University of Alberta

Developing and Evaluating Score Reports for Cognitive Diagnostic Assessment

by

Mary Patrice Roduta Roberts

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

Fall 2012
Edmonton, Alberta

Dedication

This dissertation is dedicated to my daughter, Isla, and to my husband, Brodi. Thank you for being by my side as I completed this journey of discovery and learning.

Abstract

Score reporting serves a critical function as the interface between the test developer and a diverse audience of test users. The basic requirements for score reporting are clearly identified within the *Standards for Educational and Psychological Testing* (1999). However, the methods to achieve these standards are not. There lies an implicit assumption that results are reported in a useful manner to educational stakeholders to enable their use for communicating student performance, but there has been a paucity of research in this area to confirm or disconfirm this assumption. Effective reporting of diagnostic results requires a multi-disciplinary effort and input from all target audiences. In this study, a framework was created to structure an approach for developing score reports for cognitive diagnostic assessments. Guidelines for reporting and presenting diagnostic scores were based on a review of current educational test score reporting practices and literature from the area of information design. Then, core members of Alberta Education's Cognitive Diagnostic Mathematics Assessments team applied the reporting framework to create three score reporting templates in the context of a Grade 3 diagnostic mathematics assessment. The templates were then evaluated by teachers on the dimensions of: (1) content and format, (2) understanding and interpretation, and (3) uses of and preferences for information. Results of this study revealed that all three reporting templates provided the teachers with information consistent with what was expected from a diagnostic assessment. Teachers did not have difficulties understanding and interpreting information within the report. However, suggestions were made to improve visual

organization and clarity of wording. Primary uses identified for reported information include communicating learning to parents and students, informing instructional planning, evaluating student learning, and incorporating results in summative reporting. To facilitate use of results, paper-based, classroom-level reports with an accompanying website should be considered. Limitations of the study and recommendations for future research are also discussed.

Acknowledgement

This dissertation was made possible with the support of many people, to whom I owe my deepest and sincerest gratitude. I would first like to thank my supervisor, Dr. Mark Gierl. Mark has been an important figure in my student life for the past 7 years. He has been instrumental in providing me with the right challenges to foster my independence as a researcher and opportunities to grow intellectually and professionally. I have benefitted from his guidance, wisdom, and support throughout my program. Thank you, Mark. I would also like to thank the members of my doctoral committee for their encouraging words and insightful comments on my research: Dr. Ying Cui, Dr. George Buck, Dr. Cheryl Poth, Dr. Elaine Simmt, and Dr. John Willse.

My dissertation would not have been possible without the support of Alberta Education, most notably, Bev Dekker, Darrel Tripp, Renate Taylor-Majeau, and Ken Marcellus. To the teachers, who graciously volunteered their time with me in the midst of their busy workdays to participate in my study, I am most grateful.

There are others I must thank for their support during my doctoral studies. Dr. Todd Rogers has been an important mentor and teacher during my time in CRAME. I have enjoyed working with him as a student, consultant, and teaching assistant. My thanks also goes to Dr. Stephen Norris, for providing me with learning and research opportunities. I have enjoyed our conversations and work together.

I made many friends both within and outside of CRAME during my doctoral studies. In particular, thank you to Cecilia Alves, Andrea Gotzmann, and Louise Bahry for your friendship and the many enjoyable conversations that enriched my experience while in CRAME. To my family, my sisters Lara and Victoria, and to my dear friends who supported my decision to pursue graduate studies, thank you for keeping me grounded. And to my parents, Victor and Purita Roduta, for instilling in me the importance of education and for believing in me from the very beginning, that I can accomplish whatever it is that I set out to do.

To my daughter Isla, you are my constant reminder of what is good and beautiful in this world. I hope one day you will pursue your dreams whole heartedly and without hesitation. Finally to my husband Brodi, your unending love, support, dedication, and patience has truly been the stabilizing force throughout my program. I could not have done this without you. Thank you for everything.

TABLE OF CONTENTS

LIST OF TABLES xiii

LIST OF FIGURES xiv

CHAPTER 1 INTRODUCTION 1

Context for the Study 3

Purpose of the Study 3

Organization of the Document 5

CHAPTER 2 REVIEW OF THE LITERATURE 6

Section 1: A Review of Current Test Score Reporting Practices in Education and Research on Presenting Information 6

The Standards and Features of Score Reporting 6

Reviewing Current Score Reporting Practices 9

Guidelines For Effective Score Reporting 11

Designing score reports: Why look to information design? 12

Designing Effective Text-Based Documents 13

Internal text structuring. 13

External text structuring. 14

Designing Effective Displays Of Quantitative Information 17

Choosing a format for displaying information. 18

Summary of section 1. 21

Section 2: Reporting Diagnostic Scores 22

Score Reporting And Cognitive Diagnostic Assessment.....22

The Attribute Hierarchy Method.....25

Shortcomings of Reported Research.....27

CHAPTER 3 DEVELOPMENT OF MATERIALS FOR EVALUATION 29

Information Design Guidelines Applied to Score Reports 32

Context for Development of the Diagnostic Score Reports.....32

Score report 1: Cognitive diagnostic score report with tabular and graphical representation of skill mastery without interpretive material, spring 2009.....35

Score report 2: Cognitive diagnostic score report with tabular representation of skill mastery without interpretive material, spring 2010.....39

Score report 3: Cognitive diagnostic score report with tabular and graphical representation of skill mastery with interpretive material, alternate template. 43

Score reporting templates for evaluation with teachers. 47

Summary of Chapter 3 50

CHAPTER 4 METHOD 51

Sampling and Participants 52

Inclusion and Exclusion Criteria 54

Instruments 54

Questionnaire to Evaluate the Diagnostic Score Reports 55

Interview guide to evaluate the score reports 57

Procedure 59

Overview 59

Stage 1: Administrating the questionnaire. 60

Stage 2: Conducting follow-up interviews. 60

Stage 3: Data analysis and revision of score report. 61

CHAPTER 5 RESULTS 64

Participant Recruitment and Characteristics 64

Evaluating Particpants' Understanding and Interpretation 68

Teacher Sources of Diagnostic Information 68

Teacher Perceptions of Diagnostic Assessment Characteristics 71

Diagnostic assessment should provide specific and descriptive information about student performance. 71

Diagnostic assessment is one of a battery of assessments that can be used to assess and evaluate a student's performance. 72

Diagnostic assessment should provide information to the teacher on instructional effectiveness and guidance. 73

Diagnostic assessment should be used with students who are having difficulties. 73

Evaluation of the Score Reports 73

Score report 2. 74

Score report 1. 78

Score report 3. 81

Summary of Evaluative Comments Across Templates 85

Uses of and Preferences for Diagnostic Information 88

Uses of Information 89

Score reports as a communication tool. 89

Informing the next instructional steps. 91

Evaluatimg the student's level of learning. 91

Using the results as part of summative reporting. 92

Potential for misuse of information? 93

Information Preferences for Diagnostic Score Reporting 93

Aggregate level reporting.....94

Mode of dissemination.....94

Which template was preferred?.....95

Refining the Reporting Framework 96

CHAPTER 6 DISCUSSION AND CONCLUSION 98

Restatement of Research Questions and Summary of Methods 100

Summary of Results.....103

Research Question 1: What should be reported on a diagnostic score report? How should this information be presented?.....103

Research Question 2: To what extent are the score reports understandable and interpretable by teachers?.....106

Research Question 3: Given the answers to questions 1 and 2, how useful are the score reports? How could the information in the score report be used? Is it clear what the next instructional steps should be for the student? 109

Limitations 110

Future Directions 111

Conclusion 115

REFERENCES 116

APPENDIX A: Participant research information sheet 125

APPENDIX B: Teacher questionnaire.....127

APPENDIX C: Sample interview protocol.....136

APPENDIX D: Developing code for thematic analysis……..….....141

LIST OF TABLES

Table for the 1990 8th and 12th grade NAEP science assessment......20

Alignment of AHM elements and outcomes to a general reporting framework.....31

Example attribute hierarchy and skill descriptors from the strand of Number.....34

Key feature differences between score report templates……..……..…49

Summary statistics for select demographic information (n=21) 65

Sources of teacher knowledge of educational assessment (n=21).....67

Teachers' ratings of information to inform diagnostic decisions about students (n=21, % agreement by response option)..... 69

Other identified sources of information to inform diagnostic decisions about students (n=21) 70

Teachers' classroom activities informed by diagnostic information (n=21)..... 71

Evaluation of format and content: Score report 2 (% agreement by response option) 75

Evaluation of format and content: Score report 1 (n=7, % agreement by response option) 80

Evaluation of format and content: Scorer report 3 (n=7, % agreement by response option).....83

Comparison of item means (/5) evaluating content and format across the three reporting templates 86

LIST OF FIGURES

Figure 1. Example of a formatted graph using the design principles of contrast and proximity. 21

Figure 2. Diagnostic score report 1 with graphical representation of skill mastery without interpretive material created for spring 2009. 38

Figure 3. Diagnostic score report 2 with tabular representation of skill mastery without interpretive material created for spring 2010. 42

Figure 4. Diagnostic score report 3 with graphical representation of skill mastery with interpretive material, alternate reporting template. 46

Figure 5. Interpretive material for score report 3. 47

Figure 6. Overview of presentation of results. 66

CHAPTER 1 INTRODUCTION

Educational tests should provide meaningful information to guide student learning. The recent emphasis on understanding the psychology underlying test performance has led to developments in cognitive diagnostic assessment (e.g. Leighton & Gierl, 2007a; Mislevy, 2006), which integrates cognitive psychology and educational measurement for the purposes of enhancing learning and instruction. A cognitive diagnostic assessment (CDA) is specifically designed to measure a student's knowledge structures and processing skills. In contrast with reporting a small number of content-based subscores, typical of most current educational test score reports, the results of a CDA yield a profile of scores with specific information about a student's cognitive strengths and weaknesses. This cognitive diagnostic feedback has the potential to guide instructors, parents, and students in their teaching and learning processes. 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. However, the question of how to effectively communicate such complex and detailed information on educational tests, in general, or CDA, more specifically, has been inadequately studied, to date.

Score reporting serves a critical function as the interface between the test developer and a diverse audience of test users. Despite the importance of score reports in the testing process, there has been a paucity of research in this area. The available body of research on test score reporting has centered on large-scale

reporting of aggregate-level results (i.e., at district, state, and national levels) for accountability purposes in the United States (Jaeger, 1998; Linn & Dunbar, 1992). Fewer studies have focused on student-level score reporting features (Goodman & Hambleton, 2004; Trout & Hyde, 2006). General conclusions drawn from these studies are not encouraging claiming that score reports are difficult to read and understand (Hambleton & Slater, 1997), often lead to inferences not supported by the information presented (Koretz & Diebert, 1993), and are not disseminated in a timely manner (Huff & Goodman, 2007).

As developments in CDA continue to progress, the need to address and overcome score reporting issues of comprehensibility, interpretability, and timeliness become even more urgent. Diagnostic testing information, including skills descriptions and learning concepts, is fundamentally different in purpose from information typically reported from traditional large-scale assessments, such as total number correct scores or percentile ranks. Test developers must report and present new kinds of information from these diagnostic tests. In short, the challenge of diagnostic score reporting lies in the integration and balance of the substantive and technical information needs of the educational community with the psychologically sophisticated information unique to CDA. But how can test developers present diagnostic information to a non-technical audience in a way that can be understood? How can test developers evaluate the effectiveness of their reports in presenting diagnostic information? To date, no such research on diagnostic score reporting exists to answer this question. Thus, to begin to address this gap in the literature, a framework for reporting diagnostic information is

needed. Following this, evaluation of the reports created from the framework is also needed to ensure that the benefits of a CDA are realized with their intended audience.

Context for the Study

The context for this study is the Cognitive Diagnostic Mathematics Assessment (CDMA) project funded by the Learner Assessment Branch at Alberta Education in Edmonton, Alberta, Canada. More specifically, the assessments developed for Grade 3 were used as the basis for developing and evaluating the score reports. CDMA is a curriculum-based set of assessments that can be used throughout the school year to measure students' strengths and weaknesses in thinking and learning mathematics. The information from CDMA can be used to provide valuable instructional guidance, which may be needed to design remedial instruction programs or supplemental interventions for students.

Purpose of the Study

The purposes of this study was to create a framework for promoting a structured approach to developing diagnostic score reports and to evaluate the effectiveness of the score reports developed using this framework for communicating diagnostic results for an operational cognitive diagnostic assessment. More specifically, my research questions were as follows:

1. What should be reported on a diagnostic score report? and How should this information be presented?
2. To what extent are the score reports understandable and interpretable by teachers, the target audience? What are teachers' understanding of the

diagnostic information in the score report? What information do teachers consider to be diagnostic and why?

3. Given the answers to questions 1 and 2, how useful are the score reports? Is it clear what the next instructional steps should be for the student?

To answer research question 1, a literature review was conducted on (a) current test score reporting practices to provide a context for diagnostic score reporting and (b) relevant literature pertinent to presenting information in score reports. Then, a structured approach for reporting diagnostic scores based on the literature review was created in the form of a diagnostic score reporting framework. Application of the reporting framework was illustrated in the context of one CDA procedure called the attribute hierarchy method (AHM; Gierl, Wang, & Zhou, 2008; Leighton, Gierl, & Hunka, 2004) to generate sample diagnostic reports. These sample diagnostic reports was reviewed by the test development team and then evaluated by teachers using an online questionnaire and semi-structured interview.

To answer research question 2, an online questionnaire was used to collect teachers' opinions on the content and format of the score report. Teachers were asked about what types of educational test results help them to make diagnostic decisions about their students. Questions in a semi-structured interview were used to investigate teachers' understanding of the information in the score report. During the interview, the reasons as to why a teacher considers one type of information to be diagnostic compared to another was explored.

To answer research question 3, questions in the semi-structured interview were used to collect opinions on the usefulness of the information of two different reporting schemes for informing classroom instruction and planning student remedial work.

Organization of the Document

This document is organized into six chapters. Chapter One provides an introduction to the study, the context of the study, the research questions and proposed methods for answering these questions. Chapter Two reviews the literature in educational measurement and information design that inform the development of the diagnostic score reporting framework. Additionally, a brief description of the AHM is provided as it is the context in which the diagnostic score reports were developed. Chapter Three introduces the diagnostic score reporting framework and presents three alternate reporting templates created using this framework. Chapter Four describes the methods that were used in this study. This chapter includes a description of the development of the questionnaire and interview protocol, sampling and participants, data collection, and data analysis. Chapter Five presents the results of the evaluations of the score reporting templates. Finally, in Chapter Six, a summary of the study, limitations, and directions for future study are provided.

CHAPTER 2 REVIEW OF THE LITERATURE

Chapter 2 is organized into two sections.[1] In the first section, score reporting research in the educational measurement literature is summarized. Also, relevant information design literature for presenting score report information is reviewed. In the second section, a review of reporting diagnostic scores is provided and the AHM as a form of cognitive diagnostic assessment is introduced. Diagnostic information characteristics specific to the AHM that are relevant to score reporting are identified and described. Key principles identified from the literature review are synthesized and summarized into a framework for reporting AHM diagnostic results. The diagnostic score reporting framework and example score reporting templates created from this framework will be presented in Chapter 3.

Section 1: A Review of Current Test Score Reporting Practices in Education and Research on Presenting Information

The Standards and Features of Score Reporting

Legislated and professional standards for test score reporting function, in part, to ensure some standardization of the information reported to educational stakeholders about student performance. These standards were created largely in the context of large-scale assessments. With No Child Left Behind Act of 2001 (NCLB), test developers must develop ways to present student-level results in mandated statewide assessments. NCLB requires states to:

[1] A version of this chapter has been published. Roberts, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice, 29 (3), 25-38.*

> Produce individual student interpretive, descriptive, and diagnostic reports...that allow parents, teachers, and principals to understand and address the specific academic needs of students, and include information regarding achievement on academic assessments aligned with State academic achievement standards, and that are provided to parents, teachers, and principals, as soon as practicably possible after the assessment is given, in an understandable and uniform format, and to the extent practicable, in a language that parents can understand. (NCLB, 2001, as cited in Goodman & Hambleton, 2004, p. 147)

In addition to meeting high psychometric standards, the information provided by large-scale assessments must also meet professional standards. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999) contains numerous guidelines relevant to score reporting. The role of test developers in the reporting process is exemplified within Standard 5.10:

> When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what the scores mean, and how the scores will be used. (p. 65)

The basic requirements for score reporting are clearly identified within these standards. However, the methods to achieve these standards are not. There

lies an implicit assumption that results are reported in a manner that can be readily understood and used by educational stakeholders. A structured approach to the test score reporting process is needed to ensure that relevant score reporting features are identified and reported.

Jaeger (1998) proposed a comprehensive research agenda for reporting results from the National Assessment of Educational Progress (NAEP) testing program. Jaeger proposed three questions that, when answered, should help to guide the score reporting process. First, *in what form should NAEP results be reported?* Form in this context refers to the method of summarizing student performance which includes the use of performance descriptors, obtained through item mapping, scale anchoring or achievement levels, and scale scores. Second, *how should NAEP results be displayed?* Displays of information include numeric, graphic, and narrative forms. Third, *how should results be disseminated?* For example, score reports can be paper based or web based; a standalone document or with an accompanying interpretative guide. Each of these three questions should be applied in the context of a specific audience. His report provided an important initial attempt at providing a structure for reporting NAEP results, which could be generalized to other research efforts on student-level score reporting.

Jaeger's framework for reporting NAEP results can be further refined by specifying reporting elements which are common to most reporting systems. Ryan (2003) provides a useful framework of eight reporting features or characteristics. These characteristics include: (1) audience for the report, (2) scale or metric for

reporting, (3) reference for interpretation, (4) assessment unit, (5) reporting unit, (6) error of measurement, (7) mode of presentation, and (8) reporting medium.

Reviewing Current Score Reporting Practices

Goodman and Hambleton (2004) provide the most recent comprehensive review and critique of student-level score reporting practices from large-scale assessments. Their review showed varied practices with the kinds of information reported and their presentation. In general, the type and number of overall scores and content-based subscores reported varied across testing programs and contexts. Usually, two types of overall scores were reported such as scale scores, percentile ranks, stanines, and number correct scores. Goodman and Hambleton conclude that scale scores are the most popular method of reporting as they are ideal for the purposes of comparing sets of scores across different groups of students and different test administrations. Previous studies indicate interpretation of scale scores is difficult for a number of audiences (Forsyth, 1991; Koretz & Diebert, 1993). This difficulty is best illustrated by the body of research conducted on the reporting of NAEP results (see special issue in the *Journal of Educational Statistics*, Spring 1992) which generated a line of research focusing on IRT-based item mapping methods (Lissitz & Bourque, 1995; Zwick, Sentur, Wang, & Loomis, 2001) to improve the substantive meaning behind scale scores in an attempt to increase the interpretability of the score report.

Goodman and Hambleton's (2004) review concluded that many of the student score reports had promising features such as reporting information in alternate forms (i.e., narrative, numeric, and graphic), having different reports for

different audiences, and personalizing reports and interpretative guides. However, they cite a number of weaknesses that require further attention and research including:

1. Reporting excessive amounts of information, such as many types of overall scores, but omitting essential pieces of information, such as the purpose of the test and information about how test results will and should be used.
2. Information regarding the precision of test scores was not provided.
3. The use of statistical jargon.
4. Key terms were not always defined in the reports or interpretative guides, leaving interpretations up to users and inviting inaccurate interpretations.
5. Efforts to report large amounts of information in such a small amount of space resulted in reports that appeared dense, cluttered, and difficult to read.

These weaknesses are echoed in the results of studies on score reporting conducted over the past 15 years which have consistently identified issues with the reporting of large-scale assessment results (Hambleton & Slater, 1997; Impara, Divine, Bruce, Liverman, & Gay, 1991; Koretz & Diebert, 1993). In general, these studies concluded that accurate interpretation of score reports were influenced by multiple factors, including familiarity of the reader with statistical, measurement, and assessment concepts, presentation of the results, and availability of information to support the reader in making appropriate

interpretations and inferences. More recently, criticism around the timeliness of reporting comes from Huff and Goodman (2007) and Trout and Hyde (2006). These researchers cited the time lag present between assessment and reporting of the results as a limitation for using assessment results to inform classroom instruction and learning. Turnaround time for score reports has become an issue that could potentially be resolved using technology (i.e., web-based reporting).

Guidelines For Effective Score Reporting

Numerous guidelines for effective score reporting have emerged in the educational measurement literature. Aschbacher and Herman (1991) reviewed relevant empirical literature from the disciplines of psychology, communication, and business for their set of reporting guidelines. Forte Fast and the Accountability Systems and Reporting State Collaborative on Assessment and Student Standards (2002) also created a set of reporting guidelines with a greater emphasis on the use of *universal design principles*. Universal design refers to the "design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design" (Center for Universal Design, n.d.). These studies suggest that reporting guidelines should incorporate design principles that, when implemented, yield score reports that are accessible to a majority of educational stakeholders. The similarity of the guidelines identified by Aschbacher and Herman, Forte Fast et. al., and Goodman and Hambleton, demonstrate a general agreement about how they believe this information should be presented. Goodman and Hambleton (2004) provide

specific recommendations for reporting student-level results (with slight modifications):

1. Include all information essential to proper interpretation of assessment results in student score reports;
2. Include detailed information about the assessment and score results in a separate interpretative guide;
3. Personalize the score report and interpretative guide;
4. Include a narrative summary of the student's results at the beginning of the score report;
5. Identify some things parents can do to help their child improve;
6. Include sample questions in the interpretative guides; and
7. For paper copies, include a reproduction of student score reports in the interpretative guides to explain elements of the score reports.

Concrete examples of score reports implementing these guidelines are few and varied across states and testing programs (Goodman & Hambleton, 2004). More importantly, the effectiveness of these guidelines requires validation through empirical studies in the context of operational score reporting.

Designing score reports: Why look to information design?

Rune Pettersson (2002) defines information design as the following:

> In order to satisfy the information needs of the intended receivers, information design comprises analysis, planning, presentation and understanding of a message – its content, language and form. Regardless of the selected medium, a well-designed information set will satisfy

> aesthetic, economic, ergonomic, as well as subject matter requirements. Information design is a multidisciplinary field where the goal of communication-oriented design is *clarity of communication*. (Pettersson, 2002, p. ix)

The field of information design embodies research from different disciplines, including psychology, communication studies, information technology, and aesthetics, with the focus on communicating information effectively. This research includes design guidelines applicable to developing user-friendly score report documents and effective displays of quantitative information, such as tables and graphs. Hence, information design principles may help us overcome some of the persistent problems that arise in assessment score reports.

