An Investigation of Educators’ Data Habit of Mind

Saad Chahine

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Human Development and Applied Psychology
Ontario Institute for Studies in Education, University of Toronto

© Copyright by Saad Chahine 2011
Educators are increasingly being asked to interact with data to facilitate students’ learning in the classroom. However, as an educational measurement community, we have little understanding of the factors and/or contexts that facilitate educators’ successful use of data. Educators’ use of score reports and the relationship to the intended use is integral to the concept of validity. A conceptual model, “Data Habit of Mind,” is proposed to study educators’ understanding, interpretation and potential applications of results from large-scale assessments. The metaphor, “Habit of Mind,” was originally coined by Robert Sternberg and Dan Keating, and has been applied in the education sector to describe educators’ habits of inquiry when interacting with assessments. Based on an extensive review of the literature, Data Habit of Mind is defined as a combination of statistical literacy and score report interpretation. Statistical literacy is the extent to which an individual is able to describe, organize and reduce, represent, and analyze and interpret data. Score report interpretation is the extent to which an individual is able to describe, summarize, question, and propose an application for a given set of elements on a score report. The combination of these two makes up an individual’s Data Habit of Mind.

Twenty educators were interviewed to assess their level of statistical literacy and their score report interpretation skills. A cognitive interview approach was used to capture the educators’ cognitive processes as they solved performance-based tasks, and protocol analysis procedures were used to encode the responses into the conceptual model. Descriptions of educators’ Data Habit of Mind were then generated through qualitative matrix analysis. Four
groups of educators were identified based on the patterns of relationship between their statistical literacy and score report interpretation scores. Demographic factors, including teaching experience, gender and educational background were not meaningful predictors of educators’ Data Habit of Mind. These results add to our understanding of how educators interpret and use test results and have implications for test validation processes.
Acknowledgements

I have always been fascinated with mathematics, which led me to fast-track and take Calculus at a younger-than-usual age. When I did, I became fascinated with limits. There are three main properties of limits. The first property is based on addition and subtraction, the second on multiplication, and the third on division. As I reflect on this thesis, I think of how the third property illustrates my own personal journey of completing this thesis: “If the limit of the denominator is not zero, the limit of the quotient of the two functions is equal to the quotient of their limits.”

Limits allow us to do the almost impossible. For example, the above property shows that you can take a part of an almost nonexistent whole, if that whole is a function that approaches zero, but never reaches zero. In many ways I feel that I have divided myself into a function that approaches zero and never reaches this limit. Even when I thought I had reached my limit, many people helped me further divided my efforts to find that these limits may be infinite.

My parents were a big part of this journey, though they died before I began this thesis. The struggles that they experienced in moving a young and large family around the world, always made me think that what I am doing does not even compare to what they endured. I only wish to achieve a portion of what they accomplished in their lives.

During my graduate schooling, my brothers and sisters, Adel, Farida, Farid, Samira and Samir, provided a great deal of encouragement and patience. I am not sure how I could have done all of this without their support. I also would like to acknowledge my sister who lost her life before I was born; I am sure Adela would have been an incredible mentor.
Through my PhD studies, I was fortunate to be surrounded by supportive colleagues and friends. I want to thank Ruth Beatty for our friendship and support that she provided over the years. She also did not show up for drinks one night and created an opportunity for me to get to know Tomoko Arimura; Tomoko, I think that if you were able to put up with me the past few months in completing this thesis, then we can get through a lot together in life.

I also want to thank my data savvy friends at Datahost: Olesya Falenchuck, Monique Herbert, Malcolm Cunningham, Amanda Ferguson, Zhimei Gu, Jingshun Zhang, Yunmei Xu, Hervé Jodouin, Anita Ram, Yael Sher, Lina Fung, Dianne Thomson, Orlena Broomes, Barnabas Emenogu and many others who hung around the lab while we tossed around half-baked ideas, pondered new theories and techniques, and most of all passionately argued about the work we were doing. I only wish that in my academic career, I could design such a space. There are also many friends I want to thank for their support through the years: Sonia Ben Jaafar, Jill Adelson, Sarah Morrison, Dawn Pollon, and Louis Volonte.

Lorna Earl has a special place in my heart and mind. As my master’s supervisor, she encouraged me to peruse validity research. Through the years, we have become very good friends and have tossed around complex ideas as if it were casual conversation. Thank you for your guidance and support. I also want to thank the Earl family. While I stayed at the cottage finishing my thesis, they were extremely supportive and caring.

Lorna also led me to my doctoral supervisor, Dr. Ruth Childs. In the few years I spent working with Dr. Childs, I encountered many challenges and at times felt as though I had reached my limit. But she was always there to guide me and help me realize the potential that I did not even know I possessed. The wealth of knowledge that you have and the precision that apply to your work is truly amazing. Thank you for working with me and taking me on as your
student. I believe that I will always call you my supervisor. I also want to thank Jan Owens, not just for the yummy baked goods, but for also being there to support and encourage me.

To my committee, talk about limits! Well, when I thought I was close to zero, you pushed me a little further and further.

A big thank you to Professor Richard Wolfe. To me, you are the mountain: You are wise and grand with peaks of knowledge. When approaching the mountain, you always respect the complexity of the terrain. Thank you for your efforts and the patience that you showed me throughout this process.

To Dr. Eunice Jang, my how we have argued over the years about validity! I have loved our conversations; my thesis and my work would have never come to fruition without your mind being involved. Thank you for pushing my thinking, and pushing me to the limits. As I have learned from you, without pushing the boundaries of our knowledge, thinking, and life - we will never know what we can accomplish.

I would like to also graciously thank Dr. John Hattie. I have read and watched your work for several years and touted your expertise. I am honored that you were involved in this thesis.

I would also like to acknowledge Susan Eligie’s contribution to this thesis. I thought I was at my wits end, but you carefully helped me find ‘new life’ in my thesis and my writing. Thank you.

Lastly, I would like to acknowledge the Ontario Graduate Scholarship and the College Board Fellowship programs. The financial contributions that these programs provided enabled me to carry out this work.

~Thank You~
People often remark that I'm pretty lucky. Luck is only important in so far as getting the chance to sell yourself at the right moment. After that, you've got to have talent and know how to use it.

Frank Sinatra
Table of Contents

Prelude: Perception, Interpretation, and Consequences......................................... 1

Chapter 1. Introduction ........................................................................................................... 3

Chapter 2. Literature Review ............................................................................................... 6
  Evolution of the Concept of Validity ..................................................................................... 6
  Validity and Interpretation ...................................................................................................... 7

Score Reports .......................................................................................................................... 9
  Traditional Measurement-Based Indicators ........................................................................ 10
  Standards and Codes for Test Reporting ............................................................................ 10
  Guidelines from Measurement Research .......................................................................... 12

Visual Data Displays ............................................................................................................. 13

Visual Language Theory ........................................................................................................ 16

Web Design ........................................................................................................................... 17

Dimensions for Report Design ............................................................................................... 17

Data Habit of Mind .................................................................................................................. 19
  Habit of Mind ....................................................................................................................... 20
  Structure of the Observed Learning Outcomes (SOLO) Taxonomy ................................ 20
  Statistical Literacy ............................................................................................................... 21

Score Report Interpretation ................................................................................................ 23

Proposed Applications for Score Reports ............................................................................ 24

Data Habit of Mind Conceptualization .................................................................................. 26

Chapter 3. Conceptual Framework ....................................................................................... 29
  Culture of Data Use .............................................................................................................. 30
  Motivation for Data Use ....................................................................................................... 30
  Research Questions .............................................................................................................. 31

