the correlations among the attributes ranged from a low of 0.11 (between A1 & A6 and A2 & A6) to a high of 0.70 (between A5 & A6). Attributes related to one another most closely in the hierarchy yielded the highest correlations. This result likely occurs because these attributes are designed to measure the most similar skills and had a direct relationship or dependency (i.e., one is a prerequisite to the other) within the cognitive model. The correlations were lower for attributes that were moderately related or unrelated to one another as the skills became more distinct from one another in the cognitive model. Most of the correlations were significant for the high ability group, with the highest correlation value (0.84) registered between the adjacent attributes A1 and A2 and the lowest value (-0.18) between the non-adjacent attributes A1 and A6. For the low ability group, only two correlations were not significant (A1 & A6 and A2 & A6). The highest correlation value (0.81) was obtained between the adjacent attributes A5 and A6, and the lowest (0.09) between the non-adjacent attributes A2 and A6. In sum, similar to the total sample group, the correlational pattern for both high and low groups indicates that, in general, attributes related to one another most closely yielded the highest correlations, and attributes that were only moderately related yielded the lowest correlations, which suggests convergent and discriminant evidence supporting the cognitive models, respectively.

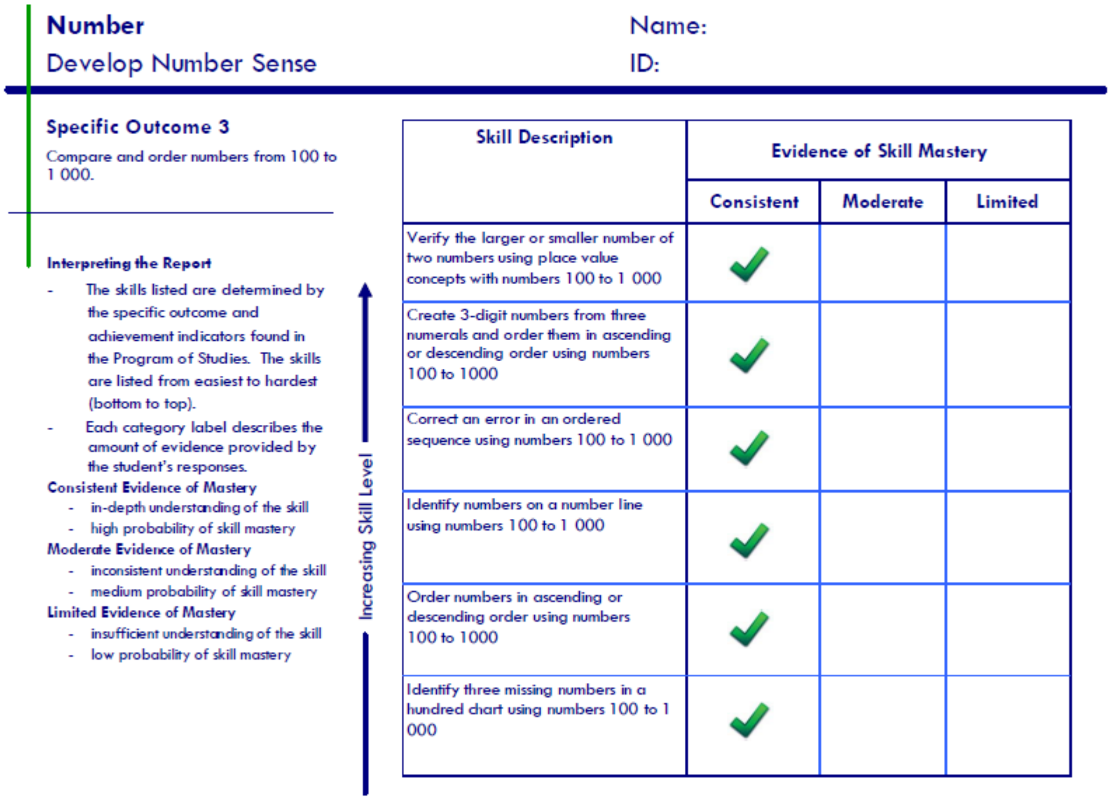
Table 9. *Correlations among attribute probability estimate as a function of the students' ability level for Subtracting 2-digit numerals.*

		A1	A2	A3	A4	A5	A6
Total	A1	1.00					
	A2	.86**	1.00				
	A3	.74**	.87**	1.00			
	A4	.67**	.74**	.85**	1.00		
	A5	.55**	.57**	.68**	.82**	1.00	
	A6	.26**	.27**	.33**	.38**	.58**	1.00
High	A1	1.00					
	A2	-.17**	1.00				
	A3	-.19**	.78**	1.00			
	A4	.00	.78**	.43**	1.00		
	A5	-.32**	.61**	.29**	.45**	1.00	
	A6	-.81**	.29**	.10	.13*	.48**	1.00
Low	A1	1.00					
	A2	.78**	1.00				
	A3	.54**	.79**	1.00			
	A4	.42**	.60**	.70**	1.00		
	A5	.30*	.41**	.50**	.79**	1.00	
	A6	.08	.20	.24	.38**	.60**	1.00

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

The correlations among the attributes in *Subtracting 2-digit numerals* produced a similar pattern of results for the three reference groups, where highest correlations were between adjacent attributes and the correlational values decreased as the distance between two attributes increased. For the total sample of students, the correlations between adjacent attributes ranged from a low of 0.58 (between A5 & A6) to a high of 0.87 (between A2 & A3). The correlations between non-adjacent attributes were lower because the attributes were more distant from one another in the cognitive model. For instance, using the total sample of students, the correlation values among A1 and A3 to A6 were 0.74, 0.67, 0.55, and 0.26, respectively. That is, as the distance between A1 and the other attribute increases, the correlation decreases. Similar to the total sample group, the correlational pattern for the high ability group indicates that, in general, attributes related to one another most closely yielded the highest correlations, and attributes that were only moderately related yielded the lowest correlations. For instance, a correlation value of 0.78 was registered between the adjacent attributes A2 and A3 and a correlation value of 0.10 was obtained between the non-adjacent attributes A3 and A6. Likewise, for the low ability group, smaller correlation values were registered between non-adjacent attributes (e.g., r_{16}=0.08) and greater values between adjacent attributes (e.g., r_{45}= 0.79).

To demonstrate how the AHM can be used to provide diagnostic feedback, three reports are presented in Figures 16to 18. The report in Figure 16 depicts a student who produced consistent evidence of mastery (all attributes were mastered); the report in Figure 17 was created for an examinee who produced moderate evidence of mastery (half of the skills were mastered); and Figure 18 was created for an examinee who produced limited evidence of mastery (only the first attribute was mastered).



Number Name:

Develop Number Sense ID:

Specific Outcome 3

Compare and order numbers from 100 to 1 000.

Interpreting the Report

- The skills listed are determined by the specific outcome and achievement indicators found in the Program of Studies. The skills are listed from easiest to hardest (bottom to top).
- Each category label describes the amount of evidence provided by the student's responses.

Consistent Evidence of Mastery

- in-depth understanding of the skill
- high probability of skill mastery

Moderate Evidence of Mastery

- inconsistent understanding of the skill
- medium probability of skill mastery

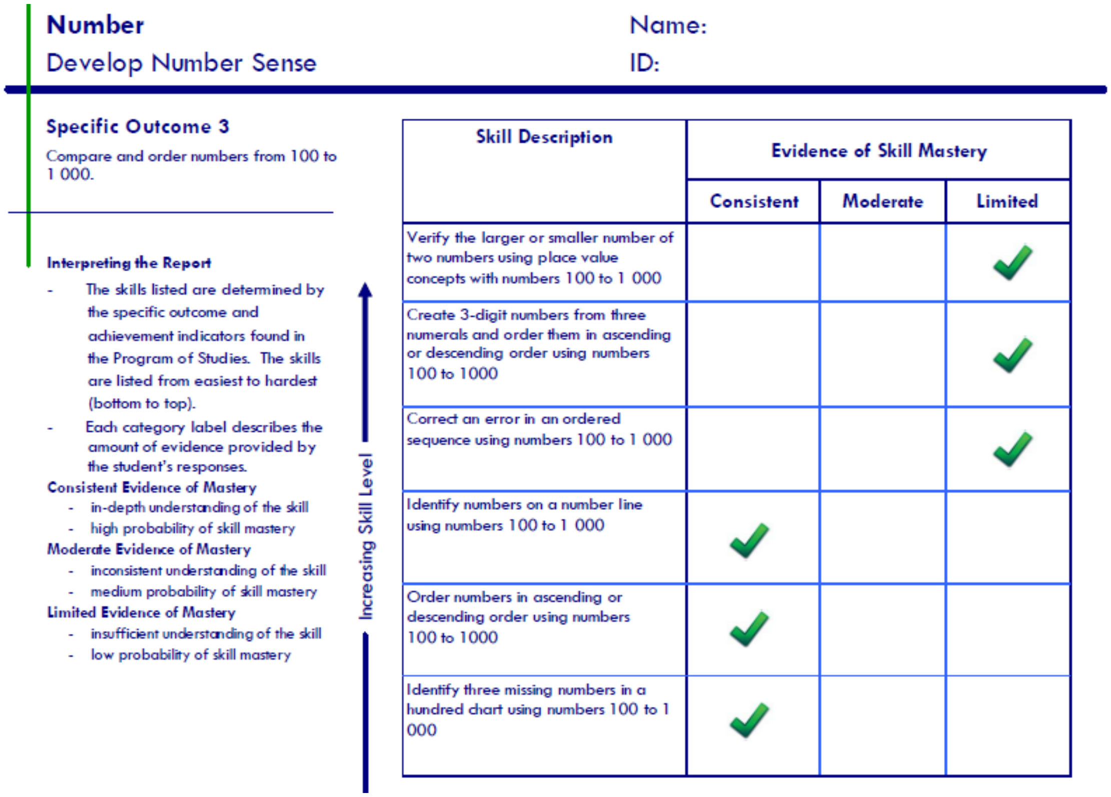
Limited Evidence of Mastery

- insufficient understanding of the skill
- low probability of skill mastery

Increasing Skill Level

Skill Description	Evidence of Skill Mastery		
	Consistent	Moderate	Limited
Verify the larger or smaller number of two numbers using place value concepts with numbers 100 to 1 000	✓		
Create 3-digit numbers from three numerals and order them in ascending or descending order using numbers 100 to 1000	✓		
Correct an error in an ordered sequence using numbers 100 to 1 000	✓		
Identify numbers on a number line using numbers 100 to 1 000	✓		
Order numbers in ascending or descending order using numbers 100 to 1000	✓		
Identify three missing numbers in a hundred chart using numbers 100 to 1 000	✓		

Figure 16. Cognitive diagnostic score report for an examinee producing consistent evidence of mastery.