Designing Effective Text-Based Documents

A number of design techniques are available to assist a reader when reading a document. More specifically, document elements can be structured to provide an organizing framework, or a *schema* in psychological terms, to present information in a coherent and logical manner. These techniques are grouped into two broad categories of internal and external text structuring (Jonassen, 1982; Pettersson, 2002). Both techniques complement and interact with each other when used to create documents that effectively communicate information.

Internal text structuring.

Internal text structuring includes techniques to organize, sequence, and provide an internal framework for understanding document content. A familiar

example of internal text structuring includes the use of paragraphs to chunk text that contain one main idea. Well written paragraphs begin with a topic sentence to signal the reader to a new concept and end with a concluding sentence to signal an end to the current topic and a possible shift in content. The use of logical connectives such as but, if, and therefore, assist with the logical flow of paragraphs. Most readers come to expect some imposed structure and organization of ideas. When this is not found, the text can be considered difficult to read and understand. This problem is more persistent with longer texts with highly technical or scientific reporting. In these instances, the use of external text structuring in combination with internal structuring techniques can be applied to assist the reader in organizing and comprehending information.

External text structuring.

Familiar examples of external text structuring techniques, such as headings and bolded font, can be found in a range of materials from textbooks to academic papers. External text structuring includes techniques such as the use of access structures, typographical cues, and spatial layout to structure text (Gribbons, 2002; Waller, 1983).

Access structures. Access structures (Waller, 1983) combine linguistic cues with typographic and spatial cues to help the reader gain access to text that is meaningfully grouped and sequenced. Waller describes access structures as having two text functions: global and local accessibility. Global accessibility can provide an overview of the content presented and assists the reader with developing a reading strategy (i.e., to search for and read specific parts of the text,

or to read the text entirely). Examples include: (a) table of contents, (b) glossary, (c) objectives, and (d) summaries. Summaries and glossaries provided after the presentation of text serve as a review aid, highlighting and reinforcing the main ideas, text organization, and sequencing.

Alternatively, local accessibility refers to techniques that signal or identify particular units of text, often providing a visual structure. Examples include: (a) headings, (b) numbering systems, and (c) lists. Depending on the purpose of the text, headings can be written in the form of a statement or in the form of a question. Swarts, Flower, and Hayes (1980) recommended that headings be posed in the form a question, signalling that the following text will answer a question the reader would like answered. Headings should be accurate, specific, and concise (Hartley & Jonassen, 1985).

Numbering systems are also often used in list structures. Lists are characterized by a string of main elements all of which contain a number of subelements (Hartley, 1987). The task in developing effective lists requires consideration not just of the wording of the elements and subelements, but also of the formatting or spatial arrangement of the list. This is an example of how local access structures serve both an identification function while also serving a visual cueing function.

Typography. Typography deals with the aspects of type which can be a letter, number, or any other character used in printing (Pettersson, 2002). Legibility is an important consideration in choosing a particular font or typeface for readable text. The point size of typeface used has important implications for

the length of written text. For example, long sentences written in very small typeface are difficult to read, whereas the use of very large typeface means fewer words per line. Typographical cues such as *italics*, **bolding**, CAPITAL LETTERS, underlining, and color, serve as an explicit visual cue or signalling device within text. These cues draw the reader's attention to important words or sections of text and can also support the spatial organization of the document (Gribbons, 1992). The judicious use of one or two typographical cues is warranted, as multiple or "over-cueing" may serve to confuse the reader and unnecessarily clutter a document. Horton (1991) and Winn (1991) argue that color can be used as a typographical cue within an organized text by maximizing type contrast and background to improve legibility. Consideration should also be given to certain color choices, given that approximately 10 percent of the North American population has difficulty distinguishing colors, including red-green and blue-yellow defects (Vaiana & McGlynn, 2002). Factors such as resources to print documents in color, as well as characteristics of the audience (i.e., color-deficits) will influence whether elaborate color schemes are used on score reports.

Document layout and organization. A combination of type, color, and spatial organization techniques can effectively structure text within a document. The use of vertical and horizontal spacing assists with reinforcing the visual hierarchical structuring characteristic of most documents. Gribbons (1992) claims the designation of vertical and horizontal cues is guided by three factors. First, horizontal positioning should accommodate the significance a reader places on information in the left-most portion of a page. Second, vertical positioning with a

common alignment should use the principle of proximity to group conceptually similar items. Third, spatial formatting should be consistent with the structure previously established using other techniques such as local access structures and typographical cues.

Designing Effective Displays Of Quantitative Information

Score reports necessitate communication of quantitative information such as test scores, percentile ranks, and error of measurement. This kind of information can be summarized narratively, or visually using a table or graph. Numerous theories of graphical perception and cognition currently exist in the literature (Bertin, 1983; Cleveland, 1994; Cleveland & McGill, 1985; Kosslyn, 1994; Wainer, Hambleton, & Meara, 1999). Two of these major theories by Cleveland and McGill (1985) and Kosslyn (1994), draw on knowledge of the brain for elucidating their theories of graphical perception and cognition. Cleveland and McGill's theory focuses on the manipulation of the perceptual features of graphs which affect the reader's associated cognitive processes of selecting and encoding. Kosslyn incorporates the perceptual theories of Cleveland and McGill, but also focuses on the cognitive processes invoked once sensory information is attended to and held in working memory.

The design principles for designing effective quantitative data displays are similar to those of designing effective text: (1) using contrast to signal important information and increase legibility, (2) using redundancy of visual cues for emphasis of presented information, such as using a combination of large typesize and color for headers, (3) using proximity to group similar elements together, and

(4) using a common alignment of elements to emphasize visual structuring of information.

Choosing a format for displaying information.

In her summary of tabular versus graphical displays, Wright (1977) aptly states that generalizations of research findings on the superiority of one format over another are difficult. The decision of whether to use one format over another requires individual consideration of the particulars to a situation including the purpose of the data display and characteristics of the intended audience.

Tables. Tufte (2001) recommends the use of tables for small data sets showing exact numerical values requiring local comparisons. When creating a table, Tufte discusses the use of vertical and horizontal formatting techniques to both structure and group numerical entries. Some evidence for this claim is provided by the documented difficulties of administrators and educators attempting to make sense of the large summary tables used in reporting NAEP results (Hambleton & Slater, 1997). Wainer (1997) drawing upon the work of Ehrenberg (1977) lists some general principles for improving tabular formats. These include:

1. Rounding digits to no more than 2 decimal places;
2. Using row or column averages to provide a visual focus and a summary;
3. Using columns rather than rows to make intended comparisons;
4. Ordering the rows and columns in meaningful ways; and
5. Using white space to group figures and to guide the eye.

The physical display of a table should be aesthetically pleasing and without excessive clutter. Tufte advocates the use of thin lines within the table, and for aesthetic considerations, varying the thicknesses of linework where applicable.

Table 1, taken from Wainer (1997), is an example of a formatted table following design principles. This table was used as an adjunct to a question from the 1990 8^{th} and 12^{th} grade NAEP science assessment. The question defined the purpose of the display: “On the basis of the information in the table, which brand do you think is the best all-purpose battery? (Assume all batteries cost the same)”. To answer this question, comparisons among the values within the table are necessary. The table presents marginal values in whole numbers and bolded to provide a visual cue to the reader allowing for quicker identification. Additionally, the rows and columns are organized in descending battery and usage averages while a visual space between the third and fourth rows signals another important organization of information: batteries with battery averages greater than 10 hours or averages equal to or less than 10 hours. The inclusion of summary values in the rows and the columns facilitates this process reducing the amount of cognitive processing required by the reader. Said another way, the reader is required to make fewer inferences from the information presented during the interpretative process.

Table 1.

Table for the 1990 8th and 12th grade NAEP science assessment.

	Battery Life in Hours				
Battery Brands	Radio	Flashlight	Cassette Player	Portable Computer	**Battery Averages**
Never Die	28	16	8	6	**15**
Electro-Blaster	26	15	10	4	**14**
PowerBat	24	13	7	5	**12**
ServoCell	21	12	4	2	**10**
Constant Charge	19	10	5	3	**9**
Usage averages	**24**	**13**	**7**	**4**	**12**

Graphs. The use of graphs over tables is preferable for readers if comparisons of the data are to be made (Shah, Mayer, and Hegarty, 1999; Wright, 1977). Graphs communicate amounts, changes, and trends in the data more accurately and can be perceived more readily. When constructed appropriately, graphical representation can reduce the cognitive load required by the reader to make accurate comparisons, inferences, and interpretations. The graphic format should be compatible with its form (Kosslyn, 1994) and its intended purpose. For instance, bar graphs are best used for static comparisons, more so than pie charts or three-dimensional figures, whereas line graphs are best used to illustrate trends. Graphics and text should be integrated in the document and not placed on separate pages, especially if the graph is meant to illustrate points discussed in the text. Labels for axes and other graphical elements should be positioned close to its referent to promote easy and accurate interpretation of information (Macdonald-Ross, 1977 as cited in Schriver, 1997).

A graph following design principles is presented in Figure 1, where the data values have been grouped for each subject area and each bar has been coded a different texture to represent males and females. Mean total scores are placed above each bar for easy reference. Labels are more specific in their description and are written in both upper and lower case.

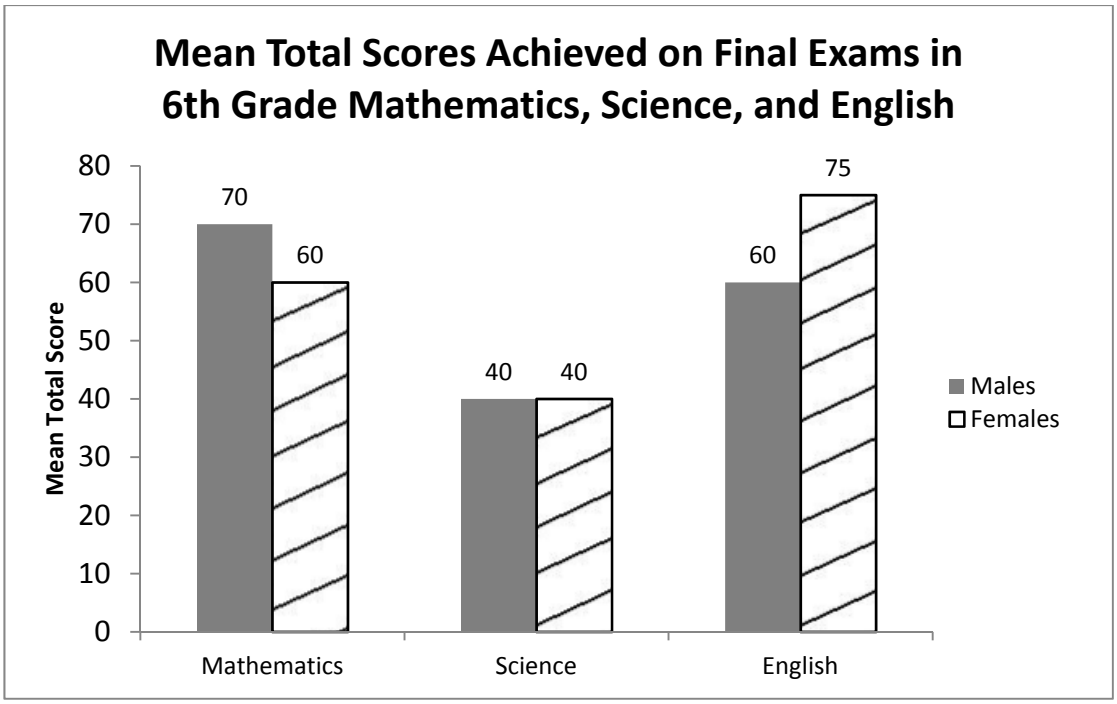


Figure 1. Example of a formatted graph using the design principles of contrast and proximity.

Summary of section 1.

Most available research on score reporting has been conducted with NAEP data or results from statewide assessments (Hambleton & Slater, 1997; Jaeger, 1998, Koretz & Diebert, 1993). These data are often reported at the aggregate level, with results often used for accountability purposes. Difficulties with reading, understanding, and interpreting score information accurately have led to strategies for creating substantive meaning for reported score information in the

form of IRT-based item mapping strategies (Beaton & Allen, 1992; Lissitz & Bourque, 1995; Zwick et. al., 2001). Sets of guidelines exist for score reporting, recommending the use of relevant design principles for improving the appearance of score reports (Aschbacher & Herman, 1991; Forte Fast et. al., 2002; Goodman & Hambleton, 2004). In particular, research reviewed from information design promotes the creation of an organizing structure or framework to help assist the reader with creating a coherent representation of the information presented, aiding in accurate comprehension and interpretation. Also, research by Trout and Hyde (2006) and Huff and Goodman (2007) identified the need to report test score information, especially for teachers, in a timelier manner for use in planning instructional activities. The use of companion websites or dissemination of results using the Internet appears to be a promising suggestion. Next, I discuss reporting diagnostic results in light of the review of current test score reporting practices, recommendations for reporting information, and designing score reports.

Section 2: Reporting Diagnostic Scores

To begin, a brief overview and rationale for the development of cognitive diagnostic assessments with implications for score reporting is provided. Then, one CDA method called the Attribute Hierarchy Method is presented as the context within which the reporting framework will be applied.

Score Reporting And Cognitive Diagnostic Assessment

Research efforts in cognitive diagnostic assessment (CDA) have been fuelled by the increasing demand, from both researchers and educational stakeholders, for more formative information from educational tests (Huff &

Goodman, 2007). Scores provided from large-scale assessments provide minimal information about a student's performance that can be used to support classroom activities. This is largely due to the dominant testing paradigm in an accountability framework with the development of large-scale educational tests that function to assess and rank order examinees based on a unidimensional latent trait. Given this focus, it is logical that large-scale assessments report only one overall total score. Diagnostic scores are often conceived as content-based subscores, which are reported from assessments originally designed to measure a unidimensional latent trait. As previously discussed, interpretation of these scores is often difficult, the scores are open to misinterpretation, and the scores usually require some context in the form of anchor items, achievement, or performance descriptors for understanding what the reported test score means in terms of student performance.

The unidimensional testing paradigm is now making way for assessments designed to model and assess multiple cognitive skills that underlie student test performance (Stout, 2002). CDA has generated a surge of scholarly interest and activity among educational measurement researchers. As testimony to this claim, the *Journal of Educational Measurement* dedicated a special issue in 2007 to IRT-based cognitive diagnostic models and related methods. Many diverse cognitive diagnostic psychometric models (CDMs) and procedures currently exist for skills diagnostic testing including the Multicomponent Latent Trait Model (Whitely (Embretson), 1980), Bayes Net (Mislevy, Almond, Yan, & Steinberg, 1999; Mislevy, Steinberg, & Almond, 2003), Rule Space Model (Tatsuoka, 1983, 1990,

1995), Unified Model (DiBello, Stout, & Roussos, 1995; Hartz, 2002), deterministic input noisy and gate model (DINA; de la Torre & Douglas, 2004; Haertel, 1989), noisy input deterministic and gate model (NIDA; Junker & Sijtsma, 2001), deterministic input noisy or gate model (DINO; Templin & Henson, 2006), Hierarchical General Diagnostic Model (von Davier, 2007), and the Attribute Hierarchy Method (Gierl, Wang, & Zhou, 2008; Leighton, Gierl, & Hunka, 2004). These models all share a common feature where the results of a complex analysis yield a profile of scores based on the cognitive skills measured by the test. In contrast to reporting one overall scaled score or multiple content-based subscores, as in large-scale assessments, cognitive diagnostic assessments produce many scores often in the form of *skill mastery probabilities*. These skill mastery probabilities serve as *scores* that are substantively meaningful because the interpretations and inferences about student performance are made with reference to the cognitive skills measured by the test. These diagnostic skill profiles can then be used to support instruction and learning.

Currently, there are few examples of cognitive diagnostic score reports. One operational example is the College Board's *Score Report Plus* for the PSAT/NMSQT which reports diagnostic information based on an analysis of examinee responses using a modified Rule Space Model. Cognitive diagnostic feedback is given in the form of the top three skills requiring improvement for each content area of Mathematics, Critical Reading, and Writing along with recommended remedial activities. Jang (2009) also created score reports as part of a study investigating the application of the Fusion Model to a large-scale reading

comprehension test for cognitive diagnosis. Jang used a reporting format similar to the *Score Report Plus*. Cognitive diagnostic feedback was provided in the form of skill descriptors, discriminatory power of items, and skill mastery probabilities. As developments continue to progress, current score reporting approaches need to be recast in light of the new kinds of information yielded by CDA.

The Attribute Hierarchy Method.

The reader is provided with a brief overview of the AHM as it will provide a context for illustrating the proposed diagnostic reporting framework. While the AHM is described in this document, CDA features such as identification and representation of cognitive skills, assessment of model-data fit, and estimation of skill mastery probabilities are not unique to the AHM, but are common across all attribute-based diagnostic testing methods. Therefore, using the AHM as an illustrative example for diagnostic score reporting will generalize to many other CDMs currently available.

The AHM is a cognitively-based psychometric method used to classify an examinee's test item responses into a set of structured attribute patterns associated with a cognitive model of task performance. An attribute represents the declarative or procedural knowledge needed to solve a task in the domain of interest. These attributes form a hierarchy that defines the ordering of cognitive skills required to solve test items. The attribute hierarchy functions as a cognitive model which in educational measurement refers to a "simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills students at different levels of learning have

acquired and to facilitate the explanation and prediction of students' performance" (Leighton & Gierl, 2007b, p. 6). The attributes are specified at a small grain size in order to generate specific diagnostic inferences underlying test performance.

Development of the cognitive model is important for two reasons. First, a cognitive model provides the interpretative framework for linking test score interpretations to cognitive skills. The test developer is in a better position to make defensible claims about student knowledge, skills, and processes that account for test performance. Second, a cognitive model provides a link between cognitive and learning psychology with instruction. Based on an examinee's observed response pattern, detailed feedback about an examinee's cognitive strengths and weaknesses can be provided through a score report. This diagnostic information can then be used to inform instruction tailored to the examinee, with the goals of improving or remediating specific cognitive skills.

Once the attributes within a hierarchical cognitive model are specified and validated, items can be created to measure each combination of attributes specified in the model. In this way, each component of the cognitive model can be evaluated systematically. If the examinee's attribute pattern contains the attributes required by the item, then the examinee is expected to answer the item correctly. However, if the examinee's attribute pattern is missing one or more of the cognitive attributes required by the item, then the examinee is not expected to answer the item correctly.

After verifying the accuracy of the cognitive model for accounting observed student response data through model-data fit analyses, attribute

probabilities are estimated for each examinee. These probabilities serve as diagnostic scores. Mastery of specific cognitive skills is determined using a neural network approach (Gierl, Cui, & Hunka, 2007; Gierl, Cui, & Hunka, in press) where higher probabilities can be interpreted as higher levels of mastery. Based on a student's observed response pattern, an attribute probability close to 1 would indicate that the examinee has likely mastered the cognitive attribute, whereas a probability close to 0 would indicate that the examinee has likely not mastered the cognitive attribute (for an example, see Gierl, Wang, & Zhou, 2008).

Shortcomings of Reported Research

Given the recent advances in cognitive diagnostic assessment (CDA), the issues related to score reporting requires attention. In particular, CDA reports different kinds of information that what is typically reported on large-scale assessments. For example, in the context of the AHM, attribute probabilities are considered to be diagnostic scores. Reporting these new types of information can pose difficulties in terms of how best to report and display such information as well as how to properly interpret and use the information provided.

To date, the major focus in the literature on score reporting is critiquing current score reporting practices post hoc. There are very few examples of score reports that are created beginning with the target audience's preferences and needs for information. In the context of CDA, this approach to developing score reports may not be the most effective because it is a relatively new assessment paradigm. That is, a teacher may not be in the best position to make recommendations on what information should be reported and how it should be presented because he or

she is not yet familiar with what CDA is. Establishing a structured approach to the test score reporting process is needed to ensure that relevant score reporting features are identified and reported. Then application of the reporting framework should be evaluated in a particular operational testing context.

The AHM is a form of cognitive diagnostic assessment designed to identify and pinpoint an examinee's areas of cognitive strengths and weaknesses in order to guide efforts to improve academic performance. This diagnostic information must be communicated through score reports in an accessible manner to a diverse audience such as students, parents, and instructors. Two important questions arise. First, *what* parts of an AHM analysis should be reported? Second, *how* should this information be presented in a score report? To answer these questions, a framework for reporting diagnostic scores that becomes the starting point for creating score reports is introduced and described in Chapter 3.

CHAPTER 3 DEVELOPMENT OF MATERIALS FOR EVALUATION

Chapter 3 presents the diagnostic reporting framework developed from a review of the current test score reporting practices, recommendations for reporting information, and designing score reports. Two diagnostic reporting templates developed in conjunction with Alberta Education are described first. Then, a third alternate template developed incorporating recommendations put forth in both the educational measurement and design literature is described next. A summary of the design and reporting considerations for each reporting scheme is provided.

The AHM yields diagnostic scores that must be communicated through score reports in an accessible manner to a diverse audience such as students, parents, and instructors. Two important questions arise. *What* parts of an AHM analysis should be reported? *How* should this information be presented in a score report? To help answer this question, an adapted reporting framework based on research by Jaeger (1998) and Ryan (2003) is proposed for reporting cognitive diagnostic scores. An example of the diagnostic reporting framework applied to elements of an AHM analysis is provided in Table 2. Inspection of the framework shows that elements and outcomes of a diagnostic analysis can be systematically identified and presented in different ways and combinations. Test developers may choose to report some or all of the content outlined in the framework in various formats and modes, however the final form will likely be influenced by the information needs of a particular audience and educational policy. Additionally, implementation of information design principles including contrast, repetition, proximity, and alignment should be applied when organizing and presenting

numerical, graphical, or text-based information on a document. The reporting framework combines both content and form considerations with design principles for presenting information as a principled approach to developing diagnostic score reports.

Table 2.

Alignment of AHM elements and outcomes to a general reporting framework.