Chapter 4. Methods ................................................................................................................ 33
  Sampling ............................................................................................................................... 33
  Cognitive Interviews ............................................................................................................ 34

Instruments ............................................................................................................................. 36
  Statistical Literacy .................................................................................................................. 36
  Score Report Interpretation .................................................................................................. 37
    Cooper’s Fitness Test Score Report ................................................................................... 38
    Jabberwocky Test Score Report ....................................................................................... 40

Educational History and Comfort with Information, Mathematics, and Statistics .......... 42

Analysis Procedures ............................................................................................................. 42
  Stage 1: Recording Verbalization ....................................................................................... 43
  Stage 2: Transcribing Interviews ....................................................................................... 43
  Stage 3: Segmenting Interviews .......................................................................................... 43
  Stage 4: Encoding Episodes ................................................................................................. 43
  Stage 5: Analysis of the Codes ............................................................................................ 45
  Stage 6: Categorizing Educators ........................................................................................ 46

Relationship of Analysis to Research Questions ................................................................. 46
List of Tables

Table 2.1 Indicators of Effective Reporting of Diagnostic Information ........................................ 18
Table 4.1 Elements in the Cooper’s Fitness Test Score Report ..................................................... 38
Table 4.2 Elements in the Jabberwocky Test Score Report ........................................................... 41
Table 4.3 Statistical Literacy: Thinking Levels, Definitions, and Examples ..................................... 43
Table 4.4 Score Report Interpretation: Thinking Levels, Definitions, and Examples ......................... 44
Table 5.1 Participants’ Demographics, Experience, and Comfort Levels ......................................... 48
Table 5.2 Participants’ Statistical Literacy Skills .............................................................................. 56
Table 5.3 Participants’ Score Report Interpretation Skills Demonstrated on the Cooper’s Fitness Test Score Report ............................................................................................................................... 61
Table 5.4 Participants’ Score Report Interpretation Skills Demonstrated on the Jabberwocky Test Score Report ............................................................................................................................... 63
List of Figures

Figure 2.1. A conceptual model of statistical literacy: Thinking levels are dependent on statistical literacy skills................................................................. 22

Figure 2.2. A conceptual model of score report interpretation: Thinking levels are dependent on score report interpretation skills................................................................. 24

Figure 2.3. A conceptual model of Data Habit of Mind: Score report interpretation (Figure 2.2) is dependent on statistical literacy (Figure 2.1) ................................................................ 27

Figure 3.1. How Data Habit of Mind may lead to improvement in score report interpretation.. 29

Figure 4.1. Olympic medal task (part of the statistical literacy tasks presented in Appendix D). 37

Figure 5.1. Part of the state population and area codes task (a statistical literacy task). .......... 51

Figure 5.2. Part of the movie stars salary task (a statistical literacy task).............................. 53

Figure 5.3. A common graph that participants tried to draw when asked to draw a graph to represent the table in Figure 5.2................................................................. 54

Figure 5.4. An element from the Cooper’s Fitness Test Score Report (the thick bar is intended to represent measurement error). ................................................................. 59

Figure 5.5. An element from the Jabberwocky Test Score Report (the rings are intended to represent measurement error). ................................................................. 60

Figure 6.1. Overall score report interpretation level by overall statistical literacy level.......... 67

Figure 6.2. Categories based on Data Habit of Mind skills.................................................. 68

Figure 6.3. Categorization of participants based on Questioning score report interpretation skill demonstrated on the Cooper’s Fitness Test Score Report by overall statistical literacy score..... 69

Figure 6.4. Categorization of participants based on Questioning score report interpretation skill demonstrated on the Jabberwocky Test Score Report by overall statistical literacy score. ........ 70

Figure 6.5. Categorization of participants based on Proposed Application score report interpretation skill demonstrated on the Cooper’s Fitness Test Score Report by overall statistical literacy score. ......................................................................................... 71

Figure 6.6. Categorization of participants based on Proposed Application score report interpretation skill demonstrated on the Jabberwocky Test Score Report by overall statistical literacy score. ......................................................................................... 72
Prelude: Perception, Interpretation, and Consequences

Two rivals, Tycho Brahe (1546-1601), an astronomer, and Johannes Kepler (1571-1630), a mathematician, established the basis of understanding planetary motion. They had conflicting beliefs about planetary motion. As Hanson (1969) describes, “Kepler regarded the sun as fixed: it was the earth that moved. But Tycho followed Ptolemy and Aristotle in this much at least: ‘the earth was fixed and all other celestial bodies moved around it’” (p. 5). Kepler based his ideas on his understanding of Platonic principles and throughout his life collected data to help confirm his belief. When Tycho died, his rival Kepler used the data (some say, stole) to mathematically derive our current understanding of planetary motion. The earth rotating around the sun, proven by Kepler, helped form the basis for Newton’s theory of gravity.

Norwood Russell Hanson (1924-1967), a philosopher of science, wrote Patterns of Discovery: An Inquiry Into the Conceptual Foundations of Science in 1958 (reprint 1969). The book focuses on a single yet critical question about these rivals: “Let us consider Johannes Kepler: imagine him on a hill watching the dawn. With him is Tycho Brahe ... Do Kepler and Tycho see the same thing in the east dawn?” (p. 5). Hanson provides a detailed physiological explanation that the process of seeing is identical in every being: “Retinal reactions ... are virtually identical; so too are our visual sense-data, since our drawings of what we see will have the same content” (p. 6). Thus, the majority of people will have the same physiological processes when viewing an object.

If Tycho and Kepler had seen the same thing, they would not be in conflict. In truth what differentiated these two men and their beliefs was an interpretation of data. Hanson argues that “people, not their eyes, see”; thus, the critical difference between people is their individual
beliefs that influence interpretation. In other words, “there is more to seeing than meets the eyeball” (p. 7).
Chapter 1. Introduction

Standardized or large-scale testing requires a series of processes and products, including the definition of the construct to be tested, item writing, pilot testing, forms development, item selection, test administration, scoring and reporting. The results or score report constitute the link between the process of testing and users of the tests, including educators, students and parents. The psychometric literature on testing is large; however, the processes around the development and use of results or score reports have not received as much attention as other components of testing. Ronald K. Hambleton, an authority on psychometrics, declared in his Keynote Speech at the 2008 meeting for the International Commission on Testing, “Reporting is the single most important topic in psychometrics today.”

Hambleton’s statement was motivated by the potential compromise in validity when educators interpret and make judgments about students from large-scale assessment reports. Lee J. Cronbach was one of the leading theoreticians and practitioners of mid-twentieth century testing. He stated that “the responsibility for valid use of a test rests on the person who interprets it” (1971, p. 445), placing reporting at the core of the testing process. Although Cronbach’s seminal work in validity has been read and quoted by many, what we know about how educators interpret scores reports is mainly conjectural. Among measurement specialists, there is a sense that educators do not have the statistical knowledge to read and interpret score reports. In the past 50 years, there has been a great deal of research devoted to ensuring scores produced are to the greatest degree possible reliable and accurate. However, very little is known about the process by which educators interpret and make judgments based on reports from large-scale testing, which directly affects our current notions of validity.
There has been some research on how teachers combine interpretation and judgments about students (Broadfoot, 1996; Moss, 2003). Consider the following scene: In a classroom, an instructor asks rapid questions that are answered by students. The instructor in this case can decipher a student’s response, interpret it, judge a student’s level of understanding, and often provide a feedback cue, within a matter of seconds. In the classroom, the distinctions between assessment, interpretation and judgment may be blurry; however, the distinctions are clear when it comes to large-scale testing because different people and products are involved at each stage. Reports from large-scale testing often provide both guidelines for interpretation and results. An educator has to then “combine these [results] with his other knowledge about the person he tests and the assignments or adjustment problems that confront them to decide what interpretations are warranted” (Cronbach, 1971, p. 445).