Number Name:

Develop Number Sense ID:

Specific Outcome 3

Compare and order numbers from 100 to 1 000.

Interpreting the Report

- The skills listed are determined by the specific outcome and achievement indicators found in the Program of Studies. The skills are listed from easiest to hardest (bottom to top).
- Each category label describes the amount of evidence provided by the student's responses.

Consistent Evidence of Mastery

- in-depth understanding of the skill
- high probability of skill mastery

Moderate Evidence of Mastery

- inconsistent understanding of the skill
- medium probability of skill mastery

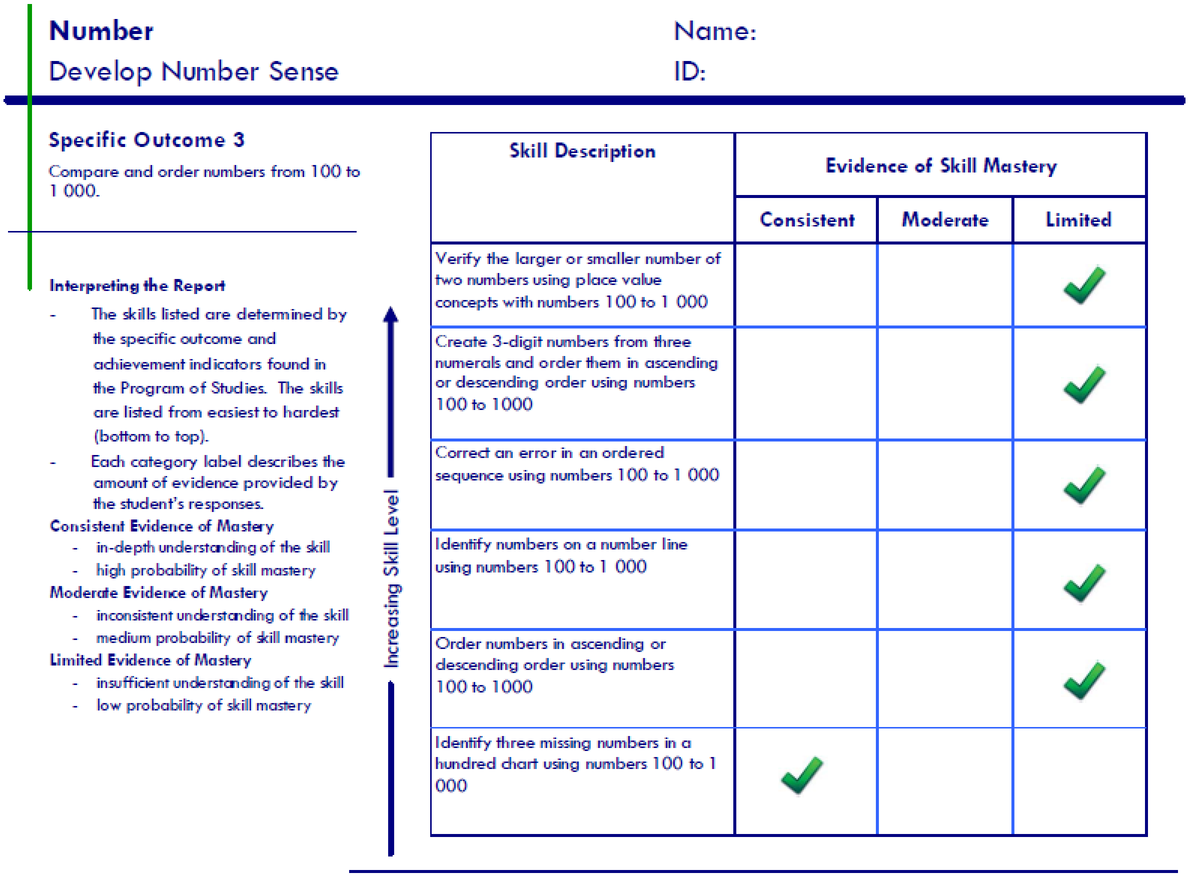
Limited Evidence of Mastery

- insufficient understanding of the skill
- low probability of skill mastery

Increasing Skill Level

Skill Description	Evidence of Skill Mastery		
	Consistent	Moderate	Limited
Verify the larger or smaller number of two numbers using place value concepts with numbers 100 to 1 000			✓
Create 3-digit numbers from three numerals and order them in ascending or descending order using numbers 100 to 1000			✓
Correct an error in an ordered sequence using numbers 100 to 1 000			✓
Identify numbers on a number line using numbers 100 to 1 000	✓		
Order numbers in ascending or descending order using numbers 100 to 1000	✓		
Identify three missing numbers in a hundred chart using numbers 100 to 1 000	✓		

Figure 17. Cognitive diagnostic score report for an examinee producing moderate evidence of mastery.



Number Name:

Develop Number Sense ID:

Specific Outcome 3

Compare and order numbers from 100 to 1 000.

Interpreting the Report

- The skills listed are determined by the specific outcome and achievement indicators found in the Program of Studies. The skills are listed from easiest to hardest (bottom to top).
- Each category label describes the amount of evidence provided by the student's responses.

Consistent Evidence of Mastery

- in-depth understanding of the skill
- high probability of skill mastery

Moderate Evidence of Mastery

- inconsistent understanding of the skill
- medium probability of skill mastery

Limited Evidence of Mastery

- insufficient understanding of the skill
- low probability of skill mastery

Skill Description	Evidence of Skill Mastery		
	Consistent	Moderate	Limited
Verify the larger or smaller number of two numbers using place value concepts with numbers 100 to 1 000			✓
Create 3-digit numbers from three numerals and order them in ascending or descending order using numbers 100 to 1000			✓
Correct an error in an ordered sequence using numbers 100 to 1 000			✓
Identify numbers on a number line using numbers 100 to 1 000			✓
Order numbers in ascending or descending order using numbers 100 to 1000			✓
Identify three missing numbers in a hundred chart using numbers 100 to 1 000	✓		

Figure 18. Cognitive diagnostic score report for an examinee producing limited evidence of mastery.

The three score reports in Figures 16 to 18were chosen to demonstrate how the results from the AHM could be used in score reporting, providing information regarding the examinees' attribute mastery. The score report has three parts: performance on skills (Evidence of Skill Mastery), the cognitive attributes measured by the test (Skill Description), and description of the performance (Interpreting the Report). In this report, the attribute probabilities were classified into three categories: "Consistent", "Moderate", and "Limited" evidence of mastery. The probability for the consistent evidence of mastery ranged from 0.80 to 1.00 and suggested in-depth understanding of the skill. The attribute probabilities for moderate evidence of mastery are higher than 0.50 and less than 0.80 and suggested an inconsistent understanding of the skill. Examinees with limited evidence of mastery received probabilities lower than 0.50, suggesting insufficient understanding of the skill. The mastery levels of the attributes are also expressed in the form of check marks. For instance, the student who received the score report depicted in Figure 16 mastered all six attributes of *Comparing and Ordering Numbers*. In Figure 17, the student showed consistent evidence of mastering the first three skills, but not the remaining attributes. Finally, Figure 18 depicts a situation where the student showed limited evidence of mastering five out of six attributes (only the most basic attribute was mastered).

Summary

Results from the attribute probability analysis for the total sample of students indicated that probably most of the attributes in the two cognitive models were mastered by the students (i.e., probability estimates greater than 0.80), an outcome that was also found when the total score and the item p-values were investigated (see first section of this chapter). In addition, the pattern of decreasing probabilities supports the structure specified in the cognitive model and operationalized by the linear hierarchy. That is, in the same way as the cognitive models were characterized by a set of attributes organized from simple to complex, the probability of mastering the attributes consistently decreased in response to this increase of complexity. Overall, the same pattern of decreasing probability estimates was also found when analyzing both the high and low ability groups, which supports the hierarchical structure specified in the cognitive models.

Another important result concerns the strong correlation between attributes sharing a direct relationship and the diminution of the correlation as the relatedness among attributes decreased. Gierl et al. (2010) proposed that this type of outcome indicates, respectively, convergent and discriminant evidence supporting the cognitive model. Convergent evidence is supported by finding that attributes most closely related to one another yield the highest correlations. Discriminant evidence is found when the lowest correlations among attributes occur for those attributes most unrelated to one another. In sum, the correlation patterns obtained using the total sample of students, as well as the high and the low ability groups, suggest convergent and discriminant evidence supporting the cognitive model for *Comparing and ordering numbers* and *Subtracting 2-digit numerals*.

Attribute Reliability

To evaluate the consistency of the decisions made with respect to the examinees' attribute mastery, reliabilities for the attributes in both models were calculated. Because the test items used in this study measured a combination of different attributes, each attribute contributed to only a part of the total item variance. When the number of items that both directly and indirectly measure a certain attribute decreases, the reliability estimates also decrease. For example, in *Comparing and ordering numbers*, attribute A1 was measured by all 15 items and thus had the highest reliability estimate (0.75). Attribute A5, on the other hand, had the lowest reliability estimate (0.44) because only four items measured this attribute. The reliability of A6 is undetermined because only one item was used to measure this attribute, thus making it impossible to compute reliability for this attribute.

Table 10. *Attribute reliability values for Comparing and ordering numbers and Subtracting 2-digit numerals as a function of the students' ability level.*

	N. of items/ ability level	A1	A2	A3	A4	A5	A6
Comparing and ordering numbers	*	15	12	9	6	4	1
	Total	0.75	0.67	0.58	0.51	0.44	NA
	High	0.20	0.17	0.25	0.34	0.35	NA
	Low	0.31	-0.01	-0.33	-0.05	0.07	NA
Subtracting 2-digit numerals		18	15	12	9	6	3
	Total	0.89	0.86	0.81	0.78	0.75	0.79
	High	0.48	0.51	0.48	0.46	0.58	0.76
	Low	0.73	0.59	0.44	0.41	0.38	0.73

Note: * Number of items that directly or indirectly require that attribute. For example, A1 is directly measured by three items and indirectly by 12 (in total by 15 items); A2 is directly measure by three and indirectly by six (in total by 12 items).