Reporting Characteristic	**AHM Analysis Element or Outcome**
Form of Reporting Results	
Scale	Attribute probabilities, total correct
Reference for interpretation	Cognitive model, criterion-referenced
Assessment Unit	Attribute level, cognitive model level
Reporting unit	Students, parents, teachers
Error of measurement	Attribute reliability
Mode of Presenting Results	
Numerical	Attribute probabilities and reliabilities
Graphical	Attribute probabilities, classification of skill mastery, attribute reliability, cognitive model
Narrative	Attribute probabilities, classification of skill mastery, summary performance descriptions, cognitive model
Medium for Dissemination of Results	Print score reports Web-based (static or interactive) score reports
Application of Design Principles	Use of contrast, redundancy, proximity, and alignment to structure text and design quantitative displays

Information Design Guidelines Applied to Score Reports

The information design literature outlines 4 general principles for effective design. These include the use of contrast, repetition, proximity, and alignment to effectively signal differences and similarities when presenting numerical, graphical, or text-based information. In the context of student-level score reporting, Goodman and Hambleton (2004) offer the following recommendations:

1. Include all information essential to proper interpretation of assessment results in student score reports;
2. Personalize the score report;
3. Include a narrative summary of the student's results at the beginning of the score report; and
4. Identify some things parents can do to help their child improve.

As stated previously, these guidelines and recommendations should be applied with an audience in mind, and that one technique may work better in some situations, but not all.

Context for Development of the Diagnostic Student Score Reports

The reports were developed for use with Alberta Education's Cognitive Diagnostic Mathematics Assessments (CDMA) in grade 3. There were three core members of the CDMA project team at Alberta Education. This team included the exam manager who was responsible for the CDMA project overall and two exam managers responsible for overseeing test development in each of grades 3 and 6. All three exam managers had extensive mathematics classroom teaching experience ranging from 15 – 32 years as well as large-scale test development

experience. The project team also had consultations with other individuals within Alberta Education such as programmers and psychometricians. However, the main points of contact were the three exam managers.

CDMA was a curriculum-based set of assessments that could be used throughout the school year to measure students' strengths and weaknesses in thinking and learning mathematics. Results from CDMA were communicated through student score reports. The information from CDMA can be used to provide valuable instructional guidance to the teacher, which may be needed to design remedial instruction programs or supplemental interventions for students. Development and field testing of CDMA in grade 3 began in 2008 and concluded in April 2011.

The purpose of the reports was to provide a summary of student performance across attributes in one skill category. This type of reporting allows the reader to compare mastery across attributes providing a diagnostic profile of cognitive strengths and weaknesses. A student or parent can use this document as a starting point for discussions with a teacher or tutor on areas requiring further instruction or study. The following reports are based on the strand of Number: Developing number sense under the skill category of "Estimate quantities less than 1000 using referents". The cognitive model for this reporting scheme is a five attribute hierarchy, depicted in Table 3. The reports presented in Figures 2, 3 and 4 incorporate the AHM reporting elements of the cognitive model, attribute scores, and attribute descriptions. The diagnostic score reports were developed

under the condition that the cognitive model adequately accounted for the observed examinee responses.

Table 3.

Example attribute hierarchy and skill descriptors from the strand of Number.

Attribute	Skill Description
A1	Apply estimation using 100 as a referent to a quantity of 100 to 1000
A2	Apply estimation using 10 as a referent to a quantity of 100 to 1000
A3	Apply estimation using 25 as a referent to a quantity of 100 to 1000
A4	Identify a justification or estimation strategy to solve a problem using a quantity of 100 to 1000
A5	Solve an estimation problem using a quantity of 100 to 1000

The reports were also developed with the intention that they could either be viewed on the web (static presentation) or printed. This decision constrained the number of pages of the report to two, so it could be printed on the front and back sides of a letter-sized page. The goal was to create a standalone document and any references in the sample report to additional resources were made for illustrative purposes, however these resources were not created. When designing the score reports, great effort was made to incorporate the design guidelines and reporting recommendations reviewed earlier in the paper. All documents were created using a program called Adobe In-Design for greater flexibility in the formatting and creation of the document.

Score report 1: Cognitive diagnostic score report with tabular and graphical representation of skill mastery without interpretive material, spring 2009.

Figure 2 represents the first reporting template created with input from Alberta Education. This score reporting template was used during the field test in Spring of 2009.

Reporting considerations. In this reporting scheme, elements specific to CDA were reported together with elements common to large-scale reporting. This approach was chosen to provide the reader with familiar reporting features while introducing relatively unfamiliar and novel diagnostic scores. For example, a total score for this skill category was provided in the top-left corner where this score could be a total correct or scaled score. Notwithstanding the limitations of reporting total scores in terms of interpretation, reporting a total score in combination with diagnostic scores can illustrate to students, parents, and teachers that the same total score can be characterized by different patterns of skill mastery. In this way, cognitive diagnostic feedback highlights student performance. Information on report contents and directions for how to read the report was placed in the top section at the beginning, serving as an overview for the reader. Also, a reminder was provided for the reader to consult the interpretive material on the second page for further detail and explanation of the score report.

The middle section of the report contained information not typically reported from large-scale assessments: student diagnostic scores and specific information on attribute-level performance. The attribute labels in the first column

corresponded to a standardized attribute descriptor which provided an abbreviated description of the cognitive skill measured by the test. Attribute-level performance was illustrated by providing information on item-level performance under the columns “Item”, “Your Answer”, “Correct Answer” and “Answer Summary”. Actual skill performance, as indicated by the attribute score, was presented in graphical form with a bar graph to report skill probabilities in reference to skill mastery classifications. For this sample score report, diagnostic scores were reported in terms of the length of the bar graph with values ranging from 0 to 1. The length of the bar was based on the estimated attribute score. This method of presenting scores does not report the actual numerical probabilities or errors of measurement related to estimation of skill mastery. Finally, a summary of the scoring was given with a breakdown of the number of questions answered correctly, incorrectly, and omitted.

The bottom section of the report provided a narrative summary, in point form, of the student’s performance across all attributes. An element of redundancy in information within the report can be helpful for understanding the major outcomes of the assessment without focusing on the details, if desired. In this section, a cognitive diagnostic summary, instead of an item-level performance summary, was provided to direct the student to areas of strengths and weaknesses based on his or her item responses. The student was provided with a short recommendation on how to improve and a reminder to consult with his or her teacher for further guidance in interpreting and using the feedback in the report.

Design considerations. This report was designed in three sections with related but different functional purposes. The top section of the report contained orienting information in the form of an overview of contents for the reader. Student identification information and a summary score was brought to the attention of the reader by placing it in a colored, boxed area in the top-left hand corner of the page, which is where the eye naturally begins when reading a document.

The middle section of the report, “Review Your Answers” contained diagnostic information regarding attribute mastery along with item-level performance. The results were based on a linear cognitive model presented in Table 3, therefore with this particular example, presenting attribute-level results vertically was consistent with the form of the cognitive model. An arrow placed beside the attribute labels pointing upwards, provided the reader with additional information about how the attributes are related in the cognitive model. Attribute labels, item-level performance, and skill performance were grouped together into three areas within this section, while variations in line thicknesses were used to visually separate attribute level results vertically. Skill performance was presented in graphical form where the horizontal bars representing skill performance were grouped together and the outcomes for each attribute were vertically aligned. This is done to assist with making comparisons among attribute performance to give a sense of where the cognitive strengths and weaknesses lie.

The bottom section of the report was structurally and visually separated from the middle section by the use of a box. This section contained mostly text-

based information using bullets with left alignment for clarity in presentation and ease of reading.

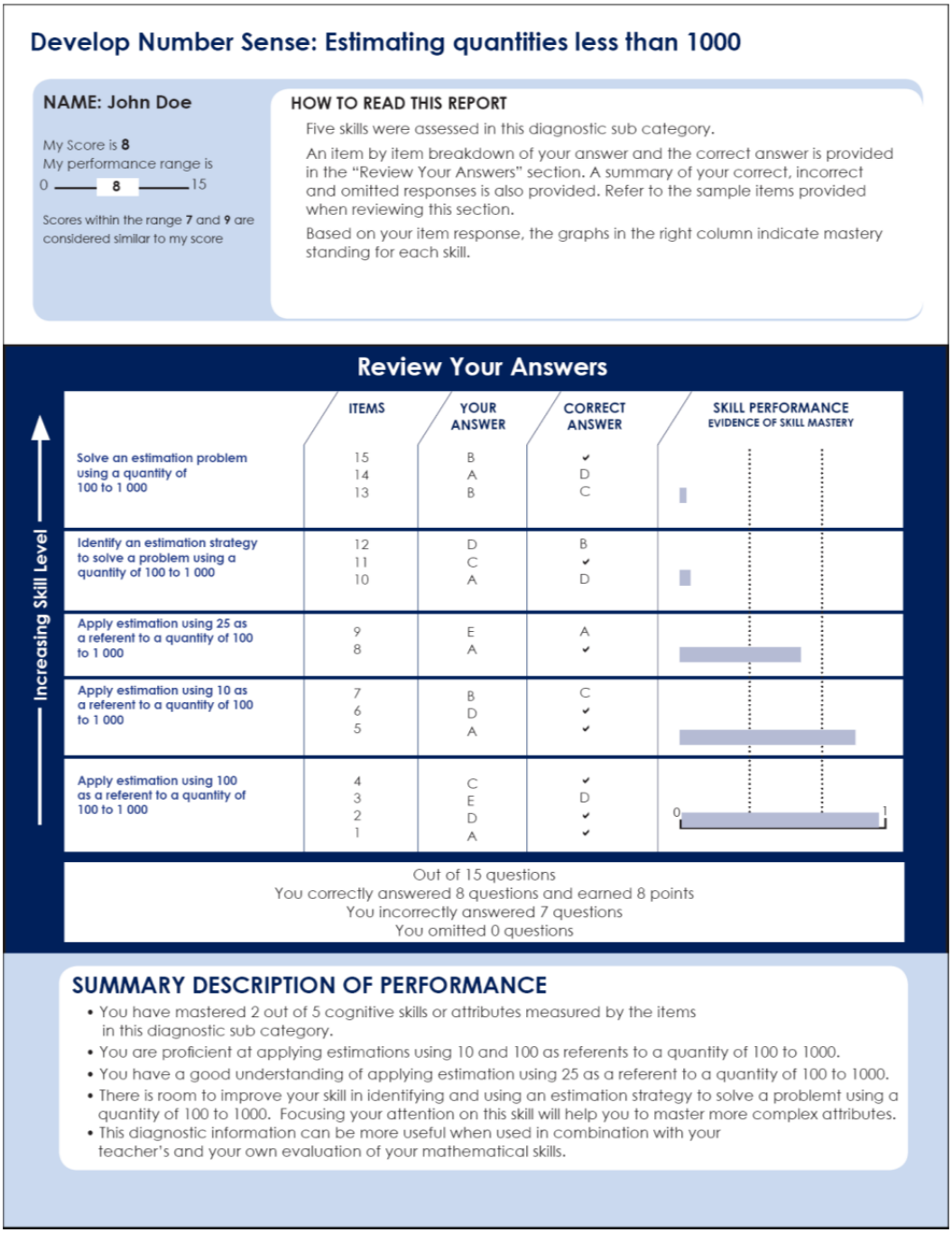


Figure 2. Diagnostic score report 1 with graphical representation of skill mastery without interpretive material created for spring 2009.

Score report 2: Cognitive diagnostic score report with tabular representation of skill mastery without interpretive material, spring 2010.

After completion of field testing in the Spring of 2009, the reporting template was modified with input from the team at Alberta Education. The alternate template is shown in Figure 3. The intent was to simplify the appearance of the score report by changing the amount of text in the reports and by changing how the diagnostic scores would be reported. Additionally, the visual aesthetic of the report changed to incorporate more color.

Reporting considerations. The project team wanted to streamline the information in the score report. This required some decisions by the team at Alberta Education with regard to what information should be reported in the new reporting template and what information should be removed. A consideration was made of what the basic information a teacher may expect to have on a diagnostic score report. Given this, the general features to appear on the report that remained included the Mathematics strand reported and the attribute descriptors. Total score, item level information such as the items aligned to each attribute, student response and correct response, were removed in this version. Orientating features such as “How to Read the Report” as well as point-form summaries of student performance were also removed.

The major change from score report 1 to score report 2 was how the attribute scores were reported and presented. Student performance was reported in terms of the evidence of skill mastery demonstrated based on his or her responses: limited, moderate, and consistent. Determining mastery probabilities associated

with each level of skill mastery would require the use of some type of standard setting procedure. For this sample score report, diagnostic scores were reported in terms of performance levels within each attribute to provide some context for interpreting the attribute scores. Checkmarks were used to denote which category the student's response was classified into. Last, a definition of each classification category was provided for the reader's reference.

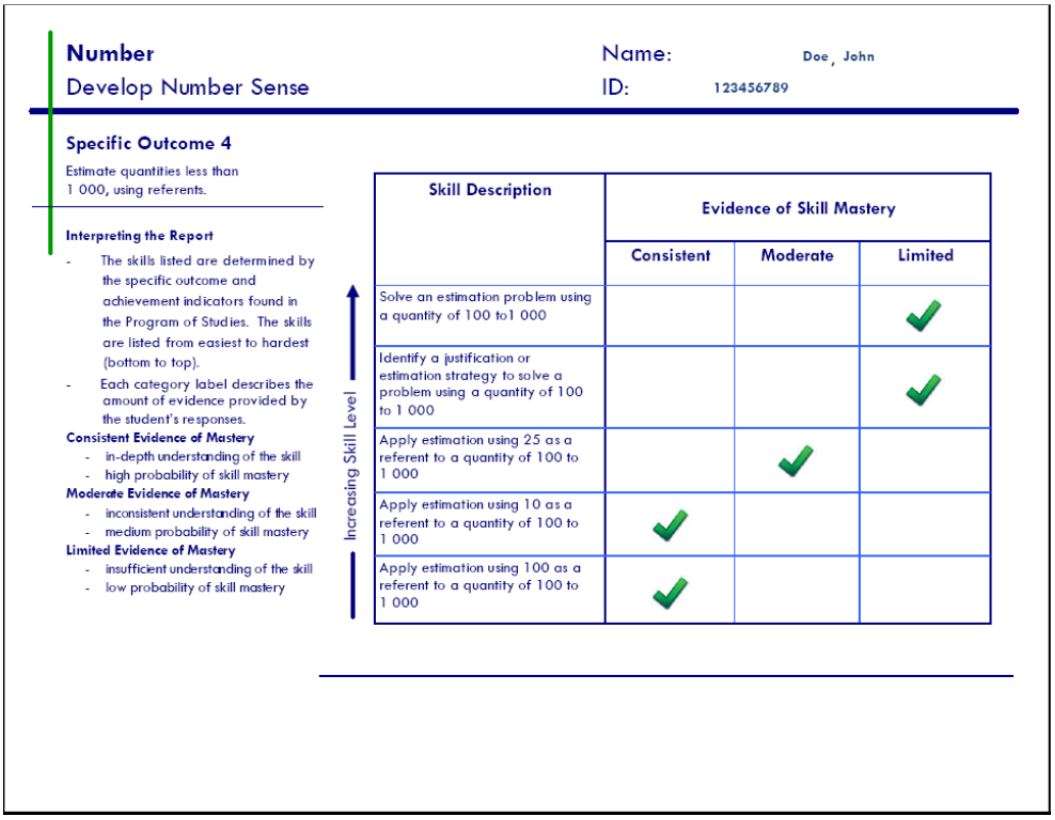
Design considerations. This report was designed in three sections with different functional purposes. The very top left section of the report contains information on the Mathematics strand tested (i.e., Number) and the name of skill hierarchy (i.e., Developing Number Sense). Student identification was also placed at the top but at the right hand corner. This information was visually separated from the remaining information in the report by the use of a horizontal line.

The bottom section of the report contained information on the student's performance. The bottom left side of the report contained reference material to aid the reader when reading and interpreting the information. First, the specific outcome from the Alberta Education curriculum that underlies the skill hierarchy was provided. Then, information on how to interpret the report including the category definitions was described.

The bottom right section of the report contained the skill scores reported in a tabular format with graphical elements. Again, the results were based on a linear cognitive model presented in Table 3, therefore presenting attribute-level results vertically was consistent with the form of the cognitive model. To reinforce the ordering of the attributes from simple to cognitively complex, an arrow was

placed beside the attribute labels pointing upwards. Skill performance was presented in tabular form with checkmarks indicating the amount of evidence of skill mastery. Checkmarks were chosen as the symbol to denote classification because its form and meaning is readily accessible and understood with teachers. The tabular format, with skill descriptors left aligned and vertical columns for each performance category, helped the reader to make visual comparisons. This format easily shows the pattern of results and the reader could quickly pinpoint where in the hierarchy of skills a student is having difficulty.

For aesthetic considerations, the bottom section of the report was visually formatted using a 1/3 – 2/3 division of space. In contrast with score report 1, this report has more white space which can help make elements of the report easier to read.



Number
Develop Number Sense

Name: Doe, John
ID: 123456789

Specific Outcome 4

Estimate quantities less than 1 000, using referents.

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
Solve an estimation problem using a quantity of 100 to1 000			✓
Identify a justification or estimation strategy to solve a problem using a quantity of 100 to 1 000			✓
Apply estimation using 25 as a referent to a quantity of 100 to 1 000		✓	
Apply estimation using 10 as a referent to a quantity of 100 to 1 000	✓		
Apply estimation using 100 as a referent to a quantity of 100 to 1 000	✓		

Figure 3. Diagnostic score report 2 with tabular representation of skill mastery without interpretive material created for spring 2010.

Score report 3: Cognitive diagnostic score report with tabular and graphical representation of skill mastery with interpretive material, alternate template.

Both score report 1 and 2 used elements of the diagnostic score reporting framework in differing degrees with regard to the amount and type of information reported. Each template was developed with input from the team at Alberta Education. However, according to the recommendation of Goodman and Hambleton (2004), score reports should have some supporting material to assist the reader with understanding and interpreting the information with the report. To this end, score report 3 incorporated the design layout of score report 1, with the reporting scheme of score report 2, and added a page of interpretive material.

Reporting and design considerations. Figure 4 reports the same information provided in score report 1 (see Figure 2) using the same design principles. However, like score report 2, actual skill performance as indicated by the attribute score is presented in graphical form with three sections on the bar denoting evidence of skill mastery: limited, moderate, and consistent. Placement of the bar was based on the estimated attribute score and the length of the bar can be adjusted to reflect the reliability associated with the mastery classification procedures (see Gierl, Cui, & Zhou, 2009). Similar to score report 1, the horizontal bars representing skill performance were grouped together, where the outcomes for each attribute were vertically aligned. As with the other two reporting templates, this presentation can assist a person to make comparisons

among attribute performances to gives a sense of where the cognitive strengths and weaknesses lie.

Interpretive material for the diagnostic score report. The new kind of information reported from a CDA was emphasized in the accompanying interpretive material which is illustrated in Figure 5. The back page of the report can be viewed in two sections. The top section provides a description of the skill category as defined by the cognitive model and the attributes. Attribute descriptions are based on the cognitive model presented in Table 3. If desired, the level of description in the interpretive guide can be simplified from the attribute descriptions provided in the cognitive model for the purposes of reporting to a target audience, such as teachers. Descriptions within the cognitive model can be written at a higher level of detail because the model represents all aspects of the intended measured construct which can act as a test blueprint for item developers in addition to serving a descriptive reporting function.

The bottom section assists with providing contextual information when interpreting the contents on the front page of the report. This information was grouped under headings of anticipated “Frequently Asked Questions” that a reader may have such as directions for how to interpret scores, how to interpret skill mastery as well as how to use this information to improve student performance. This technique was employed to decrease anticipated misinterpretations and unintended uses of diagnostic scores. Due to the large narrative component, the typeface size was kept at 10pt, line lengths were kept

short, and text was chunked into columns separated by white space to maintain legibility.