Cronbach (1971) believed there ought to be “two uses of tests: (a) for making decisions about the person tested and (b) for describing him” (p. 445). Recently a substantial body of literature has developed promoting a third use of results: to improve student learning (Black & Wiliam, 1998, 2004; Earl & Katz, 2006; Hattie, 2009; Hattie & Timperley, 2007; McDonald & Boud, 2003). Thus, educators are facing increasing pressure to use information from large-scale testing for diagnostic purposes.

Concurrent advances in statistical and measurement techniques have allowed large-scale tests to provide skill-based feedback (Huff, Laitusis, & Ewing, 2002; Leighton & Gierl, 2007). Traditionally, tests provide information on overall domains (e.g., mathematics), and sometimes provide subscale scores (e.g., geometry). Tests that are designed to produce skills-based feedback provide reports that contain information about students’ cognitive processing (Leighton & Gierl, 2007).
When the external pressures to use data from tests are combined with the statistical and measurement means to provide diagnostic information, the potential effects on improvement of student learning are large, since teachers would have the information to deliver targeted instruction. However, the strength of this combination relies on the extent to which educators can interpret and use results from the reports. Currently, little is known about how the reports, specifically those from skills-based tests, are being interpreted and used by students, educators, and schools (Roussos, Templin, & Henson, 2007). The main purpose of this study is to add to our limited understanding of the validity of test score use, because there is little research on how score reports (reports produced by large-scale testing) are interpreted by educators and how educators believe they may be able to use these reports.
Chapter 2. Literature Review

The review of literature is divided into three parts. The first is a consideration of reporting within the overall framework of the theory of validity. This section builds the rationale for this research study. The second section, on score report design, is a review of relevant literature on designing reports. This literature supported the development of the score report prompts that were used in this thesis. Finally, a section on Data Habit of Mind includes references to the literature on statistical literacy and use of assessment results.

Evolution of the Concept of Validity

Tests were already designed with validity in mind during the early twentieth century and before; for example, Binet’s tests were designed with emphasis on a high degree of interpretability including “extremely precise guides for test materials, administration, scoring, and interpretation” (van der Linden, 1986, p. 326). Binet put considerable effort into guidelines for using his tests, to reduce unnecessary variability in testing procedures that would undermine validity.

World War 2 led to a flurry of psychological and performance testing (Lemann, 2000); after the war there were substantial developments in the theory and practice of testing. In the late 1940’s and early 1950’s, validity was understood to be a measure of “how well a test does the job it is employed to do” (Cureton, 1951, p. 621), and thus a property of the test itself. There was over a decade of discussion of the meaning of validity (Cronbach, 1992) that led to Cronbach and Meehl’s (1955) paper that provided an expanded view. The paper was based on collaborative efforts of many measurement experts and outlined four types of validity: (1) predictive validity, (2) concurrent validity, (3) content validity, and (4) construct validity. The paper included systematic methods of examining each of the four types of validity. These four types of validity
were widely adopted as the definition of validity by developers and users of psychological tests as reflected in an edition of the *Standards for Educational and Psychological Testing* (APA, AERA, NCME, 1985) that was in wide use until the late 1990’s.

Standards take time to change. Cronbach’s thinking about validity, however, continued to evolve. In 1971, he presented the idea of criterion-related validity as a combination of predictive and concurrent validity, while maintaining content and construct validity:

*Criterion-related* [predictive] validation compares test scores, or predictions made from them, with an external variable [criterion] considered to provide a direct measure of the characteristic or behavior in question.

*Content validity* is evaluated by showing how well the content of the test samples the class of situation or subject matter about which conclusions are drawn.

*Construct validity* is evaluated by investigating what psychological qualities a test measures; i.e. by determining the degree to which certain explanatory concepts or constructs account for performance on the test. (Cronbach, 1971, p. 444)

Cronbach pointed out that these ideas of validity were created based on a primary use of testing for personality measures and that, given the increased use of testing in education, “it appears desirable to attempt a new statement centered on educational uses and interpretations of tests” (Cronbach, 1971, p. 445). Anastasi (1986) explained that the transition to three types of validity and their intersection constituted a move towards a more thorough understanding of the complexity of validity. Testing experts were concerned that the idea of validity that was applied in the field was too simplified; for example, if two out of the three validities were met, then the validation requirements were met (Anastasi, 1986).

**Validity and Interpretation**

Samuel Messick’s (1975) work revolutionized educational research by including values as an underlying principle in measurement and evaluation. Consistent with Cronbach’s emphasis on validity being a matter of interpretation, he made a convincing argument for the inclusion of
consequences as an element of the validation process. Messick’s and Cronbach’s influences are evident in the 1999 revised edition of the *Standards for Educational and Psychological Testing*: “Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed purpose” (AERA, APA, NCME, 1999, p. 11). Although this definition is wordy, it is essentially what Binet was attempting to do in 1905 in ensuring that his test was being administered, scored and interpreted in the way he intended.

Messick’s work was also highly influenced by Cronbach and others who believed that interpretations are a critical part of the validity argument. Messick’s subsequent (1980, 1989, 1995) discussions of validity as having four facets were the first attempts to structure interpretations as part of an argument towards validity. His validity argument structure had two bases, evidential and consequential, both of which contribute to an argument for test interpretation and test use. He believed that Construct Validity was the most critical facet in a validity argument. The three other facets, Relevance/Utility, Value Implications and Social Consequences, build on each other to represent the situational context of the test and its use.

Messick’s and Cronbach’s work influenced the development of structuring arguments. Kane, in 1992, made a convincing case for the use of practical arguments, describing elements that should be included in an argument. In 1996, Crooks, Kane and Cohen put together an argument as an eight-linked chain. Each link represents a threat to validity. This work is based on the idea of logic and practical arguments that Kane (1992) describes. In Kane’s (1992) article, he additionally describes how measurement specialists may use Toulmin’s (1958) framework for structured arguments to aid the validation process. Currently, a great deal of research (Bachman, 2005; Bachman & Palmer, 2010; Mislevy, 2003, 2005, 2007, 2009) is being conducted to apply Toulmin’s structured arguments as a process of validation.
Structuring arguments for validity is not an easy task. There are many procedures that can develop evidence towards the construct as well as many other elements of validity. The complexity is in generating evidence that test scores are being interpreted and used in the ways they were intended. Consequently, many of the definitions of validity include elements of interpretation and use. For example, Moss, Girard, and Haniford (2006) write, “Educational assessment should be able to support these professionals in developing interpretations, decisions, and actions that enhance students’ learning. Validity refers to the soundness of those interpretations, decisions, or actions” (p. 109). We have little information about how to measure “soundness of …interpretations, decisions, or actions.” This thesis is an attempt to build on current conceptions of validity by providing a model to examine educators’ cognitive processing when interpreting and proposing an application for score reports.