The reliability values for *Comparing and ordering numbers* ranged from 0.75 (A1) to 0.44 (A5) for the total sample of students. In social sciences, a widely accepted cut-off for the reliability is a value of 0.70 or higher (Nunnaly, 1978). Using this value as a threshold value for decision consistency, for the total sample, only the decisions made for Attribute 1 were determined to be consistent, given that this was the only attribute with reliability higher than 0.70. A plausible approach for improving the reliability estimates is the increase in the number of items that measure each attribute. The Spearman-Brown formula is used to estimate how much a test's reliability would increase if the test is increased by adding parallel items. The Spearman-Brown formula adapted for the AHM is specified as

$$\alpha_{AHM-SB_k} = \frac{n_k \alpha_{AHM}}{1+(n_k - 1)\alpha_{AHM}},$$

where α_{AHM-SB_k} is the Spearman-Brown reliability of attribute *k* if n_k additional item sets that are parallel to items measuring attribute *k* are added to the test. Supposing that the length of the test measuring the cognitive model *Comparing and ordering numbers* was increased three times (i.e., 3 x 15 = 45 items), the attribute reliability estimates from A1 to A5 would be, respectively: 0.90, 0.86, 0.81, 0.76, and 0.70. These results show that the reliability of the attributes would increase considerably by tripling the number of items measured by each attribute. These new reliability estimates would be above the suggested cut-off (0.70) and would yield consistent outcome interpretations.

The reliability values for the high ability group ranged from 0.35 (A5) to 0.17 (A2) on *Comparing and ordering numbers.* The reliability values for this group are not considered satisfactory for two reasons: the reliability values are very small and the pattern of increasing reliability estimates (as the number of items increases) was not obtained. A possible explanation for this occurrence is related to the fact that reliability estimates are affected by the variability of the scores and by the characteristics of the sample (Aguinis, Henle, & Ostroff, 2001). That is, the poor reliability estimates obtained when using data from the high ability students is likely to be due to the restriction of variance created by the split into two ability groups. For instance, the average observed variance for the three items associated with attribute A1 is very low (σ^2=0.04).

The reliability values for the low ability group ranged from 0.31 (A1) to -0.33 (A3) on *Comparing and ordering numbers.* Similar to the high ability group, the reliability estimates produced by the low ability students' data are very small and the pattern of increasing reliability estimates was not obtained. The variance of the observed scores for low ability students is not as low as the variance of the observed scores for the high ability students. For instance, the average observed variance for the three items associated with attribute A1 was 0.24, almost six times the observed variance of the high ability group. However, the pattern of responses produced by low ability students does not consistently match the expected response pattern as specified by the cognitive models. This divergence between actual and expected responses for the low ability group may be due to the students’ lack of skills for producing correct answers and/or to their randomly answering the test items (guessing).

Due to the problems associated with the high and low ability students’ data (e.g., small variability and random answers), reporting the reliability estimates by subgroup may be inappropriate. Interpretations based on the total sample are, therefore, preferred and adopted in this study.

For *Subtracting 2-digit numerals*, given that the reliability value for all six attributes exceeded 0.70 for the total sample, outcomes from these attributes were deemed consistent [the reliability values ranged from 0.89 (A1) to 0.75 (A5)]. The values of the attribute reliabilities are closely related to the number of items measuring each attribute. Hence, the reliability of attributes (directly or indirectly) measured by more items is expected to be higher than those measured by fewer items. However, this pattern was not observed throughout the entire hierarchy, asthe reliability for A5 (measured by 6 items) was not larger than A6 (measured by 3 items).

As with *Comparing and ordering numbers*, the reliability estimates produced using data from high and low ability students were not satisfactory and, overall, a pattern of increasing reliability estimates—as the number of items measuring a certain attribute increases—was not observed. Again, the use of the reliability estimates obtained from the total sample of students is more appropriate for making interpretations about the decision consistency for *Subtracting 2-digit numerals* possibly due to the small variability and high occurrence of random answers.

Summary

For *Subtracting 2-digit numerals*, the entire set of attribute reliability estimates based on the total sample of students were higher than the threshold (i.e., greater than 0.70), which indicated that decisions made regarding the six attributes would be considered consistent. Conversely, adopting the same threshold as the minimum acceptable reliability estimate, only one attribute in *Comparing and ordering numbers* (A1) would be considered to yield consistent test interpretations for the total sample of students. Because reliability is largely affected by the number of items on a test, an important aspect to consider is the increase in the number of items on future assessments, especially for *Comparing and ordering numbers*, which would increase the reliability of the diagnostic feedback provided to examinees, as previously demonstrated by the application of the Spearman-Brown formula. The reliability estimates for both cognitive models were not deemed satisfactory when splitting the sample into high and low achievers. The low reliability among the subgroups may be attributed to the low variability in the subgroups and high rates of random answers which, in turn, affects the reliability estimates. Therefore, the use of the subgroup reliability estimates is not recommended in this study. Instead, interpretations about the decision's consistency should be made using the reliability estimates from the total sample of students.

CHAPTER V: CONCLUSIONS AND DISCUSSION

Cognitive diagnostic assessments (CDA) can be described as an approach where the psychology of learning is combined with methods and models in statistics for the purpose of making inferences about students' specific knowledge structures and processing skills. In practical terms, a cognitive diagnostic assessment is an educational test designed to measure students' cognitive problem-solving strengths and weaknesses for diagnostic purposes (Ketterlin-Geller &Yovanoff, 2009). Since cognitive diagnosis is the process of inferring a student's cognitive state from his or her test performance (Ohlsson, 1986), one purpose of CDA is to promote formative inferences about student learning that could, potentially, allow teachers to redesign instructional approaches, evaluate instructional resources and strategies, and remediate students' weaknesses. In addition, these formative assessments have the potential to help motivate students and to empower them to take control of their own learning by providing detailed score reports to students thereby allowing them to pinpoint both their strengths and their difficulties.

Robert Stakes is quoted as saying, "When the cook tastes the soup, that's formative. When the guests taste the soup, that's summative" (quoted in Scriven, 1991, p. 169). The task of serving a delightful soup—one that meets the guests' needs and expectations while also tasting wholesome and delicious—is not an easy one. Applying this analogy in the CDA context, *improving the soup* requires the guidance from research in cognitive psychology combined with the principles in educational measurement to direct assessment design and analysis. So if tasting the soup is formative, then one needs to investigate, in advance, what kind of assessment is required, who needs to be assessed, what skills and abilities should be assessed, and what kind of information would be useful for guiding teaching and learning. A cognitive model serves a crucial role in the process of *improving the taste of the soup* as it provides a simplified description of human problem solving and a guide for item development (Leighton &Gierl, 2007a). A cognitive model is formed by different combinations of attributes—which are defined as the cognitive processing and knowledge required for solving a problem in a target domain (Tatsuoka &Tatsuoka, 1997). In addition, the use of a principled test design approach constitutes a critical step toward developing better diagnostic assessment, or using cooking terms, *improving the taste of the soup*.

This study used a four-step principled approach to test design characterized by: (1) the development of cognitive models, (2) the construction of test items according to the knowledge and skills specified in the cognitive model, (3) the use of a diagnostic psychometric analysis to assess the plausibility of the model-data fit relative to the intended underlying cognitive model and to providestudents' attribute probability estimates, and (4) the creation of detailed score reports that map examinees' mastery levels to provide more detailed information about students' problem-solving strengths and weaknesses. A cognitive model is critical in principled test design because it provides guidance for item development and for making diagnostic inferences about student test performance. Given that a cognitive model is a vital part of the design of CDAs, adhering to a principled approach to test design and analysis plays an important role in the Alberta Education Diagnostic Mathematics project.

To date, much of the research conducted on CDA has focused on retrofitted data, meaning fitting existing test items to an *ad hoc* cognitive model, rather than developing a cognition model first, and then creating items to measure the knowledge and skills specified in the model. Retrofitting is a method that uses existing test items in a CDA framework with the hope of extracting information about the cognitive attributes measured by these items, even when the items were not initially developed from a cognitive perspective. In this way, retrofitting a cognitive model to existing test items 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 occurs at all. Consequently, this retrofit model might not accurately represent students' knowledge and skills required for mastering a certain domain and, therefore, in most cases will yield unsatisfactory diagnostic classification results (Gierl et al., 2010).

While researchers in educational measurement have recently highlighted the need for explicit and clearly conceptualized cognitive models of learning (Leighton& Gierl, 2007a, 2011;Pellegrino et al., 2001), the number of practical applications in CDA has, so far, been relatively small. In addition, to our knowledge, research using the four-step principled approach to test design in the context of CDA has been non-existent. Gierl (2007) claimed that conducting these four steps "is not a standard approach to test design and it rarely, if ever, is used in operational testing situations" (p. 337). In an attempt to address this shortcoming, the purpose of my dissertation was to use the attribute hierarchy method to investigate the plausibility of modeling students' responses in a diagnostic assessment using a principled approach to test design and analysis. A principled test design approach is valuable for developing CDAs, highlighting the importance of a sound foundation—in this context, cognitive models—to guide test development, test scoring, and the inferences drawn about students based on their score report results. The Mathematics Diagnostic project serves as one of the first attempts at implementing principled test design in an operational testing program.

This chapter includes four sections. In the first section, the research questions are revisited together with a brief summary of the methods used for the present study. In the second section, a discussion of the key findings is presented. In the third section, the limitations of the study are discussed. In the fourth and final section, the recommendations for future research are outlined.

Restatement of Research Questions and Summary of Methods

The purpose of this study was to investigate the accuracy and consistency of two cognitive models in the Diagnostic Mathematics project. This assessment—characterized as diagnostic, cognitive based, and non-retrofit—began in January 2008 and was implemented and funded by the Learner Assessment Branch of Alberta Education.