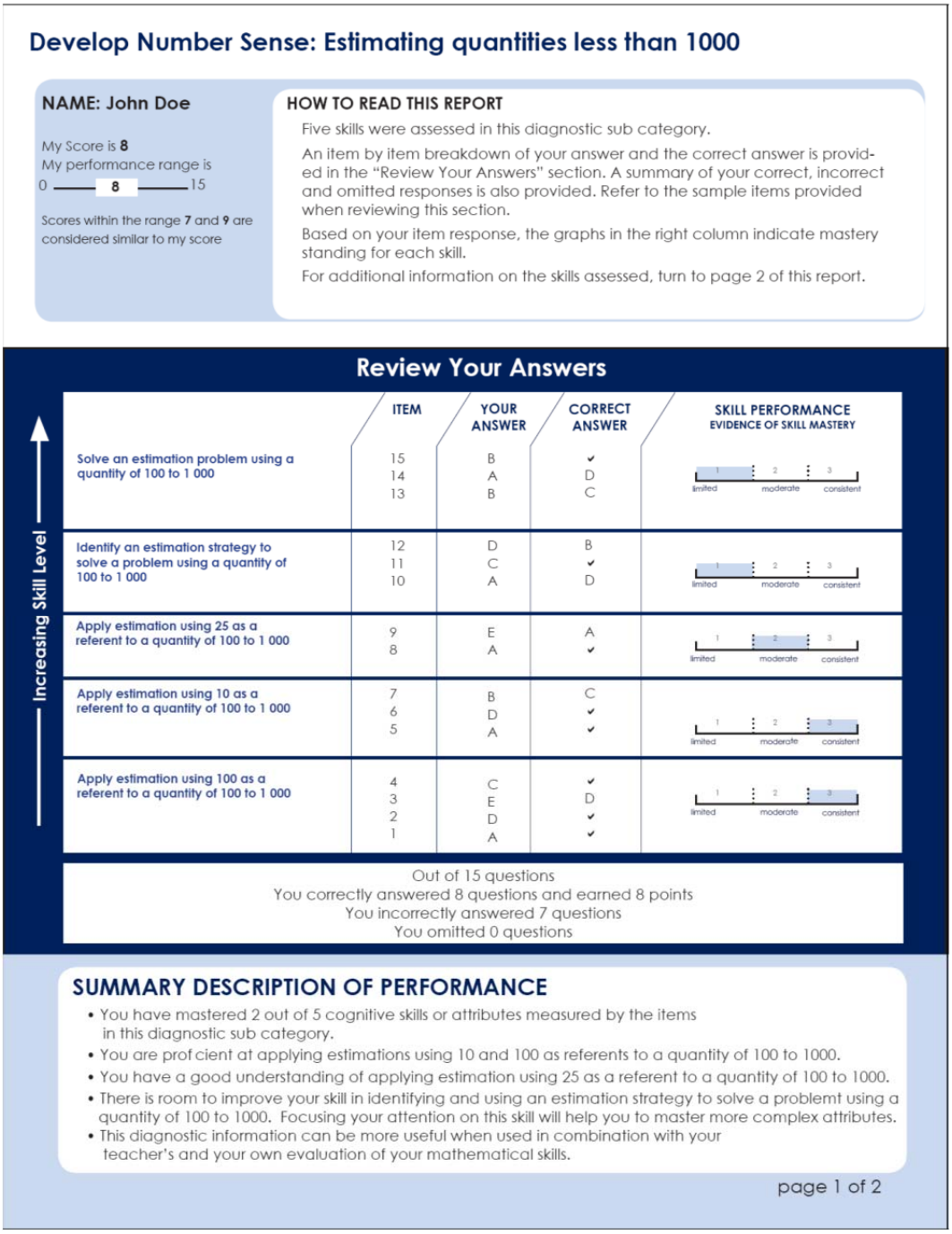
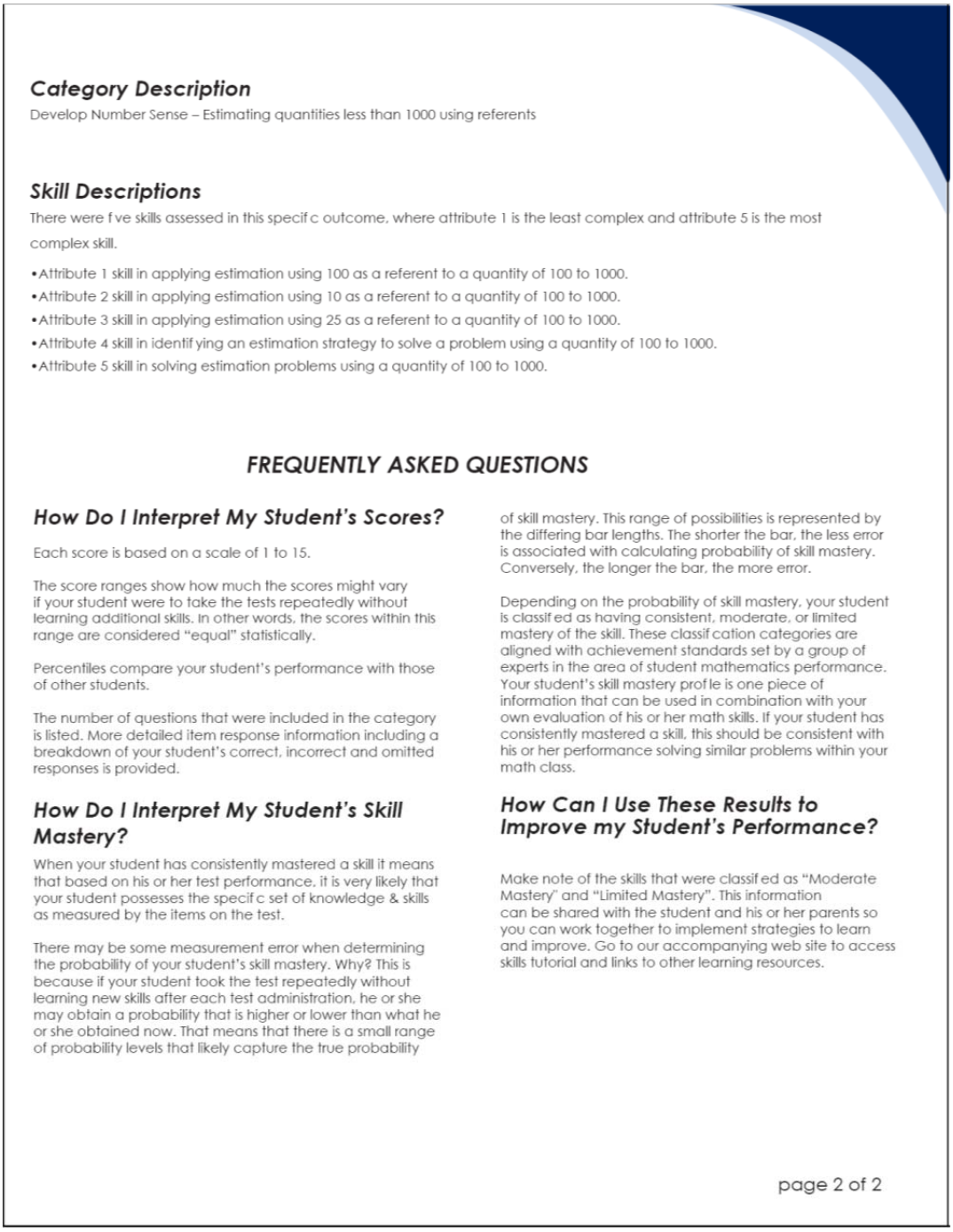


Figure 4. Diagnostic score report 3 with graphical representation of skill mastery with interpretive material, alternate reporting template.



Category Description

Develop Number Sense – Estimating quantities less than 1000 using referents

Skill Descriptions

There were five skills assessed in this specific outcome, where attribute 1 is the least complex and attribute 5 is the most complex skill.

- Attribute 1 skill in applying estimation using 100 as a referent to a quantity of 100 to 1000.
- Attribute 2 skill in applying estimation using 10 as a referent to a quantity of 100 to 1000.
- Attribute 3 skill in applying estimation using 25 as a referent to a quantity of 100 to 1000.
- Attribute 4 skill in identifying an estimation strategy to solve a problem using a quantity of 100 to 1000.
- Attribute 5 skill in solving estimation problems using a quantity of 100 to 1000.

FREQUENTLY ASKED QUESTIONS

How Do I Interpret My Student's Scores?

Each score is based on a scale of 1 to 15.

The score ranges show how much the scores might vary if your student were to take the tests repeatedly without learning additional skills. In other words, the scores within this range are considered "equal" statistically.

Percentiles compare your student's performance with those of other students.

The number of questions that were included in the category is listed. More detailed item response information including a breakdown of your student's correct, incorrect and omitted responses is provided.

How Do I Interpret My Student's Skill Mastery?

When your student has consistently mastered a skill it means that based on his or her test performance, it is very likely that your student possesses the specific set of knowledge & skills as measured by the items on the test.

There may be some measurement error when determining the probability of your student's skill mastery. Why? This is because if your student took the test repeatedly without learning new skills after each test administration, he or she may obtain a probability that is higher or lower than what he or she obtained now. That means that there is a small range of probability levels that likely capture the true probability of skill mastery. This range of possibilities is represented by the differing bar lengths. The shorter the bar, the less error is associated with calculating probability of skill mastery. Conversely, the longer the bar, the more error.

Depending on the probability of skill mastery, your student is classified as having consistent, moderate, or limited mastery of the skill. These classification categories are aligned with achievement standards set by a group of experts in the area of student mathematics performance. Your student's skill mastery profile is one piece of information that can be used in combination with your own evaluation of his or her math skills. If your student has consistently mastered a skill, this should be consistent with his or her performance solving similar problems within your math class.

How Can I Use These Results to Improve my Student's Performance?

Make note of the skills that were classified as "Moderate Mastery" and "Limited Mastery". This information can be shared with the student and his or her parents so you can work together to implement strategies to learn and improve. Go to our accompanying web site to access skills tutorial and links to other learning resources.

page 2 of 2

Figure 5. Interpretive material for score report 3.

Score reporting templates for evaluation with teachers.

The three score reporting templates were evaluated with Grade Three teachers participating in the diagnostic mathematics field test. Score reports 1 and

2 have been used on two different field test administrations. Score report 3 had not been used during the field test administrations. Score report 3 differs from the other two score reports with the presentation of the attribute scores and with the inclusion of interpretive material. Interpretive material was not developed and distributed with the previous two score reports, although it was recommended to do so. A summary of the key differences between the score reports is presented in Table 4.

Table 4.

Key feature differences between score report templates.

	Score Report 1	**Score Report 2**	**Score Report 3**
Method of reporting diagnostic scores	- Bar graph of skill probability values between 0 and 1 - Bars had a common left alignment	- Tabular format with checkmark symbols - Diagnostic scores were reported in the form of performance level categories	- Bar graph of skill probability values between 0 and 1 - Bars were centred around the skill probability value
Inclusion of item-level results	Yes. Item numbers, your answer, correct answer, and summary of performance provided.	No.	Yes. Item numbers, your answer, correct answer, and summary of performance provided.
Inclusion of additional interpretive material	Minimal. Description on how to read the report is provided.	Minimal. Definitions of each performance category is provided.	Description on how to read the report plus inclusion of an interpretive guide are provided.

The instruments used to evaluate the three score reporting templates are discussed in the next chapter.

Summary of Chapter 3

In this chapter, the diagnostic reporting framework for developing score reports was introduced. Based on this framework, three alternate score reporting templates were created for the grade 3 diagnostic mathematics assessment in consultation with Alberta Education. Score reports 1 and 2 were used during two different field test administrations, whereas score report 3 is a hybrid of reports 1 and 2 with interpretive material. As recommended by Goodman and Hambleton (2004) and in line with previous studies on score reporting (Ryan, 2001; Trout & Hyde, 2006), evaluation of the reports with the target audience is needed. The instruments and method for evaluating the score reports is presented next in Chapter 4.

CHAPTER 4 METHOD

The purpose of this study was to develop a framework from recommendations put forth in the educational measurement and information design literature, and based on this framework, create and evaluate diagnostic score reports. The grade 3 cognitive diagnostic mathematics assessment (CDMA) with Alberta Education provided the context for this study. Specifically, my research questions were:

1. What should be reported on a diagnostic score report? and How should this information be presented?
2. To what extent are the score reports understandable and interpretable by teachers, the target audience? What are teachers' understanding of the diagnostic information in the score report? What information do teachers consider to be diagnostic and why?
3. Given the answers to questions 1 and 2, how useful are the score reports? Is it clear what the next instructional steps should be for the student?

To begin to answer the first research question, a focused literature review was conducted to create a framework for developing diagnostic score reports. However, the remaining research questions required an operational context to be answered. This chapter outlines the methods used for this study. First, a description of the sampling and participants is provided. Second, the instruments used for evaluating the score reports are presented and described. Third, the detailed procedures used for answering the research questions are outlined.

Sampling and Participants

Participants were recruited using a two-stage convenience sampling procedure. The sample of teachers was drawn from a potential pool of all teachers whose classes participated in the Grade 3 Diagnostic Mathematics field test in 2009, 2010, and 2011. Prior experience with working on the diagnostic mathematics field test demonstrated that approximately 35 teachers take part in the field test each year.

The teachers participating most recently in the field test (i.e., Fall 2010 and Winter 2011) were contacted first by Alberta Education via email. These teachers were contacted first due to their recent experience with the diagnostic score reports. Alberta Education emailed potential participants a research information sheet explaining the purpose of the study and the task he or she was requested to complete. Please see Appendix A for details. Teachers wishing to participate were asked to contact me indicating their willingness to participate. Depending on the number of responses obtained at this stage, participants from 2009 were contacted. Then, a secure link to the online questionnaire was emailed to the teacher. Email reminders were sent by Alberta Education at 1.5 and 3 weeks after the original email request for participation is made.

Teachers who agreed to complete the online questionnaire were reminded of the opportunity to participate in a follow-up interview to further discuss reporting options. Therefore, a sub-sample of teachers who completed the questionnaire were recruited for the interview. Although recruitment of teachers for the interview component alone was possible, it was assumed that a teacher

would more likely consent their time to an activity that would take less time to complete (i.e., 15 minutes for the questionnaire compared to 30 minutes for an interview).

Given the research questions, proposed procedures, and sampling design, the minimum sample size was governed by the number of participants required for qualitative analysis of the interview data. Although prescriptive sample sizes are not usually defined in the beginning for qualitative studies, an estimate of the minimum number of subjects required can be specified to facilitate later analyses. Studies in usability research suggested that information saturation can occur with a sample as small as 5 subjects (Nielsen, 1993). Taking into consideration the two-stage sampling technique employed and the characteristics of the target group, which were teachers who self-selected their class to participate in the field test, information saturation was achieved with 5-8 subjects evaluating one score report.

An important factor when working with these teachers was to minimize the burden of participating in the study. Members of the Alberta Education team cautioned the use of extended questionnaires and interviews as this may deter teachers from participating. In keeping with minimizing the burden of participation, a teacher was asked to evaluate only one alternate score report instead of two. However, the teacher was provided with an opportunity to evaluate a second alternate template if he or she was interested.

Inclusion and Exclusion Criteria

In the fall of 2010, a number of Grade 4 teachers participated in the field testing of some topics that would not be covered in Grade 3 by the end of April 2011. These teachers were also invited to participate for two reasons. First, one of the purposes of the study is to evaluate the diagnostic score reports with teachers who are familiar with the content being assessed and have experienced the online diagnostic assessment. The Grade 4 teachers are in a position to provide detailed opinions on the score reports due to their experience more so than teachers who have not participated in the field test at all. Second, there are many different sources and uses of diagnostic information for a teacher. However, it is unlikely that the sources and uses are idiosyncratic to either grade 3 or grade 4. The contexts under which each teacher is working may be different and this was noted during the interview. Teachers who have not participated in the diagnostic mathematics field test were excluded from participation. While their views on score reporting are valuable, one of the purposes of this study is to evaluate the diagnostic score reports. As such, teachers who have had exposure to the diagnostic mathematics assessment and score reports are the target audience.

Instruments

The score reports for the diagnostic math assessment described in Chapter 3 were developed based on previous research in educational measurement and information design principles and in consultation with the Alberta Education team. At some point in the development phase, the intended audience of the score reports should also be consulted to provide feedback on the score reports to ensure

that the information is relevant, understandable, and useful. Two sets of instruments were developed for the evaluation component of this study. A questionnaire evaluating the content and format, sources, uses and need for diagnostic information was created. Then, an interview guide was also created with questions designed to collect feedback on the score reports and evaluate the perceived utility and uses for the diagnostic information reported.

Questionnaire to Evaluate the Diagnostic Score Reports

Previous studies examining score reporting in different contexts (Huff and Goodman, 2007; Jang, 2009; Ryan, 2003; Trout and Hyde, 2006) have used similar questions to evaluate score reports. The questions used in these studies provided the basis for constructing the questionnaire for this study. The first questionnaire was developed prior to the Spring administration in 2009 with the input and feedback of the project team at Alberta Education and at the University of Alberta. Adobe was used to format the questionnaire so teachers could choose to answer using a computer or to print it off. The questionnaire was revised again in the Fall of 2010 after discussion with the Alberta Education team about potential issues with filling out the questionnaire using Adobe PDF (i.e., saving the file and emailing the information), faxing confidential information and the time required by the teacher to participate. Based on this discussion, it was decided to transfer the questionnaire to an online environment using an internet survey provider. The online questionnaire was pre-tested with two colleagues to assess the length of time it may take to answer all questions. The questionnaire took less than 15 minutes to complete on both occasions. Given this outcome, the

questionnaire was deemed to be an appropriate length so as not to overburden the participant. Appendix B contains the Word document of the online version of the questionnaire.

The 25-item questionnaire was constructed with 4 major sections: 1) content and format of the report, 2) diagnostic information from test score reports, 3) background information, and 4) consent to a follow-up telephone interview.

Content and format. This section contained 9 Likert-scale questions; 4 questions to evaluate the content and 5 questions to evaluate the format of the score report used for this round of field testing (see Score report 2). The participant was asked to rate each statement on a scale from 1 to 5 where 1 = highly disagree and 5 = highly agree.

Diagnostic information from test score reports. This section contained 14 questions to identify teachers' sources and uses of diagnostic information, and their preferences for receiving diagnostic information in score reports. The first question asked the participant to rate on a 5-point Likert-scale whether various test results provided him or her with information to make a diagnostic decision where 1 = highly disagree and 5 = highly agree. For example, Question 11 asked the participant to rate whether overall or total subject-level score (e.g., Number = 16/24) provides information to make a diagnostic decision. The next two questions asked whether the participant expected a: 1) diagnostic score report for an individual student aggregated across hierarchies in one strand of Mathematics, and 2) at the classroom level aggregating across all students for one skill hierarchy.

The next set of questions was related to the participants' sources of and uses for diagnostic information. The first question asked what sources provide the participant with information to make decisions about a student. The next two questions asked the participant to list his or her uses of diagnostic information from assessments and how frequently it is used in his or her teaching. Finally, the last two questions asked the participant how interested he or she was in receiving information from the score report, and the preferred mode for receiving student score reports (i.e., print-based or web-based).

Background information. This section was composed of six questions. Information on the participant's gender, years of teaching, level of education, and background information on educational assessment was requested.

Consent to a follow-up telephone interview. In this section, the participant was asked whether he or she would be willing to participate in a telephone interview to further discuss their opinions on the score report as well as to evaluate an alternate reporting template.

Interview guide to evaluate the score reports.

To keep the time to complete the questionnaire at a reasonable length, open-ended questions were omitted. In-depth detailed feedback on the score report would be more appropriately collected using a semi-structured interview format. Again, using previous score reporting studies as a guide (Ryan, 2003; Trout and Hyde, 2006), a set of open-ended questions were developed. The questions are referred to as a "guide" because it was anticipated that some questions may be refined, added, or deleted based on the first few actual

interviews with the participants. The interview guide was created with the intent to have the interview completed in approximately 30-45 minutes. Appendix C contains the sample interview guide.

The interview protocol began in the context of evaluating the current score report used during the field test administration (i.e., score report 2). First, an opportunity was given to the participant to clarify his or her responses to the questionnaire. Then, a set of open-ended questions were asked to provide an overall evaluation of the content and format of the current score report. One question further explored what information a teacher considers diagnostic that is, what it means to the teacher when we say a test has diagnostic information value. Last, the participant was asked to describe and interpret the information within the report in the context of showing the report to a parent.

The context for the next part of the interview was evaluation of an alternate template to the current score report. The participant was asked to consider one of the two alternate reporting templates (i.e., score report 1 or score report 3). The same set of 9 Likert-scale questions used in the questionnaire to evaluate content and format were asked first, followed by the set of open-ended questions. The next set of questions in the interview asked the participant to consider both the current and alternative templates and choose which reporting template is preferred and the reasons why. At this stage of the interview, the participant may have chosen to make suggestions on how to improve upon the score reports or to create a different reporting template.

Then, questions on the perceived utility of the information in the score reports for informing future instruction or remediation activities were asked. Perceived utility is stated here because it was not assumed that the teacher had used or had an adequate opportunity to use the diagnostic information in his or her own classroom. The teacher was also asked whether there were concerns with how the information might be used. Finally, the interview ended with an opportunity for the participant to make any final comments.

Procedure

Overview

This study used two data sources to address the research questions. Both questionnaire and interview formats were used for two considerations. First, the use of both approaches allows for a complementary approach to answering the research questions. Findings or themes in the survey were explored in an interview format with open-ended questions. Second, the use of both a questionnaire and interview is made on pragmatic grounds. This study was meant to answer a practical question in an operational context. The use of both data sources to achieve this end is desirable because both methods together can provide more comprehensive information on issues related to score reporting.

This study was conducted in three stages. In Stage 1, the questionnaire was administered to a sample of teachers who participated in the diagnostic mathematics field test. In Stage 2, follow-up telephone interviews were conducted with a subsample of teachers from Stage 1. In Stage 3, results of the analysis from

stages 1 and 2 were used to suggest revisions to the score reporting framework and/or templates.

Stage 1: Administrating the questionnaire.

The instrument for data collection was the online teacher questionnaire and housed by SurveyGizmo. Settings on the website ensured that a secure link was established when completing the questionnaire. In addition to recording the participants' responses, SurveyGizmo recorded whether the questionnaire was incomplete, whether the participant required multiple sessions to complete the questionnaire, and the geographical location of the respondent. Each completed questionnaire was assigned an ID by the system rendering each submission anonymous unless the participant consented to a follow-up telephone interview upon which he or she was required to provide a name and contact information.

Stage 2: Conducting follow-up interviews.

Interviews were conducted over the telephone. Two days prior to the interview, participants were emailed a copy of their responses to the online questionnaire as well as the interview questions. This was done to encourage reflection on the score reports that may help generate rich responses and to maximize the interview time for discussions on the score reports. Participants were required to read through one alternate student score report and answer substantive questions about the text and data displays to assess how much of the information presented was understandable and interpretable. Teachers also evaluated the usefulness of reported information for guiding instructional planning and student learning efforts. During the interviews, participants were

encouraged to volunteer their opinions and suggestions. Interviews were audiotaped and transcribed. Identified problem areas provided valuable information about concerns encountered by users when interpreting the content and format of the diagnostic student score reports.

Stage 3: Data analysis and revision of score report.

Questionnaire. Descriptive statistics were used to describe the sample and to summarize survey responses. Comparisons of ratings for questions evaluating the content and format of the score report across reporting templates were made to evaluate whether there was a preference for a particular reporting template.

Interviews. Interview data were transcribed prior to analysis. Closed-ended questions used to evaluate the content and format of the alternate score report were summarized using descriptive statistics. Responses to open-ended questions were analyzed using thematic analysis (Boyatzis, 1998). An inductive approach was used to code and identify emerging themes in the interview data. According to Boyatzis (1998), three stages are involved in using thematic analysis with an inductive approach: 1) Deciding on the sample and subsamples from which to develop the code, 2) Developing themes and a code, and 3) Validating the code.

In Stage 1, consideration is made upon the sample from which the code will be developed. This can be accomplished one of two ways. All the data can be collected first and then a subsample used to analyze and develop themes which will be applied to the remaining sample. This technique would presuppose sampling according to some criterion to define these subgroups prior to analysis.

Given the unfolding convenience sampling design used for recruiting participants for the interview, a second approach was used. The first two or more interviews provided the basis for creating the code. Based on this code, the third interview was coded. Revisions to the code occurred with new information collected in subsequent interviews that could not be captured using the initial code. Development of the code was iterative until all of the data could be coded or accounted for.