**Score Reports**

Unlike the extensive literature readily available on test design and analysis, there is little guidance on how to best report or display test results to build a validity argument. In order to study how test score reports are interpreted and potentially used, it is important to study the best way to design reports to facilitate intended interpretations and use. Score reports provide students, educators, parents and other stakeholders with information about testing performance. Effective communication of results is needed to support the validity of the interpretation and use of the results. This part of the literature review is intended to identify effective elements of communication of results by drawing on traditional measurement sources, such as professional standards, as well as non-measurement literature, such as writing about web design. These bodies of literature may be further divided into traditional and contemporary sources.
Traditional Measurement-Based Indicators

In the field of measurement, there has been a growing focus on reporting (Goodman & Hambleton, 2004; Ryan, 2009). At the same time, there has been a growing attention to the best ways to display statistical data (Tufte, 1990, 2001, 2006, 2009). Both sets of literature are examined below in three subsections: a comparative analysis of the professional standards and codes of reporting, a summary of guidelines that are emerging from research, and a synopsis of the best ways to visually display data.

Standards and Codes for Test Reporting

The years since the adoption of widespread standardized and large-scale testing have seen the development of a number of publications designed to guide testing professionals. *The Standards for Psychological and Educational Testing* (1999) (referred to below as the *Standards*), developed by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, are widely used by educators, including teachers, administrators and developers. Baker and Linn (2002) write that the *Standards* “are widely recognized as the most authoritative statement of professional consensus regarding expectations for tests on matters of validity, fairness, and other technical characteristics of tests” (p. 4). Goodman and Hambleton (2004) and Ryan (2006) reference the *Standards* in their research. They focus on different subsets of the *Standards* to support their advice for the most appropriate method of reporting, as further explained below.

Another set of guidelines, the *Code of Fair Testing Practices in Education* (2004), was developed by the Joint Committee on Testing Practices, which included representatives from the American Counseling Association, the American Educational Research Association, the American Psychological Association, the American Speech-Language-Hearing Association, the
National Association of School Psychologists, the National Association of Test Directors, and
the National Council on Measurement in Education. Section ‘C’ is devoted to reporting and
providing stakeholders with information. The Code of Fair Testing Practices in Education states,
“Test developers should report test results accurately and provide information to help test users
interpret test results correctly. ... [and] test users should report and interpret test results accurately
and clearly” (pp. 8-9). Parallel reporting principles are provided: one for test developers and the
other for test users.

In Canada, a Joint Advisory Committee in 1993 developed The Principles for Fair
Student Assessment Practices for Education in Canada, which was endorsed by the Canadian
School Boards Association, the Canadian Teachers Federation, the Canadian Association for
School Administrators, and the Canadian Society for the Study of Education. The document is
divided into two parts: Classroom Assessments and Assessments Produced External to the
Classroom.

The National Council on Measurement in Education, in 1995, developed a Code of
Professional Responsibilities in Educational Measurement. Section ‘6’ is dedicated to reporting.
The recommendations are targeted to test developers and so include how reports should be
produced.

The Joint Committee for Standards is probably best known for their program evaluation
standards; however, they have produced an additional book, The Student Evaluation Standards
(2002). Similar to their popular program evaluation standards, this book has a section devoted to
reporting. This is a set of guidelines with sub-sections devoted to improving student learning.

An intensive analysis was conducted by the author to identify underlying themes in these
two documents. The resulting list of fourteen concepts or properties of exemplary reporting

appears below, some of which appeared in only one and some in three or more of the documents.

The full analysis is provided in Appendix A.

1. Suggestions on how to interpret results while taking into account test properties
2. Interpretation guides
3. Appropriate uses and non-appropriate uses of the scores and the report
4. How standards and/or norms were generated
5. Multiple measures to report on overall students’ performance
6. Information to allow for group comparisons
7. A rapid return time
8. Methods for monitoring progress and change
9. Information on students’ performance for parents and guardians
10. Confidence that information about a student is confidential
11. Multiple methods of reporting
12. Guidance in understanding the consequential use of scores
13. Methods to facilitate learning
14. Explanatory information on the reports

**Guidelines from Measurement Research**

The measurement literature contains suggestions about attributes of useful reporting.

Ryan (2006) describes two main purposes of score reports: (1) to account for student performance and (2) to support improvement of student performance. Ryan views subscale scores, such as Algebra and Geometry, as providing diagnostic information. He provides a list of elements that he has found common in testing reports and suggests that a potential benefit of online reporting is that reports can be more interactive.

Goodman and Hambleton (2004) conducted a major review of score reports and interpretation guides. They defined diagnostic information as: (1) Producing sub-domain scores and (2) identifying what skills student needed to improve. They limited their review to paper-based reports and interpretation guides.
The list below shows recommendations from these two studies, analyzed and grouped by the author. Ten concepts were found in reviewing these documents (details are provided in Appendix B).

1. Relevant information
2. How standards and/or norms were generated
3. Customized information based on user or group
4. Interpretation guides
5. Information on students’ performance for parents and guardians
6. Methods to facilitate learning
7. Explanatory information on the reports
8. Precision of measurement at the subscale level
9. Multiple methods of reporting
10. Accessibility to users on multiple platforms

Many of the concepts identified above and in Appendix B are consistent with those found in the standards literature. This suggests that, within many of the reports studied by the researchers, the standards are being applied. Some additional concepts, such as being able to view and interact with reports on various platforms, are new and come from innovations in technology. The researchers point out that providing customized information is important to those who are reading the reports and using the tools.

**Visual Data Displays**

Making statistical interpretation simpler has long been a focus of statisticians, applied researchers and persons working in fields that require the use of statistics. It is not only statistics that people are weary of, it is numbers in general. There are often misinterpretations by the public when reading statistical or probability-based information (Rosenthal, 2006). In the author’s personal experience teaching mathematics, it is apparent that many students and some teachers have difficulty distinguishing between a bar graph and a histogram, displays that are
intended to convey different information. However, if educators, students and parents receive visual data presentations that are easy to understand, they may be more likely to use the results and feedback to support students’ learning. As Tukey (1990) observed, when we display statistical information, we are trying to shape how people will understand a phenomenon.

Edward Tufte, a current leader in the field of visually displaying data, provides six principles for good data displays in his 2006 book, *Beautiful Evidence*: (1) comparisons, (2) causality, mechanism, structure, explanation, (3) multivariate analysis, (4) integration of evidence, (5) documentation, and (6) content. Tufte emphasizes that the way data are presented ought to facilitate comparisons, whether these be from year to year, or one group to the next. He points out that two-dimensional paper will allow only bivariate comparisons (e.g., achievement and time). With the increasing capabilities of on-line utilities, users may begin to choose variables for comparisons, and even attempt to represent multi-dimensional comparisons. Integration of evidence is critical to understanding data. When a measurement is also represented graphically it helps the user have a stronger understanding of the measurement.

Tufte emphasizes that documentation is important to support the credibility of information: Having relevant information on the quality of the test fosters readers’ confidence. The last and most important principle is content. Everyone is interested in how they performed on a test; they wait for the report. Thus, having relevant content in the reports and including accompanying tools for exploring and interpreting the data are the most critical aspects of displaying information.

Howard Wainer (1992) described three levels of questions that can be answered when looking at a geographical display of data:

1. Elementary-level questions (e.g., “What was the petroleum use in 1980?,” p.16),
2. Intermediate-level questions, which involve trends seen in parts of the data (e.g., “Between 1970 and 1985 how has the use of petroleum changed?,” p. 16), and
3. Overall-level questions, which involve an understanding of the deep structure of the data being presented in their totality, usually comparing trends and seeing groupings (e.g., “Which fuel is predicted to show the most dramatic increase in use?” or “Which fuels show the same pattern of growth?,” p.16).

His guide is useful to promote graphical representations that enable different levels of questioning. As educators vary in their levels of understanding, score reports ought to cater to all the levels to engage the reader. Wainer’s (1997a) work has also influenced tabular displays. He described four main rules that ought to be followed when presenting information in tables: rounding, sensible ordering, summarizing and spacing.