The accuracy and consistency by which cognitive diagnostic assessments classify students' test responses are key components to providing diagnostic feedback about examinees' knowledge and skills (Zhou, 2010). Four research questions were addressed in this study:

1. How did the observed student response data fit the expected response data produced by the cognitive models created by content specialists? Did the fit for the observed and expected response data differ for high and low achieving students?
2. Were the attribute probability estimates ordered from easy to difficulty across the student sample for each of the attribute hierarchies? Was the attribute order the same for high and low achieving students?
3. Did the correlations among attributes show convergent and discriminant evidence supporting the hierarchical structure of attributes? Were the correlational patterns different for high and low achieving students?
4. How reliable were decisions about the mastery of specific attributes for the students who wrote the diagnostic test? Did the reliability estimates differ for high and low achieving students?

To answer the research questions, a study using a four-step principled approach to test design was implemented. Next, I summarize the important methodological aspects adopted in this study.

This study had four major steps. The first step was the development of the cognitive models. Cognitive models were developed by content specialists in two content areas—*Comparing and ordering numbers* and *Subtracting 2-digit numerals*. Each cognitive model had six attributes, as described in Chapter III (see Figures 1 and 4). The development of the cognitive models happened chronologically before the item development process, as required by the principled approach (see Gierl, 2007). Cognitive modeling was initiated as one of the first steps in the project, following project conceptualization (i.e., establishing the purposes of the project) and methodological planning. Items were developed only after cognitive models were created. The development of the cognitive models was conducted by experienced classroom teachers (ranging from 13 to 30 years) who also had familiarity developing large-scale educational tests (ranging from one to nine years). In addition, a large number of content specialists participated in the process of developing the cognitive models. Having 10 specialists involved in cognitive model development increases the likelihood that the generated models are accurate and trustworthy because multiple perspectives from a large group of knowledgeable and experienced participants foster discussion, which, in turn, promotes reasoning and the presentation of different opinions. Working in large groups also helps specialists reach new and diverse solutions, seek consensus and agreement, and promote collaboration.

The reference group for whom the cognitive models were designed consisted of moderately high achievers, i.e., students capable of applying the mathematical concepts with some degree of proficiency to solve problems. Moderately high ability students are expected to be confident learners who are able to create their own connections and apply their previous knowledge and skills to solve routine and novel problems. They demonstrate perseverance when working through challenging problems. Students in this reference group are capable of understanding and grasping the mathematical concepts at their grade level and at an achievement level of approximately 70-80%. The assumption behind the choice of developing cognitive models based on moderately high ability students is that low ability students will eventually reach the former group by having proper instruction and learning opportunities. More information about the characteristics of this group can be found in Chapter III, Step 1.

Content specialists were instructed to follow a set of requirements—that is, granularity, ordering, measurability, and instructional relevance—when developing the cognitive models. Hence, the knowledge and skills in both models were specified at a fine level of *granularity*, meaning that the skills in these cognitive models were specific and clearly defined. *Measurability* concerns the necessity that each skill be stated clearly enough to enable construction of items. The attainment of the skill *measurability* was enhanced by the fine-level of granularity, together with a clear description of the skills. Overall, the skills from both models, *Comparing and ordering numbers* and *Subtracting 2-digit numerals*, were described in a way that allowed a test developer to create three items to measure each skill. Cognitive models were also developed using the systematic approach of organizing the skills into an *ordered hierarchy* of increasing difficulty. Overall, the difficulty levels of attributes in both models increased as higher order attributes required increasing problem-solving skills, higher-order thinking skills, and/or deeper understanding of relevant concepts. Finally, the skills included in the cognitive models were developed to be *instructionally relevant* and meaningful to teachers and students. The cognitive models on the Diagnostic Mathematics project were based on the instructional objectives and learning outcomes required by the Program of Study for Mathematics. As a result, outcomes from the assessment may have the potential to be used for planning classes, modeling instruction, measuring student progress, and offering feedback to students and parents. Defining characteristics are critical for a better fit between the data and the cognitive model since they constitute the bridge between the cognitive model and the test items.

The second step was item development. After cognitive models were developed and approved, test items were created by a team of item writers who had extensive teaching experience in mathematics (17 years, on average) and had vast experience in writing test items for Alberta Education(5 years, on average). Hence, the process of developing items for the Diagnostic Mathematics project was based on the knowledge and skills specified in the cognitive models developed in the first step. Using cognitive models as a foundation for item development together with having appropriately qualified item writers, who received proper training and who used a well-established set of principles for item writing and item review, constitute important procedures implemented in the Diagnostic Mathematics project. Test items were then field tested with a sample of 383 grade 3 students. These students were presented with items from both cognitive models: 18 multiple-choice test items from *Comparing and ordering numbers* and 18 numerical response test items from *Subtracting 2-digit numerals*.

The third step consisted of conducting confirmatory psychometric analyses. Psychometric analyses using field test data from the Diagnostic Mathematics project were conducted with the attribute hierarchy method (AHM). The present study is one of the first applications of the AHM to non-retrofit data from an operational testing program. The AHM is a sophisticated psychometric procedure that can handle complex cognitive structures with dependencies among skills. The AHM is considered a *confirmatory* approach, where the plausibility of the model-data fit relative to the intended underlying cognitive model can be assessed. This diagnostic psychometric method was used because it supports an integrated view between cognitive psychology and educational testing. In this study, the use of the AHM for evaluating the cognitive models was performed in three ways. First, the HCI estimates were used to investigate the degree to which observed examinee response patterns were consistent with the specified attribute hierarchy. Second, the average of the attribute probability estimates and correlations among attributes were used to examine the interrelationship among the attributes specified in the hierarchy. Third, the attribute reliability estimates were used to assess the consistency of the decisions made with respect to examinees' mastery of specific attributes.

The fourth step was score reporting. This was the last step in the Diagnostic Mathematics project required to meet the principled approach to test design and analysis adopted in this study. This step is critical because it specifies how diagnostic scores can be transformed into reports for test users. Score reports must present information that is useful to the students, teachers, parents, and relevant school administrators, thereby allowing examinee's performance to be linked to specific cognitive inferences about their knowledge, processes, and strategies. Score reports play an important role in synthesizing cognition and assessment by combining cognitive models with statistical methods to permit test users to make inferences about students' cognitive strengths and weaknesses.

Key Findings

The purpose of this study was to evaluate cognitive models in the Diagnostic Mathematics project in two content areas—*Comparing and ordering numbers* and *Subtracting 2-digit numerals*—using three psychometric analyses developed for the AHM: the hierarchy consistency index, students' attribute probability estimates, and attribute reliability estimates. Specifically, this study answered four research questions:

Research Question 1: How did the observed student response data fit the expected response data produced by the cognitive models created by content specialists? Did the fit for the observed and expected response data differ for high and low achieving students?

Each of the cognitive models for the Diagnostic Mathematics project assessed a small number of knowledge components and a few specific cognitive skills in two content areas. This level of granularity yielded specific diagnostic inferences about examinees' mastery of the underlying knowledge and skills required to perform competently on the assessment tasks.

For *Comparing and ordering numbers*, a median HCI value of 0.57 was obtained when the data for the total sample of students were considered. This median HCI value suggests that the fit between the cognitive model and the response data for the total sample was moderate. For *Subtracting 2-digit numerals*, a median HCI value of 0.86 indicated a strong fit between the cognitive model and the response data obtained from the total sample of students. Hence, HCI values suggest the cognitive models adequately fit the data for the total sample of students. Taking the two subgroups into consideration, the observed student response strongly fit the expected response data produced by the cognitive model created by content specialists—where high achievers data yielded a median HCI of 0.90 on *Subtracting 2-digit numerals* and a median HCI = 0.76 on *Comparing and ordering numbers*. Conversely, the HCI values for low ability students indicated a poor fit between cognitive models and response data (median HCI = 0.00 for *Subtracting 2-digit numerals* and median HCI = -0.03 for *Comparing and ordering numbers*). Hence, the fit for the observed and expected response data differ for high and low ability students. The difference in the model data fit for the two subgroups was expected, as cognitive models were designed to represent a student who understands and grasps the mathematical concepts at an achievement level of approximately 70-80%, i.e., moderately high achievers.

In sum, the median HCI estimates suggest a satisfactory model fit between the two cognitive models in the Diagnostic Mathematics project and the observed response data from the total sample and moderately high ability students. Considering Cui’s guideline for evaluating the fit between the students’ response patterns and the expected response patterns based on the hierarchical relationship among attributes—median HCI values greater than 0.80 suggest excellent fit and values greater than 0.60 suggest moderate fit—then, the findings from the two cognitive models suggest a strong model-data fit.

Research Question 2: Were the attribute probability estimates ordered from easy to difficulty across the student sample for each of the attribute hierarchies? Was the attribute order the same for high and low achieving students?

Using data from the total sample of students, the average attribute probability estimates from A1 to A6 were ordered, as expected, from least to most difficult. The pattern of decreasing probabilities was expected given the dependent relationship among the attributes—as specified in the cognitive model and operationalized by the linear hierarchy—where A1 is the simplest cognitive skill that also serves as a prerequisite to all other attributes. As a prerequisite attribute, A1 reflects the most basic skill of the cognitive model and is expected to be mastered by the majority of the students. As the complexity of the skill increases, it is expected that the probability of mastering those skills decreases.

Similar attribute order was obtained when taking the high and low ability groups into consideration. The pattern of decreasing probabilities for the high and low ability groups supports the structure specified by the cognitive models on *Comparing and ordering numbers* and *Subtracting 2-digit numerals*. That is to say, in the same way as the cognitive models were characterized by a set of attributes organized from simple to complex, the probability of mastering the attributes consistently decreased in response to this increase in complexity for both, high and low achieving students. Hence, the ordering of the attributes did not differ as a function of the performance level of the students.

Research Question 3: Did the correlations among attributes show convergent and discriminant evidence supporting the hierarchical structure of attributes? Were the correlational patterns different for high and low achieving students?

Gierl et al. (2010) proposed that strong correlations between attributes sharing a direct relationship and decreasing correlations among attributes sharing indirect relationships this type of outcome indicates, respectively, convergent and discriminant evidence supporting the hierarchical structure of attributes that form the cognitive model. Convergent evidence is endorsed by finding that attributes most closely related to one another yield the highest correlations. This result occurs because these attributes measured the most similar skills and had a direct relationship or dependency (i.e., one is direct a prerequisite to the other) within the cognitive model. Discriminant evidence is found when the lowest correlations among attributes occur for those attributes most unrelated to one another. That outcome occurs because as the skills became more distinct from one another in the cognitive model, the correlations between attributes decrease.