In Stage 2, there were three steps involved in developing the code. First, the information in the transcript was reduced. This was accomplished by creating a summary of the interview data according to the major categories reflected in the questionnaire: Content and format, understanding and interpretation, and uses of and preferences for information. Summaries were also created for responses to open ended interview questions such as teachers' perceptions of diagnostic assessment characteristics. Second, the summaries for each category were used for identifying themes within each category, when developing the code. Third, the code was developed.

In Stage 3, the code was cross-validated by coding the remaining data. A theme was considered to be valid in that it accounts for a range of differing responses in the sample and not just the response of one. An illustration of how the code was developed for summarizing teachers' perceptions of diagnostic assessment characteristics is provided in Appendix D.

Revision of the score reporting framework. Using the feedback from the questionnaire and interviews, a revision to the score report framework was made.

Where applicable, suggestions were summarized with respect to adding or removing information, organization and/or presentation of the information for the individual templates.

Summary of Chapter 4

In this chapter, the procedures and instruments for evaluating the score reports were described. Using convenience sampling, teachers whose classes completed the Grade 3 diagnostic test were recruited to complete an online questionnaire evaluating the score report template used for the recent round of field testing. A subsample of these teachers were recruited to participate in a telephone interview to follow up with some of their answers on the questionnaire and to evaluate at least one alternate score reporting template. Based on the feedback received from the teachers, proposed revisions to the current score reporting framework and individual reporting templates were made.

CHAPTER 5 RESULTS

In this chapter, the results of the evaluation for each of the three reporting templates described in Chapter 3 are presented. These reporting templates were developed to answer research question #1. This chapter outlines the results of the evaluation conducted to answer research questions #2 and #3. To begin, background information on the teacher participants is provided. Then, the results are presented in order of research question and within each question, by template. For each research question, where applicable, the quantitative results are provided first followed by the qualitative results. A visual overview of the structure of this chapter is shown in Figure 6.

Participant Recruitment and Characteristics

Teachers were invited to participate in the study via email by a member of the Alberta Education team. Of the 40 teachers participating in the 2010-2011 field test, 5 teachers (13%) responded and completed the online questionnaire. Of the 20 teachers who participated in the 2009-2010 field test, 2 teachers (10%) responded and completed the online questionnaire. Through nomination, another 19 teachers were contacted and 14 teachers (74%) responded and completed the online questionnaire. Of the total 21 teachers who responded to the online questionnaire, 14 teachers agreed to a follow-up interview of which 11 teachers (52%) participated.

Table 5 provides background characteristics of the teacher sample who completed the online questionnaire for this study. The sample was mostly female (86%) with an average of 15 years teaching experience (M=15.42, SD=7.58). The

majority of the sample (86%) were currently teaching Grade 3 at the time of participation. Two of the remaining three participants were teaching combined grade classes and one participant was teaching Grade 2. On average, teachers taught for eight years at their current grade level (*M*=7.9, *SD*=5.69). Almost half of the sample had a B.Ed. as their highest level of education (48%) and the other half of the sample had either a degree plus a B.Ed. (38%) or a M.A./M.Ed (14%).

Table 5. *Summary statistics for select demographic information (n=21)*

		n	%
Gender	Female	18	86
	Male	3	14
Total Years Teaching	<10	4	19
	10-14	7	33
	15-20	4	19
	>20	6	29
Current Grade Teaching	3	18	86
	Other	3	14
Number of Years Teaching at Current Grade	<5	6	29
	5-9	8	38
	10-14	4	19
	>15	3	14
Highest Level of Education	B.Ed.	10	48
	Degree + B.Ed.	8	38
	M.A./M.Ed.	3	14

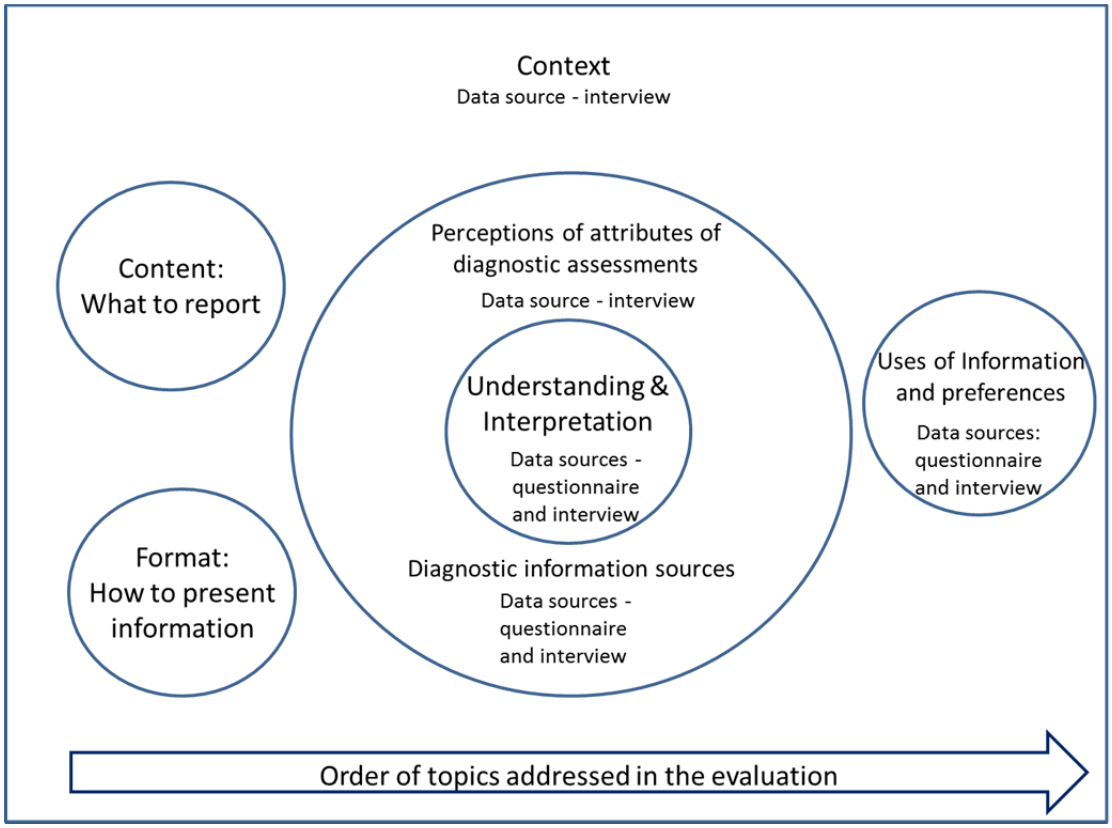


Figure 6. Overview of presentation of results.

Participants were also asked about their background in educational assessment. Table 6 outlines the areas from which the teachers learned about educational assessment. Overall, the teachers in this sample reported having knowledge about educational assessment in a formal setting, as part of their degree, as well as in experiential settings, such as the classroom. Almost all the teachers identified classroom experience (95%) and inservices or workshops as sources of knowledge around educational assessment. Almost 80% of the sample also reported taking at least one course on educational assessment as part of their degree program.

Table 6. *Sources of teacher knowledge of educational assessment (n=21)*

	n	%
Classroom experience	20	95
Inservices or workshops	19	91
University or college course as part of a teacher pre-service program	13	62
University or college course as part of a graduate or extra courses program	4	19
Newsletters or bulletins	4	19
Other: Work with Alberta Education	2	10
Other: Information from books or on the web	2	10
Other: Teacher collaboration	1	5

Evaluating Particpants' Understanding and Interpretation of Score Reports

All 21 teachers who completed the questionnaire evaluated the content and format of score report 2. All eleven teachers who agreed to participate in a follow-up telephone interview evaluated at least one additional template, either score report 1 or score report 3. Four teachers evaluated score report 1, four teachers evaluated score report 3, and three teachers evaluated both score reports 1 and 3.

To address research question 2, teachers were presented with three score reporting templates and asked to evaluate the comprehensiblenss and interpretability of the content and format of these templates. These templates were presented to the teachers in order to understand the information sources they currently access and use to make diagnostic decisions about their students . This background information is presented first as it provides a context for which the teachers evaluated the three reporting templates.

Teacher Sources of Diagnostic Information

As illustrated in Table 7, the teachers generally agreed (i.e., item means >4) that more specific information, such as descriptions of skills a student has mastered or has yet to master, item-level results, and reporting performance categories were most diagnostic of student performance. Information such as overall score (item mean = 3.38) and subject-level subscores (item mean = 3.76) were considered to have less diagnostic value. In addition to the information provided from the online diagnostic assessment, teachers drew from multiple information sources to help guide their diagnostic decisions around a student (see Table 8). These information sources for teachers arise mostly from teacher-

designed assessments (100%), classroom observations (95%), interactions (86%), oral assessments with the student (86%), and textbook unit tests (71%), and less so from a standardized large-scale assessment (38%). This finding is consistent with teachers reporting that their knowledge of educational assessment comes mostly from experiences in the classroom.

Table 7. *Teachers' ratings of information to inform diagnostic decisions about students (n=21, % agreement by response option)*

Item	Strongly Disagree	Disagree	Netural	Agree	Strongly Agree	Mean (/5)
Overall (total) score	0	33	14	33	19	3.38
Subject-level subscores	0	14	19	43	24	3.76
Descriptions of specific knowledge and/or skills a student demonstrated on the test	0	5	0	24	71	4.62
Descriptions of specific knowledge and/or skills a student should develop	0	5	0	33	62	4.52
Item-level results	0	0	5	57	38	4.33
Reporting performance levels	0	0	5	48	48	4.43

Table 8. *Other identified sources of information to inform diagnostic decisions about students (n=21)*

Information Source	N	%
Teacher designed tests	21	100
Observation assessments	20	95
Anecdotal information	18	86
Oral assessments	18	86
Classrooom textbook unit tests	15	71
Provincial achievement tests	8	38
Other: projects, labs, peer assessments	4	19

Teachers reported using diagnostic assessment information for a variety of educational classroom activities (see Table 9). Almost all teachers used diagnostic information to plan instruction and select instructional strategies (95%), and select remedial activities (90%). About 3/4 of the teachers surveyed used the information to assess their own teaching effectiveness or to give the information directly as feedback to the student. A little over half of the sample (57%) used diagnostic results for the purposes of a referral so the student could receive additional testing. Other activities included writing questions to review concepts not well understood by the students in the class, and to communicate student performance to parents.

Table 9. *Teachers' classroom activities informed by diagnostic information (n=21)*

Classroom practice activity	N	% of sample
Plan your instruction	20	95
Select your instructional strategies	20	95
Select remedial activities for your students	19	90
Assess your own teaching effectiveness	16	76
Give feedback to your students	16	76
Refer a student for further testing	13	57
Other	2	10

Teacher Perceptions of Diagnostic Assessment Characteristics

In addition to gathering information about what kinds of information would help teachers make diagnostic decisions, the 11 teachers who participated in a follow-up telephone interview were also asked about their perception of characteristics of educational diagnostic assessments. Four themes were derived from an analysis of the participants' responses.

Diagnostic assessment should provide specific and descriptive information about student performance.

Generally, teachers stated that diagnostic assessment should provide detailed and specific descriptions for reporting academic performance. One teacher commented, "*[diagnostic assessment] should have lots of descriptors for*

limited because they have lots of deficiencies and are more variable in their skills as a group." Teachers commented that diagnostic assessments should not report total score or percentage correct as these numbers are neither helpful nor descriptive. Out of the eleven teachers, eight stated that diagnostic assessments should provide information identifying a student's areas of strength and weakness. As one teacher explained, "*Good diagnostics will tell you exactly what the problem is so you know what to work on.*" Within each of the skills assessed, the level of competency should also be reported. This information is perceived to be useful as it helps the teacher know where to start when working with a particular student.

Alternatively, six out of the eleven perceived diagnostic assessment to have a normative function when reporting student performance. These teachers' describe diagnostics as informing them where the student is performing in relation to a peer group, grade level, and/or normal learning continuum. It is a picture of where the student is *right now* and where the student *should be.*

Diagnostic assessment is one of a battery of assessments that can be used to assess and evaluate a student's performance.

Five out of eleven teachers view diagnostic assessment as one tool out of many currently available to help a struggling student. As one teacher remarked, "*The more information and resources, the better to help the child.*"

Diagnostic assessment should provide information to the teacher on instructional effectiveness and guidance.

Five out of eleven teachers state that diagnostic assessments can help the teacher know if they have prepared the student adequately for the next unit. The results of a diagnostic assessment can direct the teachers' instructional focus with each student. In the words of two teachers, "*Diagnostics tells the teacher whether the child has or has not understood the concept. It tells the teacher what to re-teach before moving on.*"

Diagnostic assessment should be used with students who are having difficulties.

A small group of teachers commented on the target population of diagnostic assessment: students who are experiencing difficulties with the subject material. Groups of students who share a similar skill profile can be taught together. One teacher commented, "*Ideally to be used in small groups, combine the online CDA with observation of the students*" with the inference that the combination of assessment and observation helps paint a better picture of why students are experiencing a specific problem.

Evaluation of the Score Reports

Teachers' evaluation of the comprehensibleness and interpretability of the content and format of score reports 1, 2, and 3 were obtained. The results of the evaluation for score report 2 are presented first followed by score report 1 and score report 3.

Score report 2.

Twenty-one teachers were asked the degree to which the information reported and presented was understandable and useful for communicating information about student performance. Table 10 shows that the teachers score report 2 positively. Overall, the teachers agreed that the information contained with the checkmark version was easy to interpret (mean=4.43), understandable (mean=4.23), useful (mean=1.76), and the level of reporting was appropriate for communicating diagnostic results (mean=4.23). The teachers rated the visual formatting of the score report favorably with the information being perceived as well-organized (mean=4.38), presented clearly (mean=4.38), and visually appealing (mean=4.33). The amount of information was rated as being sufficient (mean=3.95) and not too much (mean=1.71).

During the follow-up interview, as part of the evaluation, teachers were also asked to comment on four aspects of the checkmark template: 1) most useful information, 2) least useful information, 3) information that the teacher would like but is missing, and 4) suggestions to make the score report more informative or useful.

Table 10. *Evaluation of format and content: Score report 2 (% agreement by response option)*

Item	Strongly Disagree	Disagree	Netural	Agree	Strongly Agree	Mean (/5)
1. Information reported is easy to interpret	5	0	0	38	57	4.43
2. Information is not useful	52	38	0	0	10	1.76
3. Performance level descriptors are appropriate for reporting diagnostic results	5	5	0	43	48	4.23
4. The language used in the student score report is understandable	5	5	0	43	48	4.23
5. Too much information is presented in the student score report	33	62	5	0	0	1.71
6. The amount of information presented in the student score report is sufficient	5	5	0	72	19	3.95
7. The student score report is well-organized	5	0	0	43	52	4.38
8. The information is presented clearly	5	0	0	43	52	4.38
9. The score report is visually appealing	5	0	5	38	52	4.33

Most useful information. Four out of eleven teachers identified having a description of the skills on the assessment and how these skills are tied to specific outcomes in the curriculum as being most useful. Three out of eleven teachers stated the skills identified as needing additional work was helpful. Two teachers found the hierarchical ordering of the skills from most cognitively simple to complex was most useful. Another two teachers stated the reporting categories of student performance as demonstrating consistent, moderate, or limited evidence of mastery most helpful.

Least useful information. When asked to identify what information on the score report was least useful, all eleven teachers stated they could not identify anything as all information within the score report was perceived as useful.

Information that the teacher would like but is missing. Of the eight teachers who responded to this question, the following information was thought to be useful and informative of the student's performance on the assessment. Three teachers thought that having access to the correct answers and questions would be nice to link with the skill descriptions described in the report. Two teachers stated they wanted more information on how the student was classified into the performance categories and not just a description of the categories themselves. These teachers wanted to relate the number of questions a student answered correctly for each attribute to whether or not the student was classified as demonstrating consistent evidence of mastery. One teacher identified having the date available on the each report generated. If the assessment is to be used formatively, over time, then times between administrations would be important to have to gauge student growth. Related to the date, a record of how long a student required to complete the assessment may also provide a useful context for understanding whether the student rushed through, and was therefore classfied as limited, or if the student took adequate time and still struggled. Other information identified as good to have include a denominator when reporting total score (i.e., 8 out of 15).

Suggestions on how to make the report more informative and useful. Of the five teachers who responded to this question, the following information was thought to make the report more informative and useful. Specific to the information within the score report, two teachers wanted more explanatory information to make the classification process clear and to indicate a percentage

range that corresponds to each category. Another teacher would have found it helpful to have the reporting categories match the category descriptors used in their own report cards. Two teachers made suggestions related to better support for their instruction in the classroom by asking for a classroom-level report of student performance and by providing teachers with resources on how to address deficiencies idenitified in the score report at the class level.

Unclear terms. To evaluate the clarity of the language used within the score report, teachers were asked to idenitify any unclear terms. One teacher commented on the the word "in-depth" and how does one measure or know if a student has "in-depth knowledge and understanding" of the subject matter. Another teacher was unsure whether the term "category descriptor" referred to the skill descriptions or the skill performance categories. To enhance clarity and consistency in understanding the skill performance categories, one teacher suggested having the category labels of "consistent" and "moderate" changed to the terms used with the existing provincial achievement tests.

Intepretive exercise. Given the generally positive comments about score report 2 with the presentation and ease of interpretability, teachers who participated in the follow-up telephone interview were asked to engage in a hypothetical parent-teacher interview where he or she had to explain the score report to the parent. In their explanations, most teachers described the student's performance in comparable ways by emphasizing the pattern of mastery of skills, with the easier skills at the bottom of the hierarchy and more difficult skills at the

top. One teacher's explanation that is illustrative of the other responses is provided below:

> *So Mrs. Doe this past little while we were looking at number sense within the mathematics, within math. And the number sense is broken down into five different areas. Now we can see and how we noted it was the checkmarks. And the consistent means. And I'd say that consistent means this, the moderate means this. And I would try to say he needs a lot of work, to put it in simpler terms for them that they're doing OK, or did it very, very good. And then they can see where the checkmarks are. And I say, your child needs some extra help in problem solving, from 100 to 1,000. And then they would probably say, well what kind of questions are those? And I would try to give 'em examples of what that is. Higher level, I would say their basic math is good, and I'd be that, I would show them increasing skill level. I would say the basic math is good, but as the higher the skill level gets, they're more limited.*

Some teachers mentioned the importance of having some context, not just explanation, for the assessment results. For example, it is important to note whether the child was away or ill when the assessment was administered. What time of year was the child tested and do the results reflect a child learning the material versus. mastery?

Score report 1.

In the follow-up telephone interview, a subset of seven teachers were asked the degree to which the information reported and presented in score report 1

was understandable and useful for communicating information about the students performance. Table 11 shows that the teachers rated score report 1 positively. Overall, the teachers agreed that the information contained within this version was easy to interpret (mean=4.14), understandable (mean=4.71), useful (mean=1.71), and the level of reporting was appropriate for communicating diagnostic results (mean=4.00). The teachers rated the visual formatting of the score report favorably with the information being well-organized (mean=4.43), presented clearly (mean=4.29), and visually appealing (mean=4.14). The amount of information was rated as sufficient (mean=4.14) however the teachers were not sure either way with whether there was too much or too little information reported (mean=3.14).

Table 11. *Evaluation of format and content: Score report 1 (n=7, % agreement by response option)*

Item	Strongly Disagree	Disagree	Netural	Agree	Strongly Agree	Mean (/5)
1. Information reported is easy to interpret	0	0	14	57	29	4.14
2. Information is not useful	57	29	14	0	0	1.71
3. Performance level descriptors are appropriate for reporting diagnostic results	14	0	0	43	43	4
4. The language used in the student score report is understandable	0	0	0	29	71	4.71
5. Too much information is presented in the student score report	0	29	29	43	0	3.14
6. The amount of information presented in the student score report is sufficient	0	14	0	43	43	4.14
7. The student score report is well-organized	0	0	0	57	43	4.43
8. The information is presented clearly	0	0	14	43	43	4.29
9. The score report is visually appealing	0	0	14	57	29	4.14

During the follow-up interview, teachers were also asked to comment on four aspects of score report 1: 1) most useful information, 2) least useful information, 3) information that the teacher would like but is missing, and 4) suggestions to make the score report more informative or useful.

Most useful information. Three out of seven teachers identified having the skill profiles presented in the form of a graph as being the most useful information. Two teachers found the skill descriptions the most informative, whereas the remaining two teachers thought that knowing item-level results (i.e., correct answer, student answer) was the most helpful.

Least useful information. What was identified as the most useful for some teachers was identified as the least useful by others. In particular, item-level

results such as correct answer and the student answers were perceived as unhelpful by some respondents. One teacher commented that test items should accompany the score report if item-level results are reported. Another teacher thought the column of correct answer was unecessary since a teacher should know what the correct answer is. One teacher thought that the instructions on "how to read the report" wasn't helpful because how to read and interpret the report was self-explanatory. Another teacher thought the "Summary of Performance" as presented was unnecessary as it was too much information given the results were already presented but in a different form.

Information that the teacher would like but is missing. One teacher identified similar shortcomings with score report 1 as with score report 2. Namely, this teacher wanted to have the date of administration and time of day reported, as well as highlighting the curricular outcome tested.

Suggestions on how to make the report more informative. Three teachers offered their comments on improving score report 1. Two teachers' comments were centered on formatting changes to reduce clutter and improve presentation (e.g., reduce the number and move columns around, make graphs darker). One teacher reiterated wanting to have a classroom report generated for her use.

Score report 3.

In the follow-up telephone interview, a subset of seven teachers, three of whom also evaluated score report 1, were asked the degree to which the information reported and presented in score report 2 was understandable and useful for communicating information about the students performance. As with

the other reporting templates, teachers rated score report 3 positively (see Table 12). Overall, the teachers agreed that the information contained within this version was easy to interpret (mean=4.14), understandable (mean=4.57), useful (mean=1.71), and the level of reporting was appropriate for communicating diagnostic results (mean=4.86). The teachers rated the visual formatting of the score report favorably with the information being well-organized (mean=4.71), presented clearly (mean=4.86), and visually appealing (mean=4.29). The amount of information was rated as being sufficient (mean=4.86). In contrast with score reports 1 and 2, teachers did not think that too much information was presented in score report 3 (mean=2.29).