Wainer based these suggestions on concepts of visual perception. Visual and tabular displays of data are useful in reports for two reasons: (1) condensation of information in a small space and (2) attractiveness to the reader. However, there is a risk to using a graphical or tabular display, in that the reader has greater freedom for interpretation. If you were to provide a narrative interpretation, the reader would have less freedom. However, in an age of growing internet use, people may be less likely and willing to read multiple pages of narrative.

Both Tufte and Wainer were familiar with the work of John Tukey, who also offered a list of suggestions about data displays. Tukey’s (1990) belief was that an “important aspect of impact is immediacy. One should see the intended at once; one should not ever have to wait for it to gradually appear” (p. 328). Thus, visual representation of data should immediately facilitate an understanding of phenomena, without viewers having to intensively analyze the graphic. He discusses 19 ideas at length in his article; however, he points out the most important six:

1. immediate and strong impact,
2. easy flow of attention across parallel elements,
3. planning to show phenomena, not numbers,
4. attention to both prospecting for what the data might show, and transfer (to others) of what is learned from it,
5. partnership with computation, and
6. putting disproportionate responses to work (p. 327).

Tukey’s work is both informative and revolutionary; he believed that numbers represent phenomena, and that we should not use visual displays in place of numbers, but rather use numbers in complementary ways to provide the best understanding.

Tukey, Wainer and Tufte focused their recommendations on ways to represent data. Authors in other fields have considered visual approaches to communicating meaning, including developers of visual language theory.

**Visual Language Theory**

Visual Language Theory and the study of how people understand visuals and make relations between structures is an emerging field of research. Growth has been spurred by the increasing capabilities of computers and the expanding roles of visual expressions in the media. Marriott and Meyer (1998) describe the logic behind visual communication:

> [N]ot all human communication is sequential in nature. Important components of human communication are visual languages, such as maps or diagrams. In these languages the basic symbols are not encountered sequentially but rather seen together at a glance. (p. 1)

Thus, the placement of different elements on a report is as important as the numbers that go into developing a score and a scale. If those reading a report do not follow the sequence intended by those who developed the report, some of the meaning may be lost. Marriott, Meyer and Wittenburg (1998) additionally state, “By a visual language we mean a set of diagrams which are valid ‘sentences’ in that language where a diagram is a collection of ‘symbols’ in a two or three dimensional space” (p. 5). This definition has become accepted as a definition of visual language.
Web Design

Web design is the new face of reporting, as the internet increasingly constitutes the interface between testing programs and schools. A set of indicators about the quality of web design may assist in enhancing the “user-friendliness” of these systems. In short, examining effective web design is relevant for two reasons:

1. It is the leading area of development of ways to most effectively present information, and
2. Reporting is increasingly internet-based.

Many large testing companies (e.g., CollegeBoard, ACT, Cambridge, Ontario’s Education Quality and Accountability Office) have moved to a web-based platform. The literature on how people learn in web-based environments is growing. How a person interacts with information about student results from testing is unknown.

The literature on web design is vast; however, a useful summary is provided by the U.S. Department of Health and Human Services’ on-line book, *Research-Based Web Design and Usability Guidelines* (2006). The document is essentially a qualitative meta-analysis of literature on web design and provides a summary of the most effective design elements. Study of the document reveals four main themes: (1) clarity in meaning, (2) communicating with priority, (3) displaying quantitative elements, and (4) on-line reporting.

Dimensions for Report Design

The foregoing section of the literature review identified indicators of effective score report design. Table 2.1 (developed by the author) provides a summary list of indicators of exemplary reporting and the origin of each. This table was used to design score report prompts to be studied in this thesis.
### Table 2.1

**Indicators of Effective Reporting of Diagnostic Information**

<table>
<thead>
<tr>
<th>Indicators of Effective Reporting</th>
<th>Standards &amp; Codes for Test Reporting</th>
<th>Measurement Research</th>
<th>Statistical literature on visual data displays</th>
<th>Web Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suggestions on how to interpret results while taking into account test properties</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpretation guides</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriate uses and non-appropriate uses of the scores and the report</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>How standards and/or norms were generated</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple measures to report on overall students’ performance</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information to allow for group comparisons</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A rapid return time</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methods for monitoring progress and change</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information on students’ performance for parents and guardians</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confidence that information about a student is confidential</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple methods of reporting</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guidance in understanding the consequential use of scores</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methods to facilitate learning</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explanatory information on the reports</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevant information</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Customized information based on user or group</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision of measurement at the subscale level</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accessible to users on multiple platforms</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graphs answer elementary to overall levels of questions</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Numbers on report are rounded</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information in tables is ordered meaningfully</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rows and columns in tables are summarized</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spacing between elements in report allows for ease in perception</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This table shows the complexity of expert opinions on score reporting; some recommendations come from just one source, while many have overlapping sources.

**Data Habit of Mind**

There are multiple audiences for the results of large-scale testing: students, parents and educators are the most immediate. Educators have dual roles, as they both apply test results in the classroom and interpret them for students and parents. The importance of educators’ understanding in maintaining the validity of the test through the reporting stage is critical. As described in the prelude to this thesis, individuals can have conflicting interpretations based on their knowledge of the area, the culture in which they live, and how the data are presented.

Educational researchers (e.g., Earl & Katz, 2006; Sutherland, 2004) have referred to using data as developing an “inquiry habit of mind.” They use this metaphor when working with educators,
often in professional development settings, in order to encourage educators to develop this habit, so that it becomes a routine part of their experience with data. In this thesis, this metaphor is used to explore educators’ understandings of test reports specifically and statistical literacy more generally.

**Habit of Mind**

Dan Keating (1990) coined this metaphor to describe expertise in two interrelated modes, Domain-General Habits of Mind and Domain-Specific Expertise. The metaphor of habits of mind has several advantages:

First, it presumes no particular structural outcome. Rather than reducing diversity to fit an a priori pattern, it encourages the observation of what fits with what over time.

Second, it strikes a better balance between the inevitably closed or fixed quality of structures and the apparent plasticity of development. We know a bit about habits – the longer we have them, the harder they are to change – but they are very flexible in the early stages and are never completely rigid.

Third, habits of mind are not exclusively cognitive. They can easily incorporate dispositional, affective, motivational, and personality variability as well ...

(Keating, 1990, p. 256)

These advantages identified by Keating (1990) provide a rationale for the use of this metaphor with regard to educators’ use of data. The metaphor as described can allow for cognitive development at any level of learning. Data use habits developed early will last longer; similarly, a habit of non-use of data will be difficult to change. Lastly, for educators to develop a Data Habit of Mind, supports have to be in place.

**Structure of the Observed Learning Outcomes (SOLO) Taxonomy**

There is a large literature, which is not reviewed here, on frameworks for understanding cognitive processes of both adults and children. Moseley, Elliott, Gregson, and Higgins (2005)
reviewed over 400 journals and identified 55 different thinking skills frameworks with a view to gaining a greater understanding of a framework for understanding lifelong learning. The authors found that the most appropriate framework depended on the specific application and that frameworks may be combined in order to develop a model. One of the frameworks they reviewed was the Structure of the Observed Learning Outcomes (SOLO) taxonomy by Biggs and Collis (1982, 1998). The SOLO taxonomy was used in Jones et al.’s (2000) and Mooney’s (2002) research studies to describe cognitive processing. Hattie (2004) also used the SOLO taxonomy as a foundational element in the design of “asTTle,” a computer based testing program.