Outcomes from the total sample of students indicated convergent evidence supporting the hierarchical structure of attributes that form the cognitive model as attributes related to one another most closely yielded the highest correlations. Discriminant evidence was also obtained as the correlations were lower for attributes that were moderately related to one another.

The correlational patterns for high and low achieving groups did not differ from the total sample of students. Overall, smaller correlation values were registered between non-adjacent attributes and greater values between adjacent attributes, outcome that suggests convergent and discriminant evidence supporting the hierarchical structure of attributes for both cognitive models, *Comparing and ordering numbers* and *Subtracting 2-digit numerals*.

Research Question 4: How reliable were decisions about the mastery of specific attributes for the students who wrote the diagnostic test? Did the reliability estimates differ for high and low achieving students?

A reliability estimate of 0.70 or higher is widely accepted as a threshold value for decision consistency (Nunnaly, 1978). Using this cut-off as a reference, all six attributes in *Subtracting 2-digit numerals* produced consistent interpretations about the mastery of attributes, when considering the total sample of students. Conversely, for *Comparing and ordering numbers*, only the decisions made for Attribute 1 were found to be consistent for the total sample. Using the Spearman-Brown formula it is possible to estimate how much a test's reliability would increase if the test is increased by adding parallel items in each attribute. Supposing that the length of the test measuring the cognitive model *Comparing and ordering numbers* was increased three times (i.e., 3 x 15 = 45 items), the attribute reliability estimates would be 0.70 or greater, that is, above the suggested cut-off for an assessment to yield consistent outcome interpretations for all of the attributes in this model. The values of the attribute reliabilities are closely related to the number of items measuring each attribute. Hence, the reliability of attributes (directly or indirectly) measured by more items is expected to be higher than those measured by fewer items. One possible factor that contributed to a higher reliability of the assessment measuring *Subtracting 2-digit numerals* is due to the slightly higher number of items in this assessment (18 items) than in the assessment measuring *Comparing and ordering numbers* (15 items). Another potential factor is related to the item format. Results of a study conducted by Oosterhof and Coats (1980), in which they compared the reliability of multiple-choice and completion-item formats, indicated that fewer math completion-items are required for obtaining reliability equal to that provided by multiple-choice items. This outcome is consistent with the findings in the current study, where resulting reliability estimates where higher for the numerical-response test (i.e., *Subtracting 2-digit numerals*) than the multiple-choice test (i.e., *Comparing and ordering numbers*). In sum, using the total sample of students, the decisions made with respect to examinees' mastery of attributes on *Subtracting 2-digit numerals* can be considered reliable. Because reliability is sensitive to the number of items in the sample—i.e., when the number of items is small, the calculated reliability tends to be low—one should use more caution when interpreting the results from the attribute reliability estimates for *Comparing and ordering numbers*. Alternatively, the developers could enhance their reliability estimates by increasing the number of items measuring the knowledge and skills in this model.

Regarding the second part of this research question, the reliability estimates for both cognitive models differed as a function of the students' ability level. For instance, the reliability estimates for the high ability on both cognitive models—*Comparing and ordering numbers* and *Subtracting 2-digit numerals*—are very small and the pattern of increasing reliability estimates (as the number of items increases) was not obtained. A possible explanation for this occurrence is related to the fact that reliability estimates are affected by the variability of the score and by the characteristics of the sample (Aguinis, Henle, & Ostroff, 2001). That is, the poor reliability estimates obtained when using data from the high ability students is likely to be due to the restriction of variance created by the split into two ability groups. The reliability estimates for the low ability group on both cognitive models were also very low and the pattern of increasing reliability estimates was not obtained. The variance of the observed scores for low ability students is not as low as the variance of the observed scores for the high ability students. However, the pattern of responses produced by low ability students does not consistently match the expected response pattern as specified by the cognitive models. The divergence between actual and expected responses may be due to the students' lack of skills for producing correct answers and/or to a random pattern of test items answers (e.g., guessing and skipping questions).

In sum, due to these problems associated with the high and low ability students' data (e.g., low variability and random answers), reporting the reliability estimates by subgroup may be inappropriate. Interpretations based on the total sample are, therefore, preferred and adopted in this study.

Limitations

The present study has at least four specific limitations. First, cognitive models were developed by experienced content specialists, but they were not submitted to empirical validation with students.Submitting the cognitive model to empirical validation plays an important role for CDAs as it can strengthen the validity argument about the construct being measured, and help clarify the psychology that underlies test performance thereby yielding more interpretable and meaningful test scores (Cui & Leighton, 2009). The use of inquiry methods for analyzing student thinking processes, such as task analyses and verbal protocols, are useful tools for validating the attributes required to develop a cognitive model. Relying solely on expert judgments for cognitive model development may result in models that provide an inaccurate understanding of the knowledge and skills examinees' use to solve items on a test, as it may overemphasize how content experts solve test items and underemphasize how students solve the same items. In other words, content specialists based their assertions on expectations about the cognitive path invoked by the item in the minds of their students who take the tests (Schmeiser & Welch, 2006). As a result, the cognitive models developed by these specialists may serve as a better approximation of the representations and descriptions of the knowledge and skills used by experts rather than students when solving items.

Second, the excellent fit between response data from high ability students and the cognitive model (e.g., median HCI = 0.90 in *Subtracting 2-digit numerals*) may be mainly due to the facility of the items, where nearly all high ability students completely mastered the five initial attributes in this cognitive model. The high p-value across these 15 items (0.91) and the small standard deviation (0.05) help support the conjecture that the pattern of correctly answering all items explains the excellent model-data fit for the high ability students, but not necessary the hierarchical structure outlined by the cognitive model. As a result, the order of the attributes could be changed without significantly altering the HCI of the model for students at this ability level. To enhance our understanding about how cognitive models could better represent the differences between high- and low-achievers, future research should be conducted using results from think-aloud protocols, as the data from a think-aloud analysis may help illuminate the solution paths that lead to both correct and incorrect responses.

Third, the low attribute reliability estimates, especially when performing subgroup analysis, constitutes a potential limitation in this study. Reliability estimation is critical in cognitive diagnostic assessment as it concerns to the consistency of the decisions made with respect to the examinees' strengths and weaknesses (Zhou, 2010). Specifically, the reliability estimates for the high and low ability groups were very small and the pattern of increasing reliability estimates (as the number of items increases) was not obtained on both cognitive models. This outcome may be related to the restriction of variance created by the split into two ability groups, the small number of items in the test forms, subgroup sample sizes, and random answers. Future research should be conducted to evaluate the effect of these factors on the reliability.

Fourth, the fact that most of the Alberta Education teachers participating in the cognitive model development were not *mathematicians* constitutes a potential weakness in the content specialist approach. Questions have been raised about whether elementary school teachers should be considered content specialists (National Council of Teachers of Mathematics, 2000; National Mathematics Advisory Panel, 2008; National Research Council, 1989). The Association of Mathematics Teacher Educators (2010) claim, for instance, that because most elementary teachers are generalists—that is, they study and teach all core subjects—they rarely develop in-depth knowledge and expertise with regard to teaching elementary mathematics. The current debate about the importance of formal education in *mathematics*, coupled with the fact that only 30 percent of the cognitive model developers were mathematicians, may indicate a misuse of the term “content specialist” in this dissertation. Most importantly, it could suggest that the cognitive models may not fully represent the actual knowledge and skills students use when solving test items. Perhaps, given the absence of content specialists graduated in mathematics in the Diagnostic Mathematic project, this study would benefit from having the cognitive models reviewed by external mathematicians specialized on elementary education and also by cognitive psychologists specialized in elementary mathematics.

Future Directions

Being among the first applications of the AHM to non-retrofit data from an operational testing program, the findings of this study add substantially to our understanding of the necessity of a principled approach to assessment design, and also contribute to a growing body of literature on CDA. However, more research needs to be undertaken to develop an in-depth understanding of the role of cognitive models, as well as to more fully comprehend the factors that affect how representative a cognitive model is to students with different ability levels.

The current study suggests at least four directions for future research. One line of future research is related to conducting protocol analysis to validate the cognitive models that have been developed by the content specialist approach. Empirically analyzing students' verbal responses is a vital approach to strengthen the validity argument about the construct being measured, and also helps clarify the psychology that underlies test performance which, in turn, may provide more interpretable and meaningful test scores (Cui & Leighton, 2009). The importance of conducting empirical studies on this topic is accentuated by the lack of literature introducing substantive psychological theory to support the development of cognitive models and delineating the psychological processes that reflect the construct measured by a test. A study could, for example, use the test items from the Diagnostic Mathematics project to collect verbal protocol data from a sample of students. Students would be asked to think aloud as they solved the items and have their answers audiotaped. Based on students' oral and written responses, flowcharts representing the cognitive processes reported by students would be created. These flowcharts would be used to evaluate both the item attributes and their hierarchical ordering of the Diagnostic Mathematics cognitive models.

A second line of research would involve the investigation of aspects that differentiate high- and low-achievers concerning the model fit. Because a single cognitive model does not explain the performance of different subgroups of examinees equally well (see Gotzmann, Roberts, Alves, & Gierl, 2009; Gotzmann & Roberts, 2010; Leighton et al., 2009), extending previous research by evaluating the adequacy of multiple cognitive models—developed not only by using the expertise of the content specialists, but also generated from verbal reports taken from different subgroups of students—may enhance the understanding about the subgroup differences. The use of the *combined* framework (see Conceptual Frameworks for Developing Cognitive Models section in Chapter II) may also play an important role not only in the development of cognitive models but also in their validation. Potentially, a study that combines a range of different methods for developing and validating cognitive models—for instance, cognitive theories, task analysis, expert review, verbal protocols, and interviewing students—could highlight the information necessary for understanding the differences in the way low and high-achievers perform on a specific cognitive model.