Table 12. *Evaluation of format and content: Score report 3 (n=7, % agreement by response option)*

Item	Strongly Disagree	Disagree	Netural	Agree	Strongly Agree	Mean (/5)
1. Information reported is easy to interpret	0	0	14	57	29	4.14
2. Information is not useful	57	29	14	0	0	1.71
3. Performance level descriptors are appropriate for reporting diagnostic results	0	0	0	14	86	4.86
4. The language used in the student score report is understandable	0	0	0	43	57	4.57
5. Too much information is presented in the student score report	43	14	29	0	14	2.29
6. The amount of information presented in the student score report is sufficient	0	0	0	14	86	4.86
7. The student score report is well-organized	0	0	0	29	71	4.71
8. The information is presented clearly	0	0	0	14	86	4.86
9. The score report is visually appealing	0	14	14	0	72	4.29

During the follow-up interview, teachers were again asked to comment on four aspects of score report 3: 1) most useful information, 2) least useful information, 3) information that the teacher would like but is missing, and 4) suggestions to make the score report more informative or useful.

Most useful information. Four out of seven teachers identified having the skill profiles presented in the form of a graph as being most useful information. One teacher found the skill descriptions organized hierarchically as the most informative. One teacher thought that knowing item-level results (i.e., student

answer) was the most helpful. And one teacher thought the interpretive material accompanying the report was the most useful information.

Least useful information. Six out of seven teachers identified aspects of the score report that was the least useful. Similar to what was seen with score report 1, four teachers thought item-level results such as correct answer and the student answers were perceived as unhelpful, especially without access to test items. One teacher thought listing item numbers was redundant since items were administered in increasing difficulty. Another teacher believed reporting total correct such as “Your score is 8” didn’t tell him/her very much in comparison to reporting skill profiles.

Information that the teacher would like but is missing. For this sample of teachers, no information was identified as missing.

Suggestions on how to make the report more informative. Where most of the teachers felt the report was informative, one teacher’s suggestion on improving the score report centered, again, on formatting of the report. For example, it was suggested to increase font size, put bigger graphs, and decrease the appearance of clutter on the page.

Teachers who evaluated score report 3 were also asked to comment on the content, format, and comprehensibleness of the interpretive material attached to the first page of the score reporting template. Overall, the teachers agreed that the information was helpful and good to reference periodically when needed. As one teacher commented:

Information on the first page is self-evident. Second page just says in words what is said graphically or in pictures on the first page. Good to have if that is your preference.

One teacher understood the purpose of the interpretive material, which was to enhance understanding and consistent interpretations of the information within the report.

I think it's reiterating what, it gets rid of any room for interpretation because maybe you look at it one way and thought something wrong. If you look at both of these pieces of information, then you should have a very clear picture of what your child is at and where they need to go.

One teacher also suggested that the interpretive material be provided once at the beginning and not with every student report.

As with the score reporting templates, a few teachers felt the writing could have been more concise and that organization of the material on the page could be improved. When asked whether the teacher referred to it when reading the first page of the score report, many of them stated they only glanced over the information because the information on the first page was self-explanatory.

Summary of Evaluative Comments Across Templates

Looking across the results of the responses to the questionnaire items designed to evaluate the content and format of the three score reporting templates, in general, the templates were all rated positively. The teachers agreed that the information was easy to interpret, useful, comprehensible and reported at an appropriate level of specificity for a diagnostic test. In general, the teachers agreed

that the information provided was sufficient, well-organized, presented clearly, and visually appealing. Table 13 summarizes the item means across the three reporting templates.

Table 13. *Comparison of item means (/5) evaluating content and format across the three reporting templates*

Item	Score report 2 n=21	Score report 1 n=7	Score report 3 n=7
1. Information reported is easy to interpret	4.43	4.14	4.14
2. Information is not useful	1.76	1.71	1.71
3. Performance level descriptors are appropriate for reporting diagnostic results	4.23	4	4.86
4. The language used in the student score report is understandable	4.23	4.71	4.57
5. Too much information is presented in the student score report	1.71	3.14	2.29
6. The amount of information presented in the student score report is sufficient	3.95	4.14	4.86
7. The student score report is well-organized	4.38	4.43	4.71
8. The information is presented clearly	4.38	4.29	4.86
9. The score report is visually appealing	4.33	4.14	4.29

Item 5 showed more variation in the ratings, when compared to other item means, asking participants to rate the amount of information presented in the score reporting template. When comparing the three templates to one another, the amount of information provided in the report increased from the score report 2, to score report 1, to score report 3. Looking at the teacher responses provided during the follow-up interview helps to understand why the ratings may have varied. In one way, it is possible to conclude that the more information provided in the

diagnostic testing context, the better. So, ratings of disagreement with the statement “Too much information is presented in the student score report” makes sense. The main comment teachers made that impacted their rating on whether too much information was presented was *it all depends on how you’re going to use it*. For example, if the report is mostly for the purpose of getting a quick idea of what the student has learned during a specific instructional unit, then the checkmark template (score report 2) will suffice. But, if more diagnostic information is desired for the student who is struggling, then identifying a pattern of responses with either score report 1 or 3would be more useful.

Additional comments were collected from the teachers about the content and format of the score reporting templates. Three main themes emerged from the comments. The first theme was that teachers are very busy in the classroom. Therefore, when considering what to report and how to report it, an understanding of the context in which teachers work should be considered. As one teacher remarked, “*Sometimes all you have time for is a quick look.*” The second theme was that having more information about a student is better than having less information, since the goal is to improve a student’s knowledge and skills. This theme is noteworthy because the participants saw value in having many different kinds of information, but it is not clear how these pieces would be brought together and interpreted to arrive at a cohesive picture of a student’s performance. In other words, what does a teacher do with all this information?

Since all the teachers who participated in the follow-up telephone interview evaluated the checkmark template first and then an alternate template

(i.e., score reports 1 or 3), teachers' overall comments on the content and formats of the templates were made relative to each other. Many teachers commented that the checkmark version was clear and facilitated, at a glance, reading when compared to the more detailed score report templates. As one teacher commented, "*I guess the checkmark one is the quick here you go this is what I did, and the second one is a lot more in depth.*"

Third, there appeared to be some uncertainty as to who was the intended audience of the reporting templates 1 and 3. Teachers commented that the language used in the "Summary of Performance" was child directed with a positive tone. But, when looking at the other sections of the score report, such as "How to read this report", teachers assumed that they were also the target audience. Two teachers commented that the language used in the score reports could possibly be understood by older elementary children but not with Grade 3 students, without some assistance, "*The language is appropriate for Grade 6 students maybe, but not grade 3."*

Uses of and Preferences for Diagnostic Information

The eleven teachers who participated in the follow-up telephone interview were asked to reflect on how a teacher could use the information in the score reports. Some teachers provided an example of how he or she actually used the information because there was an opportunity to do so as part of the larger Alberta Education Diagnostic Mathematics Project. Given the overall positive reception of the score reports, all eleven teachers perceived the information in all three templates to be useful for a Grade 3 teacher. The respondents were asked how a

teacher could use the information in their classroom practice. Four general tasks were identified based on the responses: communication with parents and students, informing the next instructional steps, evaluation of the student's current level of learning, and incorporating the results in summative reporting.

Uses of Information

Score reports as a communication tool.

All the teachers commented that the score report would be a useful tool for communicating a student's learning and progress in math. When asked whether the teacher would consider sharing the information within the report with parents and students, the answer was a resounding yes, with the condition that the teacher provide interpretive support. The teachers were hesitant to send the score reports home, as presented in this study. Overall, the teachers felt that the report needed additional explanation, that parents may not understand or might misinterpret the information in the report, and would be confused with what to do with the information. Sending home the score report, as is, could have unintended consequences, as described by one teacher:

> *I think sometimes you have to be careful because they might not, what I hesitate on is that some lesser educated parents not understand what it means. So all they'll look at is oh you got all limited, so that means you're dumb. When they don't really understand maybe that we haven't covered that or that, you know, this particular one everybody did poorly on. You know what I mean? Without being able to kind of justify it and explain it, I*

might hesitate to just send it home cold turkey without them having any information, background information.

In the opinion of the teachers, the best venue for sharing the reports was in a parent-teacher interview. Given that multiple outcomes may have been tested during a reporting period, teachers thought they would pull out one score report to show a parent as an example of the kinds of skills being assessed for one outcome instead of presenting multiple score reports. One teacher stated that sharing the reports can highlight both the weaknesses and the strengths of the student:

I think if you're showing it in parent teacher interviews for example, you can say it's a positive comment you know. The parent can sense how much their child's struggling in math, there's something on the report that shows OK, well they're doing really well in skip counting. I just think it gives that positive in the child.

Teachers were less likely to show the actual score report to the students because they felt the students wouldn't understand the diagnostic information without the teacher's and/or parent's assistance.

I think if you're gonna sit down and go through it with them, you'd have to kinda talk in more kids terms. You could say well I'm noticing you might have had a little bit of difficulty with your counting by 25 backwards, did you understand the question? Then it allows you to see did the question confuse them on the diagnostic? Maybe they need to just show you on the paper, pencil right. But then clearly if they can't do it there and they can't do it with you it sort of gives you that double check type of thing.

Some teachers felt that parts of the report could be shown to the student – for example, the checkmarks figure on the checkmark template. One teacher stated she did show the checkmark chart with her students and it had a motivating effect on them. The students may not have understood what consistent, moderate, or limited evidence of skill mastery meant. But the students knew what pattern of checkmarks they wanted on the chart – all checkmarks aligned in the left column.

Informing the next instructional steps.

Eight of the eleven teachers claimed that the information in the score report could be used to focus and guide instruction. As four teachers commented, "*you know where to re-teach.*" As part of informing the next instructional steps, five teachers commented that the assessment results could be used to create and work with small groups enabling differential instruction. One teacher summarized these points in her statement:

> *It shows them which hierarchy they need to focus on. And then I think it also allows you on the other end if you've got children who understand everything, I think that that allows them to not, it's not wasting your time re-teaching the concept but if I've got five kids that already know everything there is to know about skip counting, then I can take them and we can do a challenge activity.*

Evaluating the student's level of learning.

Five out of the eleven teachers stated that the information provided from the diagnostic assessment is but one piece of information considered when evaluating the student. Consistent with previously identified sources of

information to inform diagnostic decisions, and the sentiment that more information is better than less when working with diagnostics, the theme of the score reports offering one piece of the student picture is sensible. The results are an additional piece of objective evidence to support and verify the teachers' evaluation and understanding of the student's performance. The comment that the results are "objective" and not the outcomes of a teacher created test, is considered to be a positive characteristic in the teachers' opinion when communicating student achievement to parents.

Using the results as part of summative reporting.

Four out of the eleven teachers stated that the information could be incorporated into their marks for the student's report card. One aspect of the diagnostic assessment and reporting was the link between the skills tested to the specific outcomes in the curriculum. If the teacher was working in an outcome-based reporting environment in his or her district (i.e., reporting student achievement with respect to specific outcomes taken from the curriculum), then the results of the diagnostic assessment aligned closely with an existing reporting framework. Two of the teachers expressed some hesitation with incorporating the diagnostic results in report cards. The hesitation came because these teachers thought the primary purpose for administering the diagnostic assessment was formative, not summative. Using the results in a summative way would go against the underlying principles used to development the diagnostic assessment. The other two teachers stated that they did enter the results in their markbook and incorporated them into their report card grade for the student.

Potential for misuse of information?

The literature on score reporting warns test developers to anticipate any potential misuse or misinterpretation of the information contained within the score report. In the follow-up interview, teachers were asked if they had any concerns with misuse of the information reported from the diagnostic assessment. Most concerns stemmed from teachers expressing the importance of communicating the purpose of the assessment, including what is being assessed, how it is being asssessed, when the assessment should be used, and what reasonable interpretations can be made. The most common concern identified in the current study was related to inappropriate use of the information. The respondents focused on using the results of the diagnostic for instructional decision making and not as a final or summative assessment for the particular unit. Two teachers commented on who would have access to the results – would principals and superintendents have access to the results? The concern centered around using the assessment results to evaluate the effectiveness of the teachers. Last, two teachers expressed concern with misinterpretation and potential misuse of the information by parents, if the teacher did not explain the information adequately to the parent. An example provided by one teacher was a parent using the information to compare children within the classroom or to intepret the results harshly (e.g., the child is dumb because he or she only scored limited across the skills tested).

Information Preferences for Diagnostic Score Reporting

Audience preferences play a role when considering what should be reported and how the information should be presented. Audience preference plays

a large role in determining how the information should be disseminated among the target audience as well as when making changes to some aspect of the score report – format, content, or both.

Aggregate level reporting.

As mentioned previously, a number of participants commented that teacher workload and workflow should be considered when developing the score report. Teachers are busy and, in reality, have little time to leaf through multiple, detailed, diagnostic score reports for one student, let alone for a classroom of thirty students. It would be prudent to consider ways to aggregate the results to illustrate performance trends and to summarize performance across strands or across students. When asked "For diagnostic purposes, do you want a report like the sample report for each of your students, for each strand, for an entire subject area? (e.g., you would receive skill-level results summarized for Number, Shape & Space, and Data and Probability)", 18 out of 21 teachers surveyed (86%) replied with "Yes". When asked "For diagnostic purposes, do you want a report like the sample report at the classroom level, for each skill hierarchy?", 20 out of 21 teachers (95%) replied with "Yes". Based on this finding, it appears that aggregating the results for both a classroom-level report and for summarizing student performance across strands would be worth developing.

Mode of dissemination.

For the purposes of the diagnostic mathematics field test, the assessments were scored and reports were generated external to Alberta Education. The score reports were created as a PDF file and then distributed electronically to the

teachers for their review. For future implementation, teachers were asked their preference for mode of disseminating score reports. Eleven out of twenty-one teachers (52%) preferred to have a print-based score report with an accompanying website. Six out of twenty-one teachers (29%) preferred to have a web-based only reporting system and the remaining four teachers (19%) preferred to have a print-based option only. Providing a combination of a print-based option with web-based resources for reporting is the most flexible option, particularly for teachers who who may prefer one mode over the other.

Which template was preferred?

The eleven teachers who participated in the follow-up telephone interview were asked to comment on which reporting template was preferred. For implementation purposes, which score report should be used in an operational setting? After evaluating at least one alternate template, five teachers stated they still preferred the checkmark template. The checkmark template was viewed as "*quicker and easier to read, the right level of detail for what you need, not too much reading*". Five teachers stated they would like some sort of combination of the checkmark template with aspects of score reports 1 and 3. One teacher illustrated a potential combination of the checkmark template with score report 3:

> *You know again, the first look would be at the big checkmark template and if I see that we have issues then it's come to this one (score report 3). So that first one would be a good one to say you know we'd have to do something and then we'd become more and more specific.*

One teacher stated a preference for score report 3 because of the greater detail in comparison to the checkmark template. Based on these results, it appears that the checkmark template is still useful overall, but that revisions to the content and format should be explored. These revisions could involve specific changes to the checkmark score reporting template or incorporating a two-level score report where score report 1 or 3 serves as a more detailed explanation of the information in score report 2.

Refining the Reporting Framework

For this study, evaluation of the score reports was completed focusing on the content, format, and interpretability of the information . As a reminder, the score reporting framework outlined what aspects to consider with score reporting, including: what to report, how to report this information, and mode of dissemination with the goal of clarity of communcation with application of design principles. The process of creating score reports that were guided by this framework can be thought of a top-down approach. In other words, what things should be considered when developing the first draft of a score report? Based on the follow-up interviews, one more category should be added into the reporting framework: reporting context.

It seems intuitive that the reporting context should be considered, but it is not clear how or where exactly it should be incorporated into the development of the score reports and to what degree it will influence decisions on other aspects of the score report such as content and format. Based on the teachers' comments made in this study, it is clear that they evaluated the score reports with varying

backgrounds representing different geographical regions of Alberta, different district cultures, different levels of teaching experience, and different teaching philosophies. Given this variability, it is unlikely that one general score report can be created to meet the needs of all teachers equally well. What might be useful is to identify salient aspects of the context at a particular level (e.g., district level) and then make decisions around content, format, and mode of dissemination within those contextual boundaries. There may be variation among schools within a district, but a district can be thought of having their own common culture.

An example of how context shapes preferences for reporting information was illustrated by an administrator who took part in this study. In his particular district, the emphasis is on the incorporation and use of technology in their classrooms. Additionally, this district uses outcome-based reporting of achievement. Given that the diagnostic assessments are computer-based and aligned to the curriculum, the results of these assessments being disseminated online and in their current form was viewed as already working well within this district's particular context. Developing score reports that are communicative and meaningful to the target audience should incorporate user input in an iterative process. By defining the contextual boundaries, one could potentially develop a more informationally efficient and targeted score report.

CHAPTER 6 DISCUSSION AND CONCLUSION

The recent emphasis on understanding the psychology underlying test performance has led to developments in cognitive diagnostic assessment (CDA; Leighton & Gierl, 2007). CDA attempts to integrate cognitive psychological principles with educational measurement practices for the purposes of enhancing learning and instruction. The results of a CDA yield a profile of scores with specific information about a student's cognitive strengths and weaknesses. Score reporting serves a critical function as the interface between the test developer and a diverse audience of test users. Effective reporting of diagnostic assessment results is important because teachers can look to these results to help guide their instructional practice, parents often seek information on ways to help their children in identified areas of academic difficulty, and students seek feedback to validate their study and testing efforts. In short, cognitive diagnostic feedback may be used by instructors, parents, and students to guide and monitor their teaching and learning processes.

Research studies on score reporting have noted that the communication between test developers and users of educational tests is weak and requires improvement. This is evident by teachers receiving student test results too late to influence instruction, typically many months after the test administration (Huff & Goodman, 2007). Further, information is reported in ways that are difficult to read and understand (Ryan, 2003) and often without adequate supporting material to promote clear test score interpretations (Trout & Hyde, 2006). Large variability

exists in how test scores are reported to the public on educational tests (Goodman & Hambleton, 2004; Knupp & Ainsley, 2008).

As developments in CDA continue to progress, the need to address score reporting issues of comprehensibility and interpretability becomes even more pressing. Diagnostic testing information, including skills descriptions and learning concepts, is fundamentally different in purpose from information typically reported from traditional large-scale assessments, such as total number correct scores or percentile ranks. Test developers must report and present new kinds of information from these diagnostic tests. In short, the challenge of diagnostic score reporting lies in the integration of the substantive and technical information needs of the educational community with the psychologically sophisticated information unique to CDA.

Currently, there are few examples of cognitive diagnostic score reports. One operational example is the College Board's *Score Report Plus* for the PSAT/NMSQT where cognitive diagnostic feedback is given in the form of a description of the top three skills requiring improvement for each content area of Mathematics, Critical Reading, and Writing along with recommended remedial activities. As developments continue to progress, current score reporting approaches need to be recast in light of the new kinds of information yielded by CDA and the context in which this information is to be used by its target audiences.

Therefore, the purpose of this study was to create a framework for providing a structured and systematic approach to developing score reports. Then,

using this framework, score reports were created for an operational cognitive diagnostic assessment in mathematics developed using the Attribute Hierarchy Method (AHM, Leighton, Gierl, & Hunka, 2004). This study described the development and creation of the student score reports from a reporting framework and the results of a small scale evaluation with teachers using a questionnaire and semi-structured telephone interviews. Presently, there are only a small number of operational examples that illustrate the development and evaluation of student score reporting in the context of a cognitive diagnostic assessment. Hence, research in this area is sorely needed.

This chapter is organized into four sections. In the first section, the purpose of the study, research questions and an overview of the methods used to answer the research questions are provided. In the second section, a summary of the results of the study organized by research question is presented. In the third section, the limitations of this study are discussed. And in the fourth and final section, future directions for research are outlined.

Restatement of Research Questions and Summary of Methods

The main purpose of this study was to create a framework for promoting a structured approach to developing diagnostic score reports. Then, using this framework, another purpose of the study was to create and evaluate the effectiveness of the score reports for communicating diagnostic results for an operational cognitive diagnostic assessment. The context for this study is the Cognitive Diagnostic Mathematics Assessment (CDMA) project funded by the Learner Assessment Branch at Alberta Education in Edmonton, Alberta, Canada.

CDMA is a curriculum-based set of assessments that can be used throughout the school year to measure students' thinking and learning in mathematics.

Three research questions were addressed in this study:

1. What should be reported on a diagnostic score report? and How should this information be presented?

2. To what extent are the score reports understandable and interpretable by teachers, the target audience? What are teachers' understandings of the diagnostic information in the score report? What information do teachers consider to be diagnostic and why?

3. Given the answers to questions 1 and 2, how useful are the score reports? How could the information in the score report be used? Is it clear what the next instructional steps should be for the student?

To answer these research questions, a combination of a critical literature review, questionnaires and telephone interviews to evaluate the score reports were completed. Next, I will summarize aspects of the methods used in the study.

A critical review of score reporting in education was completed. This literature identified issues with score reporting including difficulties with report readability and comprehension, which can often lead to inferences not supported by the information presented. Additionally, score reports were not disseminated in a timely manner, limiting their usefulness to inform instruction and guide student learning efforts. This literature also identified problems with the presentation of information and general appearance of score reports. There were few studies that looked at both the reporting elements and the effectiveness of their presentation

with educational stakeholders. Next, a review of the information design literature was completed to identify general design guidelines to assist with the communication of test score results. This involved a review of designing effective text using typography and layout, as well as how to design effective quantitative displays of information such as tables and graphs.

The reporting frameworks available in the educational measurement literature were adapted for reporting diagnostic scores. A review of the AHM as a method for cognitive diagnostic assessment was provided to establish a context for illustrating diagnostic score reporting. Each reporting element in the framework was aligned to the specific AHM outcomes focusing on attribute probabilities as diagnostic scores. To illustrate the application of the framework, three sample score reports presenting student-level diagnostic information were created and presented following the recommendations put forth by researchers in educational measurement and information design. More specifically, the diagnostic score reports were personalized, contained basic interpretive information, provided a quick visual summary of student performance, and outlined how to use the information to guide study efforts. Although three example reports were presented here, the reporting framework provides a structured approach to developing multiple alternative reporting forms that can then be piloted with target audiences.

Creation of the score report occurred in three stages. First, three different presentations for the score report were developed based on a diagnostic reporting framework. One of these templates was then used by the project team at Alberta

Education in an early field test administration. Second, the reporting template was internally evaluated by members of the project team. Based on their feedback, revisions to the score report were made with regards to reducing the amount of information on the report, presentation of scores, and the use of color. Third, the revised score report, and two alternate templates, were evaluated with a group of Grade 3 teachers to help guide decision making around the preferred score report for use in the operational version. As part of this evaluation, questionnaires and semi-structured telephone interviews were conducted to collect data on teachers' understandings and information expectations of a diagnostic assessment, the potential uses and misuses of the information, and the actual/perceived utility of the score reports for informing a teacher's next instructional steps.