The SOLO taxonomy describes a progression from “idiosyncratic” to “analytical” processing and is not age dependent. Due to its flexibility, SOLO is an appropriate taxonomy for analyzing the Data Habit of Mind and was used in the current study. Jones et al.’s (2000) refinement of the Biggs and Collis (1982) SOLO taxonomy included the following four levels of cognition:

1. Idiosyncratic: focuses on irrelevant features when describing, representing and analyzing data.
2. Transitional: generally focuses on just one aspect of the data and at times regresses to idiosyncratic thinking.
3. Quantitative: begins to focus on more than one aspect of a data exploration task.
4. Analytical: focuses on several relevant features of a task but also makes connections among these various features. (pp. 278, 279)

These four levels are used in this study to construct scales of statistical literacy and score report interpretation and are represented on the vertical axis in Figures 2.1 and 2.2.

**Statistical Literacy**

Mooney (2002) and Jones et al. (2000) provided four hierarchal descriptive constructs of what children may be able to do as they progress through their mathematics education:
Describing, Organizing and Reducing, Representing, and Analyzing and Interpreting. The first construct, Describing, represents the ability to describe data that is presented in a table or a chart. The more precisely one is able to describe data, the higher level of fluency one has.

![Diagram showing thinking levels and statistical literacy skills](image)

**Figure 2.1.** A conceptual model of statistical literacy: Thinking levels are dependent on statistical literacy skills.

The second construct, Organizing and Reducing, involves the ability to organize and reduce data. The individual is able to take a set of unorganized information and reorder a set of points by applying some logic (e.g., smallest to largest). Additionally, the individual is able to find a central description for an entire data set by using means, medians and modes, and thereby “reduce” a data set into its “typical” properties.

The third construct, Representing, is complex, as it requires an individual to represent data in some visual format. In other words, it assesses how an individual might graph a data set based on the type of data.

The last construct, Analysis and Interpretation, is the most advanced. This construct requires individuals, educators in the current study, to compare and contrast data and/or to make
inferences based on how the data are represented. Examples might be comparing and contrasting data for males versus females, or determining what may happen in 10 years. Overall, it is assumed that the statistical literacy constructs are hierarchal; however, the third construct asks educators to draw on different and maybe more creative attributes to represent the data as presented in a table or chart.

**Score Report Interpretation**

In contrast to our understanding of statistical literacy, our understanding of data use by educators is not as well established. There is a developing body of work on how data and information may be used (Earl & Katz, 2006; Earl & Timperley, 2008; Hattie, 2009; Schildkamp & Kuiper, 2010); however, there is little in the literature about the cognitive processes involved in the interpretation of score reports by educators. Figure 2.2 was developed for the current study. It is a conceptualization similar to Figure 2.1 and identifies dimensions along which score report interpretation may be characterized. This figure is very closely tied to the statistical literacy research of Jones et al. (2000) and Mooney (2002), but has been adapted to the current application.
Figure 2.2. A conceptual model of score report interpretation: Thinking levels are dependent on score report interpretation skills.

The score report interpretation skills on the horizontal axis are similar to the statistical literacy skills described above, but have been adapted for the application of score report interpretation. Educators are not often asked to represent data, and therefore this stage was omitted from this conceptualization, and the skills of Summarizing, Questioning, and Proposed Application were added. Summarizing is when an educator can explain the report and elements of the report in her or his own words. Questioning is when an educator questions the results on the report and the possible inconsistencies. The literature of Proposed Application is described below.

Proposed Applications for Score Reports

As stated earlier, Describing, Summarizing and Questioning are verbs used to understand how an educator may be interpreting score reports. However, determining what the proposed uses may be is more complex. The relationship of interpretation to use is a major part of validity arguments (Bachman, 2005; Messick, 1980; Mislevy 2003). There is an increasing body of
literature on how educators use data to make decisions in schools (Hattie, 2009). There is also a growing literature on how to use data to improve children’s learning. This is often referred to as Assessment for Learning (AfL), a concept that is based on improving learning through critical feedback and continuous monitoring of children’s understanding. Research confirms that the potentially positive aspects of AfL are linked to increased test scores (Black & Wiliam, 1998; McDonald & Boud, 2003). As a result, many schools and districts have adopted AfL initiatives, while at the same time trying to continue everyday assessment and, as a result, they have only adopted the rhetoric without the practice of using assessment to improve children’s learning (Buckendahl, Plake & Impara, 2004; Dappen, & Isernhagen, 2005). To help educators incorporate AfL strategies and concepts within the classroom, researchers (Assessment Reform Group, 2002; Assessment Resource Center, 2006; Shute, Hansen & Almond, 2007) have developed principles or guides for AfL to help teachers use assessment to promote learning.

The approaches and guidelines do not adequately address the complexities that arise when attempting to use assessment for learning within the classroom. The identification of students’ learning progressions through formative and/or diagnostic assessment is critically important in using information to promote student learning, as a report can help both the student and the educator identify the next steps needed to move from a basic understanding to one that is more advanced.

Tests are designed with a learning progression in mind. Crocker and Algina (1986) describe two main “orthogonal properties” that need to be addressed in test development: (1) substantive content and (2) cognitive process. While the content is almost always predetermined by a set of guidelines, such as those set by a state, the cognitive processes that the
A test will elicit from children are much more complex to capture. Diagnostic testing has helped accomplish this task, as items are linked to specific skills children are expected to use.

Providing diagnostic information on a report does not mean that students will be able to directly improve their skills. All a report can do is identify to a student their strengths and weaknesses in skill development, and outline what they need to work on. Educators then can use their pedagogical expertise to help transition students from what was reported from the test to new or modified understandings. Through this process, educators use evidence as leverage in their teaching and can: (1) help students individually, (2) guide their own practice, and/or (3) aid in designing programming for a course, or cohort, etc. When combining some of the conventions of diagnostic testing with the types of information educators, students, and parents need in order to improve learning, the report may make a larger impact than simply producing scores and describing a student’s level of performance.

**Data Habit of Mind Conceptualization**

Educational measurement specialists commonly suggest that educators would be able to understand score reports if only they had greater statistical literacy, and educators wish for score reports that were easier to understand. The development of this framework is very much tied to theory, but also practice, combining the knowledge that someone needs to have to understand a report (the educational measurement specialists’ argument) with the score report design elements needed to facilitate understanding and action (the teachers’ argument). The conceptual framework for Data Habit of Mind includes three axes: (1) thinking levels, (2) statistical literacy skills, and (3) score report interpretation skills. Statistical literacy (previously illustrated in Figure 2.1) and score report interpretation (see Figure 2.2) run diagonally in two of the planes. Data Habit of Mind is the combination of the two planes (with the shared thinking levels axis)
and can be thought to run across a three-dimensional space. However, due to the difficulty of illustrating the 3-dimensional space, it is instead visually represented in two dimensions, with score report interpretation and statistical literacy on the two planes in Figure 2.3.

![Figure 2.3](image)

Figure 2.3. A conceptual model of Data Habit of Mind: Score report interpretation (Figure 2.2) is dependent on statistical literacy (Figure 2.1).

When the two planes are combined, a 3-dimensional space is constructed. The diagonal line is theoretical; it is not expected that educators will develop at the same rate on each plane. The diagonal line indicates that the Data Habit of Mind functions on these two planes and educators may be placed within this space. Some educators (possibly mathematics teachers) may have very high statistical literacy, yet not see how one might use information from score reports. On the other hand, some educators may have a very well-grounded score report interpretation due to lots of professional development, yet have low statistical literacy due to poor performance in mathematics in school.