A third line of future research should consider the investigation of how factors such as students' response variability, number of items in the test forms, sample sizes, and random answers affect the reliability as a function of the subgroup analysis (i.e., high and low ability groups). Previous research conducted by Zhou (2010) manipulated the numbers of items on a test (12, 24, 36, and 48 items), the sample sizes (e.g., 250, 500, 750, and 1000), and levels of discrepancies between expected and observed responses (10%, 15%, 20%, and 25% of model-data misfit). However, Zhou's study did not consider the variability of the students' answers and the subgroup analysis (high and low groups) as variables of the study. Knowing how subgroup group variability affects the consistency of an assessment may help us to understand how to develop more reliable tests providing diagnostic information about examinees' knowledge and skills. Therefore, future simulation studies including these factors are recommended.

A fourth line of future research should develop and compare linear and non-linear cognitive models in mathematics. Due to the complex nature of learning, the linear model can be criticized for being too simplistic and for not representing the complex phenomenon in the teaching and learning of concepts in mathematics that underlie test performance. Besides regarding human abilities as hierarchically organized, non-linear models have the advantage of being able to form increasingly complex networks where the complexity is adjusted according to the cognitive problem-solving task (Leighton et al., 2004). The use of non-linear models may also add substantially to our understanding of the psychological processes that underlie test performance in Mathematics. A study could, for example, use think-aloud reports to generate cognitive models of task performance by administering a set of test items to students and collecting their reports as they solve the items. In addition, theories in a content domain could help to identify the required cognitive skills and the relationships among these skills. The four forms of hierarchical structures (namely, linear, convergent, divergent, and unstructured) suggested Leighton et al. (2004) could be used as the reference for comparison among the models. After the four cognitive models were elaborated, test items should be developed (using the respective cognitive model as basis) and administered to comparable sample of students. The AHM analyses could then be used to evaluate whether or not non-linear attribute hierarchies reflected the cognitive attributes employed by the examinees better than the linear hierarchy. Future studies on the topic are therefore recommended.

BIBLIOGRAPHY

Aguinis, H., Henle, C. A., & Ostroff, C. (2001). Measurement in work and organizational psychology. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology* (Vol. 1, pp. 27–50). London: Sage.

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

Alberta Education (2008).*Diagnostic Mathematics Item Writer's Manual*. Edmonton, AB: Alberta Education.

Alberta Education (2009).*Diagnostic Mathematics Item Writer's Manual. Edmonton*, AB: Alberta Education.

Alberta Education (2007).*The Alberta K-9 Mathematics Program of Studies with Achievement Indicators*. Edmonton, AB: Alberta Education.

American Association for the Advancement of Science (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist, 51,* 355-365.

Anderson, R. S. (1998). Why Talk About Different Ways to Grade? The Shift from Traditional Assessment to Alternative Assessment. *New Directions for Teaching and Learning, 74*, 5-16.

Association of Mathematics Teacher Educators (2010).*Standards for Elementary Mathematics Specialists: A Reference for Teacher Credentialing and Degree Programs*. San Diego, CA: AMTE.

Bailey, K. M. (1998). *Learning about language assessment: dilemmas, decisions, and directions*. Heinl & Heinle: US.

Baker, E. L. (1997). Model-based performance assessment. *Theory into Practice, 36*(4), 247-254.

Baxter, G. P. & Junker, B. (2001).*Designing Cognitive-Developmental Assessments: A Case Study in Proportional Reasoning*. Paper presented at the annual meeting of the National Council for Measurement in Education, Seattle, WA.

Begle, E. G. & Gibb, E. G. (1980). Why do research? In R. J. Shumway (Ed.), *Research in mathematics education (*pp. 3-19). Washington, DC: National Council of Teachers of Mathematics.

Béland, A., &Mislevy, R. J. (1996). Probability-based inference in a domain of proportional reasoning tasks. *Journal of Educational Measurement*, *33*, 3-27.

Berg, C. A. (2000). The development of adult intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 117-137). New York, NY: Cambridge University Press.

Birenbaum, M., Kelly, A. &Tatsuoka, K. (1993).Diagnosing Knowledge States in Algebra Using the Rule-Space Model. *Journal for Research in Mathematics Education*, *24*, 442–459.

Birenbaum, M., Tatsuoka, C., & Yamada, T. (2004). Diagnostic Assessment in TIMSS-R: Between-Countries and Within-Country Comparisons of Eighth Graders' Mathematics Performance. *Studies in Educational Evaluation, 30*, 151-173.

Birenbaum, M. & Tatsuoka, K. K. (1993). Applying an IRT-based cognitive diagnostic model to diagnose students knowledge states in multiplication and division with exponents. *Applied measurement in education, 6*(4), 225-268.

Bishop C. M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

Black, P., & Wiliam, D. (2001). *Inside the Black Box: Raising Standards Through Classroom Assessment*. BERA short Final Draft. Retrieved from http://www.collegenet.co.uk/admin/download/inside%20the%20black%20box_23_doc.pdf

Board on Testing and Assessment (BOTA) (2009). *Letter Report to the U. S. Department of Education on the Race to the Top Fund*. Retrieved from: http://www.nap.edu/catalog/12780.html

Bolt, D. (2007). The present and future of IRT-based cognitive diagnostic models (ICDMs) and related methods. *Journal of Educational Measurement, 44*(4), 377-383.

Briggs, D. (2007). Assessing what students know, how they know it, or both?. *Measurement: Interdisciplinary Research and Perspectives, 5*(1), 62-65.

Briggs, D. C. & Alonzo, A. C. (2009). *The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression*. Paper presented at the Learning Progressions in Science (LeaPS) Conference, Iowa City, IA.

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006). Diagnostic Assessment With Ordered Multiple Choice Items. *Educational Assessment, 11*(1), 33-63.

Casey, G. & Moran, A. (1989). The computational metaphor and cognitive psychology. *Irish Journal of Psychology, 10,* 143-161.

Castro, B. V. (2008). Cognitive Models: The Missing Link to Learning Fraction Multiplication and Division. *Asia Pacific Education Review, 9* (2), 101-112.

Chen, Z., & Siegler, R.S. (2000). Intellectual development in childhood. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 92–116). New York: Cambridge University Press.

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

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 1–185.

Cui, Y. (2007). *The hierarchy consistency index: Development and analysis*. Unpublished Doctoral Dissertation. University of Alberta: Edmonton, Alberta, Canada.

Cui, Y., & Leighton, J. P. (2009).The Hierarchy Consistency Index: Evaluating Person Fit for Cognitive Diagnostic Assessment. *Journal of Educational Measurement*, *46*, 429–449.

Cui, Y., Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2006). *The Hierarchical Consistency Index: A person-fit statistic for the Attribute Hierarchy Method.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Cui, Y., Leighton, J. P., & Zheng, Y. (2006). *Simulation Studies for Evaluating the Performance of the Two Classification Methods in the AHM*. Paper presented at annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Daniel, R. C & Embretson, S. E. (2010). Designing Cognitive Complexity in Mathematical Problem-Solving Items. *Applied Psychological Measurement*, *34*(5) 348–364.

Dawber, T., Rogers, W. T., & Carbonaro, M. (2004). *Robustness of Lord's Formulas for Item Difficulty and Discrimination Conversions between Classical and Item Response Theory Models*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Dawson, M. R. W. (1998). *Understanding cognitive science*. Oxford, UK: Blackwell.

DeCarlo, L. T. (2010). On the Analysis of Fraction Subtraction Data: The DINA Model, Classification, Latent Class Sizes, and the Q-Matrix. *Applied Psychological Measurement, 35 (1),* 8-26.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333-353.

de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595-624.

de la Torre, J. & Song, H. (2009). Simultaneous Estimation of Overall and Domain Abilities: A Higher-Order IRT Model approach. *Applied Psychological Measurement*, *33* (8), 620-639.

DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds, *Handbook of statistics: Vol. 26. Psychometrics* (pp. 979-1030). Amsterdam, Netherlands: Elsevier.

Dictionary (1971). *The American Heritage Dictionary of The English Language*. Boston: American Heritage Publishing Co., Inc. and Houghton Mifflin Company.

Dietal, R. J., Herman, J. L., & Knuth, R. A. (1991). *What does research say about assessment*? North Central Regional Educational Laboratory. NCREL, Oak Brook. Retrieved from: http://methodenpool. uni-koeln.de/portfolio/What%20Does%20Research%20Say%20 About%20Assessment.htm

Dikki, S. (2003). Assessment at a distance: Traditional vs. Alternative Assessments. *The Turkish Online Journal of Educational Technology, 2* (3), 13-19.

Dimitrov, D. M. (2007). Least Squares Distance Method of Cognitive Validation and Analysis for Binary Items Using Their Item Response Theory Parameters. *Applied Psychological Measurement, 31* (5), 367–387.

Dimitrov, D. M. &Raykov, T. (2003). Validation of Cognitive Structures: A Structural Equation Modeling Approach. *Multivariate Behavioral Research, 38* (1), 1-23.

Dogan, E. & Tatsuoka, K. K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R. *Educ Stud Math, 68,* 263–272.

Downing, S. M. (2006). Selected-response item formats in test development. In S.M. Downing, and T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 287-301). Mahwah, NJ: Lawrence Erlbaum.

Earl, L. M. (2003). *Classroom Assessment for Deep Understanding: Shifting from Assessment Of Learning to Assessment for Learning and Assessment as Learning*. Excerpted and adapted from Earl, L. (2003) Assessment as Learning: Using Classroom Assessment to Maximize Student Learning. Thousand Oaks: Corwin Press. Retrieved from http://www.npbs.ca/06-elements/deep-understanding-earl.pdf

Eisenhart, M. A. (1991). Conceptual frameworks for research circa 1991: Ideas from a cultural anthropologist; implications for mathematics education researchers. In *Proceedings of the 13th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 1, pp. 202–219). Blacksburg, VA. Retrieved from http://www.colorado.edu/education/faculty/margareteisenhart/Docs/Eisenhart_Conceptual%20Frameworks%20for%20Research.pdf

Embretson, S. (1995). A measurement model for linking individual learning to processes and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement*, *32* (3), 277-294.

Embretson, S. &Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38* (4), 343-368.

Eysenck, M. W., & Keane, M. T. (2000). *Cognitive psychology: A student's handbook* (4th ed.). Philadelphia: Psychology Press/Taylor and Francis.

Ferrara, S. & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K-12.In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 579-621). Westport, CT: National Council on Measurement in Education and American Council on Education.