Summary of Results

The primary purpose of this study was to create a framework for promoting a structured approach to developing diagnostic score reports. A second purpose of this study was to create and evaluate three score reports developed from the reporting framework for communicating diagnostic results for an operational cognitive diagnostic assessment. Specifically, this study answered three research questions:

Research Question 1: What should be reported on a diagnostic score report? How should this information be presented?

Initial considerations around the content and format of the diagnostic score report were guided by the reporting framework developed from a review of the educational measurement and information design literature. Three main questions

were considered when developing the score reports: 1) What should be reported? 2) How should the information be presented? and 3) What is the mode of dissemination of information? The framework describes the potential combination of what could be reported from a diagnostic test. Application of the framework in the context of the Diagnostic Math project resulted in three score reporting templates. The content and format of these score reports also incorporated feedback from the project team through an internal review. Decisions were based on consideration of what basic information a classroom teacher *may expect* to have on a diagnostic score report.

For the score reports used in field testing (i.e., score report 2), student performance was reported in terms of the evidence of skill mastery demonstrated based on his or her responses: limited, moderate, and consistent. Classification into each category was determined as follows: “Limited” is associated with probabilities below 0.50, “Moderate” is associated with probabilities between 0.50 and 0.80, and “Consistent” is associated with probabilities of over 0.80. Checkmarks were used to denote the category into which the student’s response was classified. Last, a short description of each classification category was provided for the reader’s reference. Other general features include the Mathematics strand reported, the specific outcome tied to the curriculum and the skill descriptors. The two alternative score reports (i.e., score reports 1 and 3) developed using the framework and then evaluated by Grade 3 teachers varied from the score report used in the field test in three distinct ways: 1) the method for reporting diagnostic scores, 2) the inclusion of additional sources of diagnostic

information such as item-level results, 3) and the inclusion of a separate interpretive guide.

The overall impression was very positive for score report 2. Reporting diagnostic scores in terms of performance categories was simple, effective and facilitated “at a glance” reading and interpretation of the report. The one-page layout of the score report using colored checkmark symbols to denote classification was viewed as easy to read and follow. In the opinion of the teachers surveyed, they thought the content of score report 2 was sufficient for reporting and interpreting the outcomes of the diagnostic test. Score reports 1 and 3 with their additional information in the form of tables, graphs, and written text, were received with mixed success. Whether a teacher preferred the more detailed score report or a combination of the field test version and alternate version depended upon the context within which the teacher was working. This context included factors such as the district and school culture, teacher workload, and teacher individual preferences. While considerations of a teacher’s busy schedule can be kept in mind when making decisions around the amount of information to report, reconciling different user preferences for format and presentation when creating the score reports may be difficult, if not impossible, to do.

The interpretability of the alternate templates was influenced by both the additional content (i.e., reporting of multiple sources of diagnostic information) and alternate presentation of scores in the form of bar graphs. In order to fit additional content while keeping the score report relatively short at two pages, certain formatting decisions needed to be employed with respect to the proportion

of white space on the page and font size. In the opinion of the teachers, the presentation of information in the alternate templates was “busy”, making it difficult to find the information they were looking for and making them less likely to read all sections of the report.

Based on the feedback received on the score reports, the possibility of providing two score reports should be considered. One summary version facilitates “at a glance” reading and if desired, the teacher can access a second, more detailed report providing item-level results to illuminate the information reported in the summary version. Overall, teachers preferred to minimize the amount of reading required to interpret the score report. Teachers liked visual representations of the diagnostic scores especially if it can communicate patterns of performance clearly and succinctly.

Research Question 2: To what extent are the score reports understandable and interpretable by teachers, the target audience?

Overall, the teachers found the score report used in the field test to be understandable and interpretable. Teachers did not identify any unclear terms stating the wording used in the score reports were consistent with what is used in the Program of Studies. Some teachers stated they wanted more information to help them to understand how a student was classified into their respective performance categories. Understanding *how* the student demonstrated consistent, moderate, or limited evidence of skill mastery was important to these teachers for illustrating what a consistent, moderate, or limited student looked like. Without a clear understanding of *how* and not just *what* the labels of consistent, moderate,

and limited means, a teacher may impose their own views of what these category labels mean to them. Addressing the how question brings to bear issues related to scoring diagnostic tests. While this information can be made available, its technical nature may preclude it from being accessible and understandable to the teacher population.

With regards to the alternate reporting templates, difficulties with interpreting diagnostic scores presented in the form of graphs were found. Interpretation of the graphs was difficult for some teachers because of unfamiliarity with scores presented as skill probabilities. All the teachers who participated in this study received and were familiar with diagnostic scores reported with reference to performance categories. By reporting skill probabilities, an extra level of complexity was introduced relating to understanding what a skill probability is, how is it calculated, and how it relates to whether a student has the skill being measured or not. While the specificity and accuracy of the diagnostic score can best be captured in a bar graph, interpretation of the bar graphs was perceived to be more cumbersome and confusing.

What are teachers' understandings of the diagnostic information in the score report?

As part of the study, teachers were asked to interpret the information within the field test score report and to explain the results as if they were with the parent. On the whole, the teachers grasped the important concept that the skills were hierarchically organized in increasing cognitive complexity and that interpretation of student performance in light of this ordering was necessary. For

example, if the pattern of mastery was erratic across the three categories, then this provided the teacher with some evidence that the student may not have been engaged in the assessment or that the student may have been guessing. Teachers also interpreted the results appropriately in the context of when the assessment was administered. For example, if the assessment was administered before the content was covered as compared to administration afterwards, then the expected pattern of mastery should be different.

What information do teachers consider to be diagnostic and why?

Results of the small scale evaluation revealed that teachers drew on varied sources of information to help them make a diagnostic decision. In general, diagnostic reporting of the knowledge and skills a student demonstrated and should develop as well as reporting in terms of performance categories was effective. Some teachers thought that total score and content-based subscores also provided them with adequate information to help make a diagnostic decision about a student. The main idea reported by the teachers was that having more information is not necessarily better, but in the case of trying to diagnose student problems, *any* information is better than nothing. Some debate remains on whether reporting a total score with skill level results should be done given that the assessment is meant to be formative in nature.

Research Question 3: Given the answers to questions 1 and 2, how useful are the score reports? How could the information in the score report be used? Is it clear what the next instructional steps should be for the student?

Regardless of which score report was being evaluated, either the field test version or the alternate templates, all the teachers in the study felt that the score reports were useful. The degree of usefulness was influenced by contextual factors such as intended use of the information and preference for certain kinds of information (e.g., item-level results; having access to test items). Overall, this group of teachers were interested in the results of the diagnostic assessment because it was viewed as supporting several important uses including: 1) communicating student performance to parents and to a lesser extent, the students themselves, 2) differentiating their instruction, and 3) identifying and/or verifying where the student is at with their math performance. Eight out of the 11 teachers stated that the information within the reports was sufficient for informing their next instructional steps with the student in the field test score report sample.

Although this was not explicitly stated in the stated research questions, one other result of this study is a proposed modification of the reporting framework. Currently, the framework identifies four major areas for consideration when developing score reports: 1) content, 2) format or presentation, 3) mode of dissemination, and 4) application of design principles. Given the responses of the participants in this study, it would be appropriate to add one additional, overarching factor: context. While considering context when developing score reports may be viewed as common sense, its inclusion in the reporting framework

makes it explicit and highlights its influence when making decisions around the reporting components of content, format, and mode of dissemination.

Limitations

There are at least three limitations of this study. First, it was noted earlier that timeliness of reporting was an issue and that web-based reporting is a promising solution. However, the literature review presented in this paper did not discuss the design and cognitive implications of a web-based environment for score reporting. In this study, score reports were made available in a static, non-interactive electronic format. A more thorough literature review concerning dynamic and web-based communication should be conducted prior to designing online score reports. This kind of score reporting will likely become the norm as computer-based testing continues to grow in demand and use.

Second, the diagnostic score reports were developed primarily from the perspective of one person's interpretation of the literature. Although the score report incorporated information design research recommendations, it represents one of a possible number of equally acceptable forms that can arise from application of the proposed reporting framework. Ideally, the development of score reports would involve a number of disciplines working together with the intended audiences of these reports. The initial sample score reports were developed to respond to anticipated information needs identified in the literature and were not initially developed with the input of the target audience. The incorporation of a combined "top-down" and "bottom-up" approach to developing the score reports is a limitation which could be addressed in a future study.

Third, the participants in this study were a voluntary sample of Grade 3 teachers who were involved in the Diagnostic Mathematics project. Recruiting teachers who were familiar with the CDMA was desirable for evaluating the score reports to help collect the most relevant feedback. These teachers provided valuable and context-specific user feedback which could then be incorporated into revisions of the score report. However, the participants in this study are a limited sample and given this constraint, caution must be exercised with generalizing the results of this study to other Grade 3 teachers across the province.

Future Directions

There are at least four directions for future research. First, the main focus of this dissertation was on reporting student-level cognitive diagnostic results. This discussion was primarily devoted to the development of diagnostic score reports. As a follow-up, a small-scale evaluation was completed with the score reports developed as part of the Diagnostic Mathematics project. Future studies should focus on a larger-scale evaluation when developing score reports that implement the recommendations put forth for reporting CDA results. This evaluation should involve multiple educational stakeholders, including administrators, teachers, parents, and where appropriate, students, to determine the effectiveness of the reports for imparting meaningful and useful information to support instruction and learning. Promising methods for evaluating diagnostic reports with target audiences of the information include: the use of focus groups and/or individual semi-structured interviews, think-aloud methods with simulated reports to identify problematic areas of the score report, and experimental studies

to systematically evaluate the effectiveness of different reporting schemes (Goodman & Hambleton, 2004).

Second, future studies should look at the potential for adapting the information in one score reporting template (i.e., for teachers) to enable appropriate interpretation and use with another user group (i.e., parents). In practical terms, being able to adapt an existing score report may be a better alternative to creating different reports for each target audience. The results of this study suggest that making small modifications to the score report used by teachers and with the inclusion of interpretive information, the score report may be shared with parents and older students in a comprehensible way. As recommended, user input and feedback is an important inclusion in the score reporting development process. Although recommendations and suggestions were made by the participants for improving the score reports for particular audiences, revisions to the score reports were not made. Future studies should also focus on how user input and feedback can be systematically collected and then incorporated with each iteration in the score reporting design process.

Third, an avenue of research can be pursued to further investigate how diagnostic score information is *actually* used by teachers, parents, and students to help with instruction and learning in the classroom context over time. In this study, many of the participants stated potential uses of the information as compared to reporting actual uses in their classroom. Score reporting can provide an opportunity for student learning by providing specific feedback on test performance. Both teachers and parents can assist the student in interpreting this

information and helping the student set learning goals informed by CDA results. An example of how the function of score reporting can expand in this direction is the provision of a printable “Learning Goals” sheet. Areas requiring improvement can be selected and printed in by the student, or in the case of an interactive web-presentation, areas identified as needing improvement can be directly linked to the document. A “Learning Goals” sheet provided with the score report capitalizes on the diagnostic feedback, can initiate discussions between the student and teacher/parent in setting concrete action plans for remediation of areas of weakness, and can encourage the student to be an active participant in his or her learning.

Fourth, given the importance of computer-based testing, future studies should investigate score reporting strategies in an online environment. Two major differences between paper-based score reports and computer-based score reports are access to information and flexibility of presentation. Web-based environments have the capability to manage and organize large amounts of information (Nielsen, 2000) using tools, such as ribbons and hyperlinking, which are not available in print-based documentation. Score reports developed in an online environment have a greater potential to serve multiple information needs. For example, user-driven menus can tailor displays of information with respect to presentation (e.g., larger font or different colors), language and aggregation of results (e.g., classroom-level results). When described this way, developing online score reports can become modular in nature. An online interface can be designed using universal design principles and providing the basic information components

as outlined in the reporting framework. The machinery behind this interface that allows for user-driven commands to aggregate data, alternate between data presentations or even languages can be built systematically over time and integrated into the system.

Perhaps the greatest potential for online score reporting is with communicating technical information in ways other than text to promote comprehension and accurate interpretations. For example, multimedia components can be integrated into an online score reporting interface. While narrative explanations should still be made available, these can be presented in the form of a video or an interactive tutorial where a user could evaluate their own understanding of the concepts described within the report (e.g., reliability and error of measurement). In the context of CDA, there has not been enough focus on describing and explaining highly technical cognitive diagnostic models in non-technical terms. This may be communicated to educational stakeholders using strategies such as plain language, pictures, or stories (Sireci & Fast, 2012). In particular, interactive 3-D representations of models can help users understand complex ideas such as what a neural network looks like, the logic behind training it and determining weights, and then how it applies to scoring a diagnostic assessment. Development of modules on neural networks can promote transparency in the scoring process while also contributing to validity arguments of where diagnostic test scores come from and the supported inferences that can be made about a student's knowledge and skills.

Conclusion

The basic requirements for score reporting are clearly identified within the *Standards for Educational and Psychological Testing* (1999). However, the methods to achieve these standards are not. There lies an implicit assumption that results are reported in a useful manner to educational stakeholders to enable their use for communicating student performance. Effective reporting of diagnostic results requires a multi-disciplinary effort and input from all target audiences. Score reporting should be viewed as a form of communication between the test developer and test user, aspiring to achieve the goal of *clarity of communication.* The good news is that there are many new tools and technology available to assist us with this task. However, in order for CDA to realize its potential for informing instruction and guiding student learning, more research is required to further explore reporting strategies with different audiences who have specific but diverse information needs in a variety of educational contexts.

REFERENCES

American Educational Research Association (AERA), American Psychological Association, National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, D.C.: AERA.

Aschbacher, P.R., & Herman, J. L. (1991). *Guidelines for effective score reporting* (CSE Technical Report 326). Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing.

Beaton, A. E. & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17,* 191-204.

Bertin, J. (1983). *The Semiology of Graphics.* Madison, WI: University of Wisconsin Press.

Boyatzis, R. E. (1998). *Transforming Qualitative Information.* Thousand Oaks, CA: Sage.

Center for Universal Design (n.d.). *The principles of universal design.* Retrieved March 3, 2008, from http://www.design.ncsu.edu/cud/about_ud/udprinciplestext.htm.

Cleveland, W. S. (1994). *The elements of graphing data.* Monterey, CA: Wadsworth.

Cleveland, W. S. & McGill, R. (1985). *Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of American Statistical Association 79,* 531-554.

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

DiBello, L., Stout, W., & Roussos, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 361-389). Hillsdale, NJ: Erlbaum.

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational measurement: Issues and practice 10,* 3-9, 16.

Forte Fast, E., & The Accountability Systems and Reporting State Collaborative on Assessment and Student Standards. (2002). *A guide to effective accountability reporting.* Washington, DC: Council of Chief State School Officers.

Gierl, M. J., Cui, Y., & Hunka, S. M. (2007, April). *Using connectionist models to evaluate examinees' response patterns on tests using the attribute hierarchy method.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

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

Gierl, M. J., Cui, Y., & Hunka, S. M. (in press). Using connectionist models to evaluate examinees' response patterns on tests. *Journal of Modern Applied Statistical Methods.*

Gierl, M. J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *Journal of Technology, Learning, and Assessment, 6* (6). Retrieved January 28, 2008, from http://www.jtla.org.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretative guides: Review of current practices and suggestions for future research. *Applied Measurement in Education 17,* 145-220.

Gribbons, W. M. (1992). Organization by design: Some implications for structuring information. *Journal of Technical Writing and Communication 22,* 57-75.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 333-352.

Hambleton, R. K., & Slater, S. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Teaching.

Hartley, J., & Jonassen, D. H. (1985). *The role of headings in printed and electronic text: The technology of text.* Englewood Cliffs, NJ: Educational Technology Publications.

Hartz, S. M. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practicality.* Unpublished doctoral dissertation.

Horton, W. (1991). Overcoming chromophobia: A guide to the confident and appropriate use of color. *IEEE Transactions on Professional Communication, 34,* 160-173.

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.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. A. (1991). Does interpretive test score information help teachers? *Educational Measurement: Issues and Practice, 10,* 16-18.

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. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26,* 31-73.

Jonassen, D. H. (Ed.). (1983). The technology of text: Principles for structuring, designing, and displaying text. Englewood Cliffs, NJ: Educational Technology Publications.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions and connections with nonparametric item response theory. *Applied Psychological Measurement, 12,* 55-73.

Koretz, D. M., & Diebert, E. (1993). *Interpretations of National Assessment of Educational Progress (NAEP) anchor points and achievements levels by the print media in 1991*. Santa Monica, CA: RAND.

Kosslyn, S. M. (1994). *Elements of graph design.* New York, NY: Freeman.

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

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

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41,* 205-237.

Linn, R. L. & Dunbar, S. B. (1992). Issues in the design and reporting of the National Assessment of Educational Progress. *Journal of Educational Measurement, 29,* 177-194.

Lissitz, R. W. & Bourque, M. L. (1995). Reporting NAEP Results Using Standards. *Educational Measurement: Issues and Practice, 14,* 14-23, 31.

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

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K. B. Laskey & H. Prade (Eds.), *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 437-446). San Francisco, CA: Morgan Kaufmann.

Mislevy, R. J., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Nielsen, J. (1993). *Usability engineering.* San Diego, CA: Academic Press.

Nielsen, J. (2000). *Design web usability: The practice of simplicity.* Indianapolis, IN: New Riders Publishing.

O'Callaghan, R.K., Morley, M.E., & Schwartz, A. (2004, April). Developing skill categories for the SAT Math section. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Pettersson, R. (2002). *Information design: An introduction.* Philadelphia, PA: John Benjamins Publishing.

Ryan, J. M. (2003). *An analysis of item mapping and test reporting strategies.* Retrieved August 14, 2007, from http://www.serve.org/Assessment/assessment-publicationh1.php#StApub.

Schriver, K. A. (1997). *Dynamics in document design.* New York, NY: John Wiley.

Shah, P., Mayer, R. E., & Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph construction. *Journal of Educational Psychology 91,* 690-702.

Sireci, S. G., & Forte, E. (2012). Informing in the information age: How to communicate measurement concepts to education policy makers. *Educational Measurement: Issues and Practice, 31,* 27-32.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67,* 485-518.

Swarts, H. L., Flower, J., & Hayes, J. (1980). *How headings in documents can mislead readers.* Technical Report no. 9. Washington, DC: AIR, Document Design Project.

Tatsuoka, K. K. (1980). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345-354.

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

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (pp. 327-359). Hillsdale, NJ: Erlbaum.

Templin, J. L. & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305.

Trout, D. L., & Hyde, E. (2006, April). *Developing score reports for statewide assessments that are valued and used: Feedback from K-12 stakeholders.* Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.

Tufte, E. R. (2001). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Vaiana, M. E. & McGlynn, E. A. (2002). What cognitive science tells us about the design of reports for consumers. *Medical Care Research and Review, 59*, 3-35.

von Davier, M. (2007). *Hierarchical general diagnostic models* (Research Report No. RR-07-19). Princeton, NJ: Educational Testing Service.

Wainer, H. (1997). Improving tabular display with NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics, 22*, 1-30.

Wainer, H. (1992). Understanding Graphs and Tables. *Educational Researcher, 21*, 14-23.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement, 36*, 301-335.

Waller, R. (1983). Text as diagram: Using typography to improve access and understanding. In D. H. Jonassen (Ed), *The technology of text: Principles for structuring, designing, and displaying text* (pp. 137-166). Englewood Cliffs, NJ: Educational Technology Publications.

Whitely, S. E. (1980). Multicomponent latent trait model for ability tests. *Psychometrika, 62,* 495-542.

Winn, W. (1991). Color in document design. *IEEE Transactions on Professional Communication, 34,* 180-185.

Wright, P. (1977). Presenting technical information: A survey of research findings. *Instructional Sciences, 6,* 93-134.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20,* 15-25.

APPENDIX A

Participant research information sheet

Research Information Sheet for teachers who have participated in Alberta Education's Diagnostic Mathematics field test 2009-2011

Project Title: Developing and Evaluating Score Reports for Cognitive Diagnostic Assessments

Thank you for considering being a participant in my research study in which I am investigating alternative ways of diagnostic reporting. My name is Mary Roduta Roberts, and I am conducting research on the effectiveness of diagnostic reporting as part of my Doctoral dissertation. I would very much appreciate your participation.

The context of my study is the Alberta Education Grade 3 online diagnostic mathematics test. CDAs are designed to measure specific knowledge and skills that are aligned with curriculum goals. The results of a CDA yield a profile of scores which provides detailed information about a student's cognitive strengths and areas requiring improvement. This feedback may be used by teachers to guide and monitor their own students' learning processes.

A score report has been created to report the CDA results. Part of my dissertation is to evaluate this report. An online questionnaire was developed to 1) evaluate the format and content of the sample diagnostic score report, 2) collect information about teachers' sources of, uses and preferences for diagnostic information, and 3) collect background information. I am hoping you will be able to complete this online questionnaire.

In order to participate, I will email you a link to an online questionnaire. You will be asked about your perceptions and opinions of the current diagnostic score report. The questionnaire should take no more than 15 minutes to complete.

I developed two other diagnostic score reports as alternatives to the current score report. Hence the second part of my dissertation is to evaluate each of these reports and compare the results with the results for the current report. If you would like to help with this evaluation, then I would like to talk with you about reporting and the three reports. The interview should take about 30 minutes. If you would like to participate in these interviews, check the box at the end of the online questionnaire and provide me with your contact information.

The plan for this study has been reviewed and approved by the Faculties of Education, Extension, Augustana and Campus Saint Jean Research Ethics Board (EEASJ REB) at the University of Alberta. For questions regarding participant rights and ethical conduct of research, contact the Chair of the EEASJ REB c/o (780) 492-2614. In the case of any concerns, complaints or consequences at any point during the study you may contact me or my supervisor Dr. Mark Gierl at (780) 492-2396, or the Department Chair of Educational Psychology at (780) 492-5266.

Thank you for considering my request to participate in my doctoral study. If you are willing to participate, please contact me by mail at the address below or by email so that I can provide you with the link to the questionnaire.