Habits of mind do not exist in a vacuum. Keating (1990) and Sternberg (1989) describe cognitive activity as needing to be examined in the context of where a person is learning.
Similarly, ability to use data and to think analytically about data is influenced by both pressures and supports for data use. Further, educators act within a community, and are often governed by policy and mandates that affect the culture of data use within a school. Linn states (1998), “most of the uses of assessment programs where the need for an evaluation of the consequences is greatest are determined by governmental bodies such as legislature and school boards” (p. 28). His interpretation is that there is a need to examine how school cultures operate in terms of how they use data in light of provisions for data use and the consequences associated with testing. Thus, the next chapter is an initial conceptualization of what this relationship may look like.
Chapter 3. Conceptual Framework

Investigating Data Habit of Mind as a concept is larger than the single interaction between a score report and an educator interpreting the report. The conceptual drawing in Figure 3.1 is a proposed hypothesis of how data use can impact student scores. The elements (shown as circles) are hypothesized to work together through influencing one another. The main connection is between Data Habit of Mind and the way in which an educator uses data enacted in the classroom. However, influences may be mediated through the culture of data use within the school and/or the design of score reports. As well, the regional policies to which the school is bound also impact the culture of data use.

Figure 3.1. How Data Habit of Mind may lead to improvement in score report interpretation.
Culture of Data Use

Understanding the data use policy within a school board, district or region is complex. Ben Jaafar and Earl (2008) developed a framework that can highlight the intricate differences amongst different jurisdictions. The framework is comprised of five dimensions: Testing Structure, Standard Setting, Consequential Use of Data, Reporting, and Professional Involvement. The 5-dimensional framework has been successfully used to characterize accountability policy across Canada, among school districts in Ontario and Alberta, across the United Kingdom, in Qatar, and comparatively between the United Kingdom and Canada.

While the dimensions of the framework are intended for multiple types of tests and assessments, the framework in this study will be used mainly for the Consequential use of data dimension. The dimension is designed to examine the way in which data are used, either positively or negatively, as described in the jurisdictional policy. Using this dimension, consequences and the context of data use may be differentiated in understanding how policy may then affect the motivators of data use within a school.

Motivation for Data Use

Developing a Data Habit of Mind is often influenced by contextual motivators. Thus, it is imperative to understand the motivational factors within a school community. Two aspects of motivation that contribute to data use have been highlighted by researchers (Deci & Ryan, 1987; Sutherland, 2004): enthusiasm and commitment. Educators working within a school culture that is enthusiastic about how reports may enhance their understanding of children are more likely to use the data. As well, educators in a school culture that holds great value in and commitment to examining reports are more likely to collaboratively use the data to enhance their own teaching practice and student’s learning.
Essentially, engaging in conversations that are centred on the discussion of data is about making time and respecting the information that can be learned from data. These are more general aspects and have to do with the culture and nature of the school environment. Earl and Timperley (2008) write that,

Knowledge is created through dialogue or conversations that make predispositions, ideas, beliefs and feelings explicit and available. It is in these conversations that new ideas, tools and practices are created, and mutual knowledge is either substantially enriched or transformed during the process. (p. 2).

They suggest an understanding that is continually evolving through interaction with others.

Research Questions

As described earlier, the main purpose of the current study was to understand how score reports are interpreted by educators and to examine educators’ beliefs about how score reports may be used to promote student learning. Thus, the research questions focus on the highlighted circle in Figure 3.1 (i.e., Data Habit of Mind), which was described in the literature review as a combination of statistical literacy and score report interpretation.

Research Question 1. How are educators proficient in statistical literacy?
   a. How do educators apply different skills to solve statistical tasks?
   b. Are there patterns in how educators apply skills in solving statistical tasks?

Research Question 2. How do educators interpret and propose to use test score reports?
   a. How do educators apply different skills when interpreting and proposing a use for test score reports?
   b. Are there patterns in how educators apply skills when interpreting and proposing a use for test score reports?
Research Question 3. What are the relationships between statistical literacy and score report interpretations to model Data Habit of Mind?

   a. How may Data Habit of Mind skills be interrelated?
   b. How may educators be categorized on their Data Habit of Mind?

Research Question 4. How may an educator’s educational history and level of comfort with information, mathematics, and statistics contribute to Data Habit of Mind?

   a. How may educational history contribute to Data Habit of Mind?
   b. How may level of comfort with information, mathematics, and statistics contribute to Data Habit of Mind?
Chapter 4. Methods

The research questions were investigated through qualitative methods. Often, in research, the method is tied to the researcher’s paradigmatic stance, particularly the decision to use qualitative or quantitative methods. It is standard in program evaluation for researchers to state their explicit values in conducting research (see the Utility Standard in Yarbrough, Shula, Hopson & Caruthers, 2011). The researcher views himself as a pragmatist and applies a methodological approach that suits the study rather than being driven by a stance. The stance of pragmatism has grown from the field of mixed methods, and has come to describe individuals who endorse, among many aspects, practicality (Johnson & Onwuegbuzie, 2004). Thus, the methodology chosen for this study was intended to be practical to implement, and the results are hoped to be practical for researchers, practitioners and policy makers. There are four parts to this chapter: (1) Sampling, (2) Cognitive Interviews, (3) Instruments, (4) Analysis Procedures.

Sampling

Participants for the interviews were recruited through advertisements. These included a flyer (see Appendix C) posted in buildings at the University of Toronto, email messages to professional contacts, and advertisements on Facebook and LinkedIn. The majority of participants came through word of mouth within schools.

The sample was very diverse. Although originally it was expected that all individuals in the sample would be teachers, some of the participants were working in non-teaching roles, although all were trained as teachers and were working within education. One participant is a vice principal, another is a school lead for her grade level and works on many board initiatives, yet another is a special education consultant and works with several schools within the board, another is a literacy consultant, and one educator is responsible for all home schooling. This
research study adhered to the University of Toronto Ethical Protocol, and all participants who participated in the study (including in the pilot) signed a consent form to participate (see Appendix C). Additionally, all participants were given a $35 gift card for a local bookstore.

**Cognitive Interviews**

The terms “cognitive interviews” and “think-aloud verbal protocols” are often used interchangeably. The differences between the two are the purpose and the social interaction that may happen between interviewer and participant. Think-alouds are usually used to understand a cognitive process when solving problems and the social interaction is standardized and controlled; it is expected that the interviewer does not veer outside the script and the probes are often minimal, without an indication of judgment (Ericsson & Simon, 1993). Cognitive interviews, in contrast, are intended to reveal the underlying cognitive process. Willis (2005) points out that this process has an investigative focus:

> The interviewer can focus on particular areas that appear to be relevant potential sources of response error by actively searching for problems. Put another way, this allows us to examine what may be thought but left unstated, even under a think aloud. (p. 3)

Cognitive interviews allow the interviewer more flexibility during the interview process to adapt probes based on the individual interview to gain insight into the participant’s thought process (Grabowski, Bailey, Bensky, & Link, 2009; Willis, Lawrence, Thompson, Kudela, Levin, & Miller, 2005). The purpose of this study is to understand educators’ cognitive processing when interacting with score reports and, thus, cognitive interviews are more appropriate than the think-aloud method. Willis’s (2005) approach is followed in this study.

The interview questions (Appendix D) were first pre-pilot tested with 3 graduate students (2 females and 1 male), then in a pilot study with 12 pre-service Master’s of Arts students (11 females and 1 male) at the University of Toronto. The pilot study students were all from a
mathematics teaching methodology course, taught by the same instructor, and all were planning on teaching in the future. Changes were made to the interview questions based on practicality.