Fisher, W. P. (2009). *Distinguishing between Consistency and Error in Reliability Coefficients: Improving the Estimation and Interpretation of Information on Measurement Precision.* Available at SSRN: http://ssrn.com/abstract=1685556

Frisby, C. L. (1999). Straight Talk About Cognitive Assessment and Diversity. *School Psychology Quarterly, 14* (3), 195-207.

Garfield, J. B. (1994). Beyond Testing and Grading: Using Assessment To Improve Student Learning. *Journal of Statistics Education, 2* (1). Retrieved from: http://www.amstat.org/publications/jse/v2n1/garfield.html

Gibbs, G., & Simpson, C. (2004). Conditions Under Which Assessment Supports Students' Learning?. *Learning and Teaching in Higher Education, 1, 3-31.*

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the Rule-Space model and Attribute Hierarchy Method. *Journal of Educational Measurement, 44*(4), 325–340.

Gierl, M. J., Alves, C., & Taylor-Majeau, R. (2010). Using the Attribute Hierarchy Method to make diagnostic inferences about examinees' 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 & Perspective*, *6*(4), 263-268.

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

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., Leighton, J. P., Wang, C., & Tan, X. (2005). *Technical Report #2: Evaluating primary skill categories*. Unpublished technical report. New York: The College Board.

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., & Leighton, J. P. (2007). Linking Cognitively-based models and psychometric methods. In C. R. Rao and S. Sinharay (Eds.) *Handbook of Statistics, Vol. 26: Psychometrics,* (pp. 1103-1106). Elsevier Science B. V.: The Netherlands.

Gierl, M. J., Roberts, M., Alves, C., & Gotzmann, A. (2009). *Using judgments from content specialists to develop cognitive models for diagnostic assessments*. In J. Gorin (Chair), How to Build a Cognitive Model for Educational Assessments. Paper presented at the 2009 annual meeting of the National Council on Measurement in Education, April 2009, San Diego, CA.

Gierl, M. J., Tan, X., & Wang, C. (2005). *Identifying content andcognitive dimensions on the SAT* (Research Report No. 2005-11). 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), 4-50.

Gierl, M.J., Zheng, Y., & Cui, Y. (2008). Using the attribute hierarchy method to identify and interpret cognitive skills that produce group differences. *Journal of Educational Measurement,45*(1), 65 – 89.

Gierl, M. J., Zhou, J., Alves, C. B. (2008). Developing a Taxonomy of Item Model Types to Promote Assessment Engineering. *The Journal of Technology, Learning, and Assessment, 7*(2), 1-51. Available from http://www.jtla.org.

Good, C.V. (Ed.) (1973). *Dictionary of Education*. New York: McGraw Hill Book Company.

Gorin, J. S. (2006). Test Design with cognition in mind. *Educational Measurement: Issues and Practice, 25*(4), 21-35.

Gott, S. (1990). Assisted learning of strategic skills. In N. Frederiksen, R. L. Glasser, A. M. Lesgold, and M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (p. 453-486). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Haladyna, T. M. (2004). *Developing and Validating Multiple- Choice Test Items* (3rd ed.).Mahwah, New Jersey: Lawrence Erlbaum Associates.

Haladyna, T. M. & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Educational*, *1*, 37-50.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-309.

Halford, G. S., Wilson, W. H. & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences, 21*, 723–802.

Hayes, N.A., & Broadbent, D.E. (1988). Two modes of learning for interactive tasks. *Cognition, 28*, 249- 276.

Henson, R. A., Templin, J. L., &Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191-210.

Huebner, A. (2010). An Overview of Recent Developments in Cognitive Diagnostic Computer Adaptive Assessments. *Practical Assessment, Research & Evaluation*, *15* (3), 1-7.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.)*, Cognitive diagnostic assessment for education: Theory and applications* (pp. 19-60). Cambridge, UK: Cambridge University Press.

Huff, K., Steinberg, L., & Matts, T. (2009). *The Promises and Challenges of Implementing Evidence-centered Design in Large Scale Assessment*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Jaeger, R. M. (1998). *Reporting the results of the National Assessment of Educational Progress* (NVS NAEP Validity Studies). Washington, DC: American Institutes for Research.

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y. -R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117-131). Ames, IA: Iowa State University

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258-272.

Kato, K. (2009). *Improving Efficiency of Cognitive Diagnosis by Using Diagnostic Items and Adaptive Testing*. Unpublished Doctoral Dissertation. Faculty of the Graduate School of the University of Minnesota.

Ketterlin-Geller, L. R., Jung, E., Geller, J., & Yovanoff, Y. (2008). *Project DIVIDE Instrument Development* (Tech. Rep. No. 08-10). College of Education, Behavioral Research and Teaching. Eugene, OR: University of Oregon.

Ketterlin-Geller, L., & Yovanoff, P. (2009). Cognitive diagnostic assessment in mathematics to support instructional decision making. *Practical Assessment, Research, & Evaluation, 14 (16)*, 1-11.

Klahr, D. (1992). Information processing approaches to cognitive development. In M. H. Bornstein & M. E. Lamb (Eds.), *Developmental Psychology: An advanced textbook* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Knol, A. B., Slottje, P., Sluijs, J. P., & Lebret, E. (2010). The use of expert elicitation in environmental health impact assessment: a seven step procedure. *Environmental Health, 9*, 1-16.

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

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation, 35,* 64–70.

Lai, H. & Gierl, M. J. (2010).*Two User-Friendly ANN Approaches for CDA Scoring*. Paper presented as a requirement of the course EDPY 699. Educational Psychology, University of Alberta.

Law, B., & Eckes, M. (1995).*Assessment and ESL: On the Yellow Big Road to the Withered of Oz*. Winnipeg, Man.: Peguis.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(4), 6–15.

Leighton, J. P. (2008). Where’s the psychology? A commentary on ‘‘Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art’’. *Measurement - Interdisciplinary Research and Perspectives*, *6*, 272-275.

Leighton, J. P., Cui, Y., & Cor, M. K. (2009). Testing expert-based and student-based cognitive models: An application of the AHM and HCI. *Applied Measurement in Education*, *22*, 1-26.

Leighton, J. P., & Gierl, M. J. (2011). *The Learning Sciences in Educational Assessment: The Role of Cognitive Models.* Cambridge, UK: Cambridge University Press.

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., Heffernan, C., Cor, M., Gokiert, R., & Cui, Y. (2008). *An experimental test of student verbal reports and expert teacher evaluations as a source of validity evidence for test development*. Edmonton, Canada: University of Alberta - Edmonton, Centre for Research in Applied Measurement and Evaluation.

Lester, F. K. (2010). On the Theoretical, Conceptual, and Philosophical Foundations for Research in Mathematics Education. In B. Sriraman & L. English (Eds.), *Theories of mathematics education: Seeking new frontiers. Advances in Mathematics Education,* Springer Science & Business, Berlin/London.

Lewin, K. (1952). *Field theory in social science: Selected theoretical papers by Kurt Lewin*. London: Tavistock.

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. Retrieved from http://www.csie.ntnu.edu.tw/~jlkoh/publications/pakdd2010.pdf

Lohman, D. E. (2000). Complex information processing and intelligence. In R. J. Sternberg (Ed.) *Handbook of human intelligence* (2nd ed., pp. 285-340).New York, NY: Cambridge University Press.

Luecht, R. M. (2007). *Assessment Engineering in Language Testing: from Data Models and Templates to Psychometrics*. Invited paper (and symposium) presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M. (2008). *Assessment engineering in test design, development, assembly, and scoring*. Invited keynote and workshop presented at the East Coast Organization of Language Testers (ECOLT) Conference. Retrieved from: http://www.govtilr.org/ Publications/ECOLT08-AEKeynote-RMLuecht-07Nov08[1].pdf

Luecht, R. M., Gierl, M, Tan, X., & Huff, K. (2006). *Scalability and the development of useful diagnostic scales*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Masters, J. S. (2010). *A Comparison of Traditional Test Blueprinting and Item Development to Assessment Engineering*. Unpublished Doctoral Dissertation. Faculty of the Graduate School at University of North Carolina at Greensboro.

McGlohen, M. K. (2004). *The Application of a Cognitive Diagnosis Model via an Analysis of a Large-Scale Assessment and a Computerized Adaptive Testing Administration*. Unpublished Doctoral Dissertation. Faculty of the Graduate School of University of Texas at Austin.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215–237.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

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). Westport, CT: American Council on Education/Praeger Publishers.

Mislevy, R. J. (2008). How Cognitive Science Challenges the Educational Measurement Tradition. *Measurement: Interdisciplinary Research and Perspective*, *6*(1 & 2), 124. Available online athttp://bearcenter.berkeley.edu/measurement/docs/CommentaryHaig_Mislevy.pdf

Mislevy, R. J., Behrens, J. T., Bennett, R. E., DeMark, S. F., Frezzo, D. C., Levy, R., Robinson, D. H., Rutstein, D. W., Stanley, K., Winters, F. I., & Shute, V. J. (2010). On the roles of external knowledge representations in assessment design. *Journal of Technology, Learning, and Assessment, 8*(2). Retrieved from http://www.jtla.org.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Mahwah, NJ: Erlbaum.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002).On the roles of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.*), Item generation for test development* (pp. 129-157). Mahwah, NJ: Lawrence Erlbaum

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

Mitchell, R. (1992). *Testing for Learning: How New Approaches To Evaluation Can Improve American Schools*. New York: The Free Press.

Mulder, G. (1983). The information processing paradigm: Concepts, Methods and Limitations. *Journal of Child Psychol. Psychiat., 24*, 19-35.

National Council of Teachers of Mathematics (NCTM) (2000). *Principles and Standards for School Mathematics*. Reston, VA: NCTM.

National Mathematics Advisory Panel (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. Retrieved March 20, 2008, from the U.S. Department of Education Web site: http://www2.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf

National Research Council (1989). *Everybody Counts: A Report to the Nation on the Future of Math Education*. Washington, DC: National Academies Press.

National Research Council (1996). *National science education standards*. Washington, DC: National Academy Press.

National Research Council (2001). *Knowing what students know: the science and design of educational assessment*. Committee on the Foundations of Assessment. Pellegrino, J.W., Chudowsky, N., and Glaser, R. (Eds.) *Division of Behavioral and Social Sciences and Education*. Washington, DC: National Academy Press.