Mary Roduta Roberts
PhD Student
6-110 Education North
Department of Educational Psychology
University of Alberta
E-mail: mroberts@ualberta.ca

APPENDIX B

Teacher questionnaire

The purposes of this study are to develop and evaluate student score reports for Grade 3 cognitive diagnostic mathematics assessments (CDMA). The CDMA is a curriculum-based set of assessments that can be used throughout the school year to measure students' strengths and weaknesses in thinking and learning mathematics. Results from CDMA are communicated through student score reports. The information from CDMA can be used to provide valuable instructional guidance, which may be needed to design remedial instruction programs or supplemental interventions for students.

The available literature on score reporting in education has shown that score reports are difficult to read and understand, and often lack adequate supporting material to assist the reader in making appropriate interpretations. Effective score reporting is important because the feedback may be used by teachers to guide and monitor their own teaching and their students' learning processes. To help develop student score reports that are understood, interpretable, and used, additional information from teachers, the users of the reports, is needed. It is hoped that the results of this study will lead to improved score reporting practices for diagnostic assessments and increase our understanding of how teachers may interpret and use the diagnostic information in the classroom.

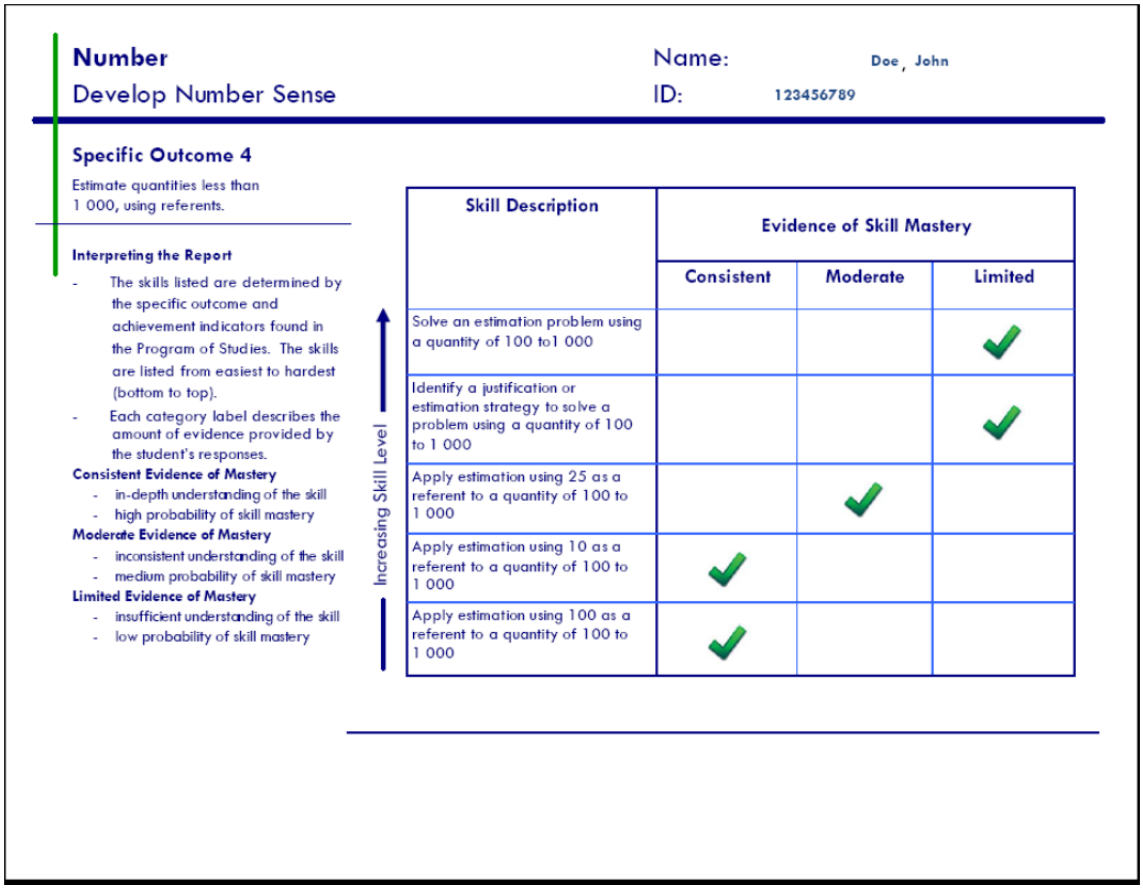
There are three major sections to this questionnaire: Content and Format of the Score Report, Diagnostic Information from Student Score Reports, and

Background Information. Please answer all questions. At the end of the questionnaire, there is an invitation for you to participate in a follow up telephone interview. If you are willing to participate in the telephone interview, please complete the required fields.

The questionnaire should take no more than 10-15 minutes to complete. All of your information will be kept private and study results will be kept confidential. Upon giving consent, you have the right to withdraw from the study at any time without prejudice. If you choose to withdraw from the study, your information will not be used in the results. If you have any concerns or questions related to the study, please contact Mary Roduta Roberts by email at mroberts@ualberta.ca or by telephone at (780) 453-2523.

Thank you for taking the time to complete this questionnaire.

B. An example Grade 3 student diagnostic score report



Number
Develop Number Sense

Name: Doe, John
ID: 123456789

Specific Outcome 4

Estimate quantities less than 1 000, using referents.

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
Solve an estimation problem using a quantity of 100 to1 000			✓
Identify a justification or estimation strategy to solve a problem using a quantity of 100 to 1 000			✓
Apply estimation using 25 as a referent to a quantity of 100 to 1 000		✓	
Apply estimation using 10 as a referent to a quantity of 100 to 1 000	✓		
Apply estimation using 100 as a referent to a quantity of 100 to 1 000	✓		

C. Content and Format of the Score Report

Please refer to the sample student diagnostic score report when answering the following questions.

On a five-point scale, please indicate the degree to which you agree with the following statements. The scale ranges from 1=highly disagree to 5=highly agree.

Content of the Score Report

1. The information reported is easy to interpret.

2. The information reported is not useful.

3. Performance level descriptors are appropriate for presenting individual student-level diagnostic results.

4. The language used in the student score report is understandable.

5. Please identify any unclear terms:_____

Format of the Score Report

6. Too much information is presented in the student score report.

7. The amount of information presented in the student score report is sufficient.

8. The student score report is well-organized.

9. The information is presented clearly.

10. The score report is visually appealing.

D. Diagnostic information from Test Score Reports

11. Different kinds of information can be included in a diagnostic report. Please indicate the degree to which each of the following types of results would help you make a diagnostic decision. The scale ranges from 1=highly disagree to 5=highly agree.

a. Overall (Total) subject-level score (e.g., Number = 16/24)

b. Subject-level subscores.

c. Descriptions of specific knowledge and/or skills a student demonstrated on the test (e.g., Student correctly answered questions that require application of skip counting of 100 forward, starting point from 100 to 1000)

d. Descriptions of specific knowledge and/or skills a student should develop (e.g., Student incorrectly answered questions that require application of skip counting of 100 forward, starting point from 0 to 1000)

e. Item-level results

f. Reporting performance levels (e.g., Limited, Moderate, Consistent Evidence of Skill Mastery)

g. Other: please specify

12. For diagnostic purposes, do you want a report like the sample report for each of your students for each strand for an entire subject area? (e.g., You would receive skill-level results for Number, Shape & Space, and Patterns & Relations, and Statistics & Probability on one score report for each of your students).

Yes No

13. For diagnostic purposes, do you want a report like the sample report at the classroom level, for each skill hierarchy? (e.g., For Numbers, Developing

Number Sense, you would receive a report summarizing the performance of your class for each skill).

Yes No

14. What are the current information sources that you use to make decisions about individual students? Please select all that apply.

Previous administration of the provincial achievement test for yourself and for the next teacher

Classroom textbook unit tests

Teacher designed tests

Anecdotal information

Observation assessments

Oral assessments

Other: Please specify _____

15. Diagnostic information from assessments can serve to guide different educational practices in the classroom. Do you currently use diagnostic information to (please select all that apply):

Then, for each of the educational practices selected, please indicate how often you use diagnostic information from assessments. The scale ranges from 1=Rarely (2-3 times per year) to 3=Frequently (at least once a week).

Plan your instruction

Select your instructional strategies

Assess your own teaching effectiveness

Give feedback to your students

Select remedial activities for your students

Refer a student for further testing

Other: Please specify _____

16. How important is it to you to receive diagnostic information from these student score reports?

Very important

Somewhat important

No opinion

Somewhat unimportant

Very unimportant

17. What is your preferred mode for receiving student score reports? Please select only one option.

Print-based

Print-based with accompanying website

Web-based only

Other (please specify): _____

18. Do you have any additional comments about the student score report?

E. Background Information

19. Gender

Male

Female

20. How many years have you been teaching?

21. At what grade are you currently teaching?

22. How many years have you been teaching at your current grade?

23. What is your current level of education?

Bachelor of Education

Degree (e.g., BA, BSc) plus Bed

MA/Med

PhD/EdD

Other: Please specify _____

24. My background in educational assessment comes from: (Please select all that apply)

University of college course as part of a teacher preservice program

University or college course as part of a graduate or extra courses program

Inservices or workshops

Newsletters or bulletins

Classroom experience

Other: Please specify _____

F. Consent to participate in a follow-up telephone interview

Would you be willing to participate in a follow up telephone interview to give us more detailed information and your opinions on diagnostic test score reports? We would like to follow-up with your responses to this questionnaire in a telephone interview. Also, you will be asked to evaluate alternate diagnostic score reporting templates and comment on the utility of the score reports for supporting your instruction. The interview should take no more than 45 minutes to one hour of your time.

Yes. Please provide your name and email address so that we can contact you to arrange a convenient time for the interview. Please provide your phone number that we can call you at for the interview.

No. Thank you for taking the time to complete this questionnaire.

APPENDIX C

Sample interview protocol

INTRODUCTION AND CONSENT

Thank you for agreeing to be interviewed. I appreciate it very much.

"The purpose of the interview is for you to tell me more about your opinions on alternate diagnostic student score reports and to expand on the answers you gave in the online questionnaire. The information you provide will help me to revise or refine the current student score report to better reflect your preference and needs for information.

I would like to record the interview. All of your information will be kept confidential and you may choose not to answer any question that you are not comfortable with. You can request that I turn off the recorder at any time. Do you agree to continue?"

I. EVALUATION OF THE CURRENT STUDENT SCORE REPORT

First, we will start by reviewing your responses to the online questionnaire used to evaluate the currently used template for this recent round of field testing.

In the questionnaire, you were asked to rate whether different kinds of assessment results would help you to make a diagnostic decision (see Section II Diagnostic Information from Test Score Reports).

- I am interested in knowing more about your reasons why certain kinds of assessment information have more or less diagnostic value to you.
- What does it mean to you when we say a test is diagnostic?
- What are your information expectations from a diagnostic assessment?

If you wish, you may want to make additional comments about the current reporting template or to expand upon your answers in the questionnaire.

Then, we will move on with the following open-ended evaluation questions for the current student score report.

A. Open ended evaluation questions

1. What information is MOST useful to you?

2. What information is LEAST useful to you?

3. Is there any information that you would like, but is missing?

4. Do you have suggestions on how to make the report more informative/useful?

5. How else can the information in the score report be presented?

6. What are your biggest concerns with this score report?

7. Do you have any additional comments?

B. Interpretive exercise

Next is the interpretive exercise. Please review the student diagnostic score report from the online questionnaire (see attachment - Current Template).

John has completed the online diagnostic math assessment and you have this student score report printed.

- Would you consider sending this score report home with John? Why?
- Would you consider showing this score report to John's parent? Why?
- Imagine that I am the parent of John Doe. I would like you to walk me through the score report. Considering all the information presented in the score report, how is John doing in Math?

Now, we will move on to an alternate reporting template. **Consider the alternate template only** when answering the following questions and try not to refer to the previous reporting template.

II. EVALUATION OF THE ALTERNATE STUDENT SCORE REPORT

A. Closed-ended answer questions

Please refer to the **alternate student score report template** when answering the following questions.

On a five-point scale, please indicate the degree to which you agree with the following statements. The scale ranges from 1=highly disagree to 5=highly agree.

Content of the Score Report

1. The information reported is easy to interpret.

2. The information reported is not useful.

3. Performance levels descriptors are appropriate for presenting individual student-level diagnostic results.

4. The language used in the student score report is understandable.

5. Please identify any unclear terms:_____

Format of the Score Report

6. Too much information is presented in the student score report.

7. The amount of information presented in the student score report is sufficient.

8. The student score report is well-organized.

9. The information is presented clearly.

10. The score report is visually appealing.

B. Open ended evaluation questions

1. What information is MOST useful to you?

2. What information is LEAST useful to you?

3. Is there any information that you would like, but is missing?

4. Do you have suggestions on how to make the report more informative/useful?

5. How else can this information be presented?

6. What are your biggest concerns with the score report?

7. Do you have any additional comments?

III. EVALUATION OF THE CURRENT AND ALTERNATE STUDENT SCORE REPORTS TOGETHER

Considering the two student score reports simultaneously,

1. Which report do you prefer the most? Why?

2. Do you have any other concerns about the two reports that have not yet been addressed?

IV. UTILITY OF THE SCORE REPORTS

1. Would a Grade 3 teacher find the information in these score reports helpful?

    On a four-point scale, please rate the helpfulness of the information in each report. The scale ranges from 1 = No use to teachers to 4 = Very useful to teachers.

2. How could a teacher use this information?

3. Might there be any problems with how the information is used? Why?

V. FINAL COMMENTS

Do you have any additional comments you would like to make?

Thank you for taking the time to participate in this interview!

APPENDIX D

Creating code based on teachers' responses to characteristics of diagnostic assessments

Steps taken:

1. For each participant, sectioned out part of the transcript where the question about diagnostic assessment was asked all the way to the open ended evaluation questions.

> **Participant 1:** Well see in the definition in the school system, diagnostic to me means some kind of task is going to assess something. So when you think about diagnosis of something, it's similar. So when you're thinking diagnostics, it's just one more tool that I can use to evaluate performance. I think if it's more specific because I think the more you can define the mastery level and the expectations for mastery, it's easier. And also I think it's, as far as the limited ones, 'cause my concern usually, with mastery children they're pretty easy to identify. You don't need a lot of diagnosis…
>
> I was just saying at the end of that in answer to your specific, like I think it's more important certainly to have lots of descriptors for limited, we were just discussing that, and that has to be more specific information we can get because I think that's a broader range.
>
> **QUESTION:** A broader range?

Participant 1: In terms of what limited can mean. Like they come with a lot of deficiencies whereas when you get to the mastery level, that narrows. I think anyway because there's not a lot of discrepancy in terms of what skills could have, like they have reading, they have a good sense of confidence. All these things that you see in mastery kids. They're all pretty homogenous, but when you get down to the limited end…Variable in their skills, yeah. Lots of more rote ,less being able to transfer abstract to concrete or concrete to abstract. Well to me diagnostic, I don't need it for the high end kids. I know they're doing well. Really. I need diagnostic, I need more tools to get me, to get that child set up, especially in Grade 3, to meet the criteria that you need to be able to function in Grade 4. That's what, if you ask any teacher, I want something that will make sure that I know I've done my job and I've prepared this child for the next grade. That's really what we're aiming for and so if I have a limited child, the more information I can get about that child and resources to help that child, the better off I'm going to be. 'Cause I think as a teacher, we spend most of our time in that area with diagnostic and testing and assessment with the limited and moderate kids.

2. For each participant's answer, main ideas were listed.

For Participant 1:

Diagnostic means assessing something.

It's one more tool to evaluate performance.

It’s specific – defines mastery levels and expectations for mastery

For limited children

Should have lots of descriptors for limited because they have lots of deficiencies and are more variable in their skills as a group.

Diagnostic helps the child meet the criteria needed to function in grade 4.

Lets the teacher know they’ve prepared the child for Grade 4

The more information and resources the better to help the child.

3. Similar statements were looked for across all participants. These statements were then grouped conceptually. An example of a conceptual grouping of statements across a sample of teachers is provided next.

Participant 1

Diagnostic helps the child meet the criteria needed to function in grade 4.

Let’s the teacher know they’ve prepared the child for Grade 4

Participant 2

Diagnostic tells the teacher whether the child has or has not understood the concept

Tells the teacher what to re-teach before moving on.

Participant 4

Diagnostics gives specific details to let the teacher know where to focus their teaching efforts.

4. Labels, definitions, and indicators were identified for each theme. A list of the themes for teachers' perceptions of diagnostic assessment characteristics is provided next.

Conceptual Group 1 Summary:

Diagnostic assessments provide information about a student's areas of strength and weakness.

The person states diagnostic assessments identify [specific, pinpoint] areas of weakness and strength. Within each area it determines the level of competency for each skill. Diagnostic information in the form of strengths and weaknesses is information to go and work with [specific] children.

Coded when a participant writes diagnostic assessment provides information about a student on "areas of mastery or non-mastery" "areas of strengths and weaknesses", "areas needing work or improvement".

Participants coded with this theme present (8)

Statements taken/paraphrased from the transcripts:

- It's specific – defines mastery levels and expectations for mastery
- Diagnostics pinpoint areas of mastery or areas of weakness
- Educational diagnostics don't just tell us where we're failing.
- Educational diagnostics tells us specific areas where we need to change or grow to correct it.
- Good diagnostics tells you why they are a year behind. What is missing?
- Good diagnostics will tell you exactly what the problem is so you know what to work on.
- Understand where the child needs the most help.
- Diagnostic reporting can show areas of strength to the child and parents.
- Diagnostic is what are the weaknesses.
- Diagnostic reports show strengths and weaknesses
- Diagnostics is specific, identifies areas needing work.
- Students need help and this is where they need it (specificity at the level of every curriculum point).
- *Based on the child's answers, determine the level of competency for each skill
- Fine tunes things for the teacher to see specifically where the kids have trouble
- Diagnostic gives info to go and work with kids.

Conceptual Group 2 Summary:

Diagnostic reporting should provide descriptive information on performance.

The person characterizes diagnostic assessments as having detailed descriptions for reporting performance levels. For example, a lot of detail should be provided on what a limited student looks like in terms of the different kinds of deficiencies and skills. It is good to know which questions the student got right and wrong. It is also good to know the characteristics of the items the student got correct – was the item easy or hard? Diagnostic assessments should not report

total score or percentage correct – numbers may not be needed because it's not helpful. Diagnostic reporting should be simple.

Coded when a participant writes diagnostic assessment provides should provide information about "which items a student got right and which ones he/she got wrong", "no numbers should be reported", "no percentages or total correct", "detailed description of performance levels".

This grouping differs from group 1 in that all other kinds of information considered to be diagnostic besides areas designated as being a strength or weakness.

Participants coded with this theme present (6)

Statements taken/paraphrased from the transcripts:

- Should have lots of descriptors for limited because they have lots of deficiencies and are more variable in their skills as a group.
- Having numbers in the form of total score is not helpful.
- Should know which ones the child got right or wrong.
- Descriptive performance labels on the report give a clear idea of which child you need to work with.
- how they did on the questions and reports the results in a simple manner.
- Diagnostics is NOT a percentage, it's NOT a count of correct or incorrect responses.
- Diagnostic specifics including which questions did they get right?
- Know if they got easy or hard items right.
- Detailed descriptor labels for performance levels.

Conceptual Group 3 Summary:

Diagnostic assessment places the student's performance in relation to a pre-defined normative group.

The person describes diagnostic assessments as telling us where the student is performing in relation to the [peer group, grade level, and normal learning continuum]. It is a picture of where the student is right now and comparing it to where they should be.

Coded when a participant writes diagnostic assessment tells us "where the student is right now", "where the student is in comparison to their peers" "grade-level"

This grouping differs from groups 1 and 2 in that diagnostic is viewed in the traditional normative framework.

Participants coded with this theme present (6)

Statements taken/paraphrased from the transcripts:

- Tells me where the student is right now in terms of grade level/percentile
- Diagnostic compared students now vs. where they should be
- Where and at what level is the child achieving?
- But reading diagnostics tells you are reading at grade level? This math diagnostic – how does it fit with the new curriculum?
- Diagnostics gives more and pertinent information about where students are
- Diagnostics is diagnosing where they are on the normal learning continuum.
- Diagnose where they are at that moment
- Diagnostics to see their ability of the level they are at.
- Diagnostic tells you where they are at in relation to grade level

Conceptual Group 4 Summary:

The results of a diagnostic assessment can provide a teacher with information on instructional effectiveness and instructional guidance.

The person states that diagnostic assessments can help the teacher know if they've prepared the student for the next grade. It tells the teacher whether or not the students have understood what has been taught and if not, then to re-teach. Diagnostic assessments can direct the teacher on where to focus their teaching efforts and with which students. It brings to mind considerations of how the teacher can address the weaknesses. Diagnostic assessments can also tell the teacher if the student is meeting the outcomes [assumption: the teacher is teaching from the outcomes – alignment of curriculum and instruction].

Coded when a participant writes diagnostic assessment can tell us if I need to "re-teach" "did the students understand", "focus my teaching".

Participants coded with this theme present (5)

Statements taken/paraphrased from the transcripts:

- Diagnostic helps the child meet the criteria needed to function in grade 4.
- Let's the teacher know they've prepared the child for Grade 4
- Diagnostic tells the teacher whether the child has or has not understood the concept
- Tells the teacher what to re-teach before moving on.
- Diagnostics gives specific details to let the teacher know where to focus their teaching efforts.
- *Diagnostics is how are we going to address the weaknesses?

- Diagnostic expectations – tell me are students meeting the outcomes. If not, reteach.
- Use the results to work with those who have trouble.
- . Tells the teacher what the child “did not get”.

Conceptual Group 5 Summary:

Diagnostic assessment is another tool used to measure and evaluate students’ skills and abilities.

Diagnostic assessments are one more tool among others available to the teacher to evaluate performance and to provide evidence of learning. It is a source of information and a resource to help the child. Diagnostic assessment measures skills and may be thought of as a formal assessment of ability.

Coded when a participant writes diagnostic assessment is another “tool”, “evaluate”, “measures skills, knowledge”, “ability”.

Participants coded with this theme present (5)

Statements taken/paraphrased from the transcripts:

- Diagnostic means assessing something.
- It’s one more tool to evaluate performance.
- The more information and resources the better to help the child.
- Measures skills
- Diagnostic math is another tool, evidence of learning.
- Diagnostics – formal assessment of ability