During the pilot, the researcher had to skip some of the statistical literacy questions to be able to complete the rest of the interview. Also, during the pilot, participants were nervous about answering the questions, as it seemed to them more like a test than an interview. Lastly, during the pilot, a participant started crying as a result of not being able to solve the second task. To remedy these problems, only a few statistical literacy tasks (the shortest ones) were chosen and the following points were added to the interview instructions:

1. We are interested in understanding your thinking, so please don’t worry about getting the right answer.
2. You don’t have to complete the question, just start and I will stop you and ask you what you are thinking.
3. This is not a test to see how well you do, but to understand your thought process to solve these questions.

The relationship between the questions during the interview, the research questions, and coding schemes is shown in Appendix E. The statistical literacy questions are related to the first research question to understand educators’ levels of statistical literacy and are coded to correspond to the conceptual framework presented in Figure 2.1. The score report interpretation questions are related to the second research question to understand how educators read and interpret score reports and are coded based on the conceptual framework presented in Figure 2.2. The questions related to experiences and beliefs, which relate to the third research question to understand if there is a relationship between background and statistical literacy or score report interpretation, were coded descriptively (e.g., number of years of experience, highest degree).
Instruments

Statistical Literacy

The first section is an adaptation of a series of statistical literacy tasks developed by Mooney (2002) to characterize middle school children’s cognitive processing. Mooney consented to the use of these tasks for the purposes of this thesis (Appendix F). In his study, children were given each task and asked to explain how they solved the task. After reviewing Mooney’s (2002) article and piloting the questions, a subset of questions was chosen to garner the most information from the shortest questions (Appendix D). The tasks were ordered from the simplest to the most complicated. The first task asked participants to complete a bar graph, then answer a series of questions. The second task asked participants to reorder a set of numbers in male/female groupings, draw a graph based on the data and then answer a series of questions. The third task is a pictograph; here participants were to complete the graph, draw a different graph, and compare and contrast the two. The last task asked participants a series of questions based on a scatterplot. Figure 4.1 is the first task (all four tasks are in Appendix D).
Olympic Medal Task

The graph shows the medals won at the 1998 Winter Olympics for five countries. The graph is not finished. The table shows the medals won by Italy and China.

![Medals Won at the 1998 Winter Olympics](image)

<table>
<thead>
<tr>
<th>Country</th>
<th>Gold</th>
<th>Silver</th>
<th>Bronze</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>China</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

Source of data: CBS Sportsline (http://www.sportsline.com).

Interviewer: The first one is called the Olympic medal task, the graph show the medals won at the 1998 Olympics for 5 countries the graph is not finished the table shows the medals won by Italy and China. So can you use the information in the tables to complete the graph?

Interviewer: Explain how you completed the graph.
Interviewer: How do the countries compare in medals won?
Interviewer: Which country won the most gold medals?
Interviewer: How can you tell?
Interviewer: And which country won the fewest silver?
Interviewer: How can you tell?
Interviewer: What would you say is the typical number of gold medals won by the 5 countries?
Interviewer: Can you tell me what your thinking was?

Figure 4.1. Olympic medal task (part of the statistical literacy tasks presented in Appendix D).

Score Report Interpretation

Two fictional score reports were designed for this study, based on a review of score reports from several testing organizations. The first report was intended to be formative, and was
developed to resemble reports that provide students with individual skill-based feedback. The second report was based on the local provincial testing agency conventions.

**Cooper’s Fitness Test Score Report**

The first report was formative in nature and presented fictional data on the 12-minute run, a familiar test to educators, since many children in Canada take the 12-minute run as a fitness test in elementary and secondary school (Appendix D). Also, the concept of running is understood by most people. A report on a 12-minute run might be something a student would get. Table 4.1 is an adaptation from Table 2.2 in the literature review and highlights the elements that are intentionally included in the Cooper’s Fitness Test Score Report. The elements in Table 4.1 are also the common elements found in many reports that are designed to be used formatively.

Table 4.1

*Elements in the Cooper’s Fitness Test Score Report*

<table>
<thead>
<tr>
<th>Indicators of Effective Reporting of Diagnostic Information</th>
<th>Guidelines from Standards &amp; Codes for Test Reporting</th>
<th>Guidelines from Measurement Research</th>
<th>Experts’ Suggestions for Visually Displaying Data</th>
<th>Web Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suggestions on how to interpret results while taking into account test properties</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpretation guides</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple measures to report on overall students’ performance</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information to allow for group comparisons</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple methods of reporting</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Methods to facilitate learning</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Explanatory information on the reports</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Relevant information</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Indicators of Effective Reporting of Diagnostic Information</td>
<td>Guidelines from Standards &amp; Codes for Test Reporting</td>
<td>Guidelines from Measurement Research</td>
<td>Experts' Suggestions for Visually Displaying Data</td>
<td>Web Design</td>
</tr>
<tr>
<td>----------------------------------------------------------</td>
<td>-----------------------------------------------------</td>
<td>--------------------------------------</td>
<td>-----------------------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Customized information based on user or group</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision of measurement at the subscale level</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graphs answer elementary to overall levels of questions</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Numbers on report are rounded</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information in tables is ordered meaningfully</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rows and columns in tables are summarized</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>More important elements are proportionally larger and more evident than others</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communicating with priority</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Display of quantitative elements</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Similar to many reports, the title and score are disproportionally large, intended to draw the reader’s attention. The tabular format is intended to replicate a breakdown of test items, with increasing difficulty. Maximum and average speeds are presented as an additional summary. The distribution bars are intended to be cross comparative. Within the distribution bars, the individual’s score is represented as a range, intended to suggest to the reader the possibility of error in the student’s score. Finally, the element that differentiates this report from many others is an area devoted to improvement, with suggestions based on the individual student’s performance. The second page was devoted to regional comparisons, similar to the Ontario provincial testing reports.
**Jabberwocky Test Score Report**

The next fictionalized report is based on a nonsense concept of Jabberwocky and structurally based on the score reports used by the provincial testing program in Ontario (Appendix D). Jabberwocky is a beast in Lewis Carroll’s famous poem written in 1872. Since its publication, there have been many uses and references to the poem and even a dedicated website (www.jabberwocky.com). The poem removes our association with words and lets us focus on the rhythm and structure of language. It is often used in English classes to help children develop structure in their writing. Two sub-components, translating and combining, were included in the report, instead of the traditional reading, writing, mathematics or science. Translating is being able to translate words into nonsense words and combining is being able to combine those words to make sentences that are coherent but essentially nonsensical. Table 4.2 highlights the design elements of the Jabberwocky Test Score Report.
As described earlier, the report is designed based on Ontario’s provincial testing score reports. There are two main features of this report: a main chart that outlines the individual student’s levels and a summary report at the provincial level. This report was presented after the Cooper’s Fitness Test Score Report. When the order was reversed (in the pilot), participants
focused on the political aspects of provincial testing; this was not the case when the Jabberwocky Test Score Report came second.

**Educational History and Comfort with Information, Mathematics, and Statistics**

The second section of the interview consisted of two open-ended questions about the participant’s educational history (i.e., their undergraduate degree major) and the second question asked about their comfort level with information, mathematics, and statistics. The intention in placing educational history and comfort level questions after the statistical literacy tasks was to have as confident a baseline reading of the participant’s statistical knowledge as possible, prior to forming an interviewer-participant relationship. Specific wording of the questions used appears in Appendices D and E.

**Analysis Procedures**

A modified protocol analysis was used to examine the cognitive interviews. Protocol analysis