National Research Council (2009). *Risk of Vessel Accidents and Spills in the Aleutian Islands: Designing a Comprehensive Risk Assessment*. Special Report, Transportation Research Board.

NAEYC (2002). *Early childhood mathematics: Promoting good beginnings*. National Association for the Education of Young Children (pp. 1-21). Retrieved from: www.naeyc.org/positionstatements/ mathematics

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*(4), 575–603.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (Eds.). (1995*). Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.

Nichols, P. D. &Joldersma, K. (2008). Review of Cognitive Diagnostic Assessment for Education: Theory and Applications by Leighton, J. P., & Gierl, M. J. *Journal of Educational Measurement*, *45* (4), 407–411.

Norman, D. A. (1981). Twelve issues for cognitive science. In Norman D. A. (Ed) *Perspectives on cognitive science*, pp. 265-295. Hillsdale, NJ: Erlbaum.

Nunnaly, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

Ohlsson, S. (1986). Some principles of intelligent tutoring. *Instructional Science, 14,* 293-326.

Oosterhof, A. & Coats, P. K. (1981). *Comparison of difficulties and reliabilities of math-completion and multiple-choice item format.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Palmer, S. E., & Kimchi, R. (1986). The information processing approach to cognition. In T. J. Knapp & L. Robertson (Eds.*), Approaches to cognition: Contrasts and controversies* (pp. 37-77). Hillsdale, NJ: Erlbaum.

Pashler, H. (1995). Attention and visual perception: Analyzing divided attention. In S. M. Kosslyn and D. N. Osherson (eds.) *Visual Cognition: An invitation to cognitive science* (pp 71-100), Vol 2. Cambridge, Mass: MIT Press.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999).Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad and P. D. Pearson (Eds.), *Review of research in education (Volume 24)* (pp. 307–353). Washington, DC: American Educational Research Association.

Pellegrino, J. W., Chudosky, N., & Glaser, R. (Eds.) (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington D. C.: National Academy Press.

Professional Testing Inc. (2006). *Step 9. Conduct the item analysis. Retrieved from* http://www.proftesting.com/test_topics/steps_9.php.

Roberts, M., Alves, C., Gotzmann, A., & Gierl, M. J. (2009). *Development of a mathematics construct map to promote diagnostic inferences about student performance*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Roberts, M. & Gierl, M. J. (2009). *Development of a Framework for Diagnostic Score Reporting.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Roberts, M. & Gierl, M. J. (2010). Developing Score Reports for Cognitive Diagnostic Assessments. *Educational Measurement: Issues and Practice, 29,* 25-38.

Russell, M., O’Dwyer, L., & Miranda, H. (2009). Diagnosing Students’ Misconceptions in Algebra: Results from an Experimental Pilot Study. *Behavior Research Methods, 41* (2), 414-424.

Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response theory models. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment in education: Theory and practice* (pp. 205-241).

Ryan, J. M. (2003). *An analysis of item mapping and test reporting strategies*. Retrieved August 14, 2007, from http://www.serve.org/Assessment/assessmentpublicationh1.php#StApub.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119–144.

Schmeiser, C. B. & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 307-353). Westport, CT: National Council on Measurement in Education and American Council on Education.

Scriven, M. (1991). *Evaluation thesaurus* (4thed.). Newbury Park, CA: Sage.

Sengur, A., Turkoglu, I., & Ince, M. C. (2007). Wavelet packet neural networks for texture classification*. Expert Systems with Application, 32*, 527-533.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics, 31*(1), 1-34.

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

Siegler, R. S. (2003). Implications of cognitive science research for mathematics education. In Kilpatrick, J., Martin, W. B., & Schifter, D. E. (Eds.), *A research companion to principles* *and standards for school mathematics* (pp. 219-233). Reston, VA: National Council of Teachers of Mathematics.

Simonson, M., Smaldino, S., Albright, M., & Zvacek, S. (2000). Assessment for distance education. In *Teaching and Learning at a Distance: Foundations of Distance Education.* Upper Saddle River, NJ: Prentice-Hall.

Shute, V. J., Graf, E. A., & Hansen, E. (2006).*Designing adaptive, diagnostic math assessments for individuals with and without visual disabilities*.ETS Research Report, RR-06-01 (pp. 1-46), Princeton, NJ.

Slatt, L. M., Steiner, B.D., Hollar, D. W., Chessman, A.W., Xin J., & Hedgpeth, M.W. (2011). Creating a multi-institutional family medicine clerkship examination: lessons learned. *Family Medicine*, *43*(4), 235-9.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wager (Eds), *Mind in context: Interactionist perspectives on human intelligence*, pp. 3–37. Cambridge, MA: Cambridge University Press.

Snow, R. E., &Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement (3rd Edition),* (pp. 263–330). New York: Macmillan.

Solso, R. L., MacLin, M.K, & MacLin, O.H. (2005). *Cognitive Psychology*. Boston: Pearson.

SPSS (2008). *Statistical Package for the Social Sciences 16.0*.SPSS Inc., Chicago, IL.

Stiggins, R. (2002). Assessment crisis: the absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758-765. Retrieved from http://electronicportfolios.org/afl/Stiggins-AssessmentCrisis.pdf

Struyven, K., Dochy, F., & Janssens, S. (2003). Students' perceptions about new modes of assessment: a review. In M. Segers, F. Dochy, & E. Cascallar (eds), *Optimising new modes of assessment: in search of qualities and standards*. Boston: Kluwer.

Struyven, K., Dochy, F., &Janssens, S. (2008). The effects of hands-on experience on students' preferences for assessment methods. *Journal of Teacher Education, 59*(1), 69-88.

Tanner, H., & Jones, S. (2003). *Marking and Assessment.* London, New York: Continuum.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 34-38.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto, (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453-488). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive Assessment: An Introduction to the Rule Space Method* (Multivariate Applications Series). Routledge Academic.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of Diagnosed Mathematical Content and Process Skills in TIMSS-R across a Sample of 20 Countries. *American Educational Research Journal, 41* (4), 901-926.

Tatsuoka, K. K., & Tatsuoka, M. M. (1992). *A psychometrically sound cognitive diagnostic model: Effect of remediation as empirical validity* (Research report). Princeton, N J: Educational Testing Service.

Tatsuoka, K. K., & Tatsuoka, M. M. (1997). Computerized adaptive diagnostic testing: Effect on Remedial Instruction as Empirical Validation. *Journal of Educational Measurement, 34,* 3-20.

VanderVeen, A., Huff, K., Gierl, M. J., McNamara, D. S., Louwerse, M., & Graesser, A. C. (2007). Developing and validating instructionally relevant reading competency profiles measured by the critical reading sections of the SAT. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 137-172). Mahwah, NJ: Erlbaum.

Wang, C. (2007). *Investigating the Cognitive Processes Underlying Student Performance on the SAT Critical Reading Subtest: An Application of the Attribute Hierarchy Method*. Unpublished Doctoral Dissertation. University of Alberta: Edmonton, Alberta, Canada.

Wang, C., & Gierl, M. J. (2007).*Investigating the Cognitive Attributes Underlying Student performance on the SAT Critical Reading Subtest: An Application of the Attribute Hierarchy Method.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Winterton, J., Delamare Le Deist, F. & Stringfellow, E. (2006).*Typology of Knowledge, Skills and Competences: Clarification of the concept and prototype*. CEDEFOP Reference series 64, Luxembourg: Office for Official Publications of the European Communities.

Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. In J. P. Leighton., & M. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education*. New York: Cambridge University Press

Ye, F. (2005). *Diagnostic Assessment of Urban Middle School Student Learning of Pre-Algebra Patterns*. Unpublished Doctoral Dissertation. The Ohio State University, USA.

Zhou, J. (2010). *Estimating Attribute-Based Reliability in Cognitive Diagnostic Assessment*. Unpublished Doctoral Dissertation. University of Alberta: Edmonton, Alberta, Canada.

Zoanetti, N. (2010). Interactive computer based assessment tasks: How problem-solving process data can inform instruction. *Australasian Journal of Educational Technology*, *26*(5), 585-606.

APPENDIX

Performance level descriptors for the diagnostic mathematics project.

Diagnostic Mathematics **Performance Level Descriptors**		
Limited Evidence of Mastery A student who performs at the **Limited Evidence of Mastery** level requires ongoing support, instruction, and practice to understand many concepts and processes. This student is typically characterized by the following traits:	**Moderate Evidence of Mastery** A student who performs at the **Moderate Evidence of Mastery** level demonstrates inconsistencies in his/her understanding. Mastery of most concepts is achieved through repeated practice and/or support. This student is typically characterized by the following traits:	**Consistent Evidence of Mastery** A student who performs at the **Consistent Evidence of Mastery** level has an in-depth understanding of concepts. Mastery is achieved quickly. This student is typically characterized by the following traits:
• Independently solves some familiar problems • Lacks prerequisite knowledge and requires repeated instruction of new concepts • Lacks strategies or chooses inappropriate strategies • Performs processes inefficiently or inaccurately • Demonstrates little perseverance • Unable to correctly solve the majority of questions • Unable to justify solutions • Inconsistent or limited understanding of mathematical concepts and related processes • Has difficulty and requires support to translate between concrete, pictorial, and symbolic modes • Solves some low complexity problems independently and some moderate complexity problems with support	• Independently solves familiar problems and solves unfamiliar problems with some support • Understands new concepts and begins to integrate them with previously learned concepts • Chooses appropriate strategies • Performs processes accurately, most of the time • Perseveres to successfully complete simple questions • Solves the majority of questions correctly • Justifies solutions • Understands basic concepts and related processes • Translates between concrete, pictorial, and symbolic modes inconsistently • Solves low and moderate complexity problems independently and some high complexity problems with support	• Independently and confidently solves both familiar and unfamiliar problems • Understands new concepts and integrates them with previously learned concepts • Chooses efficient strategies • Performs processes accurately and efficiently • Perseveres to successfully complete simple and complex questions • Rarely solves a question incorrectly • Justifies and explains solutions • Understands basic and complex concepts and related processes • Translates between concrete, pictorial, and symbolic modes readily and independently • Solves low, moderate, and high complexity problems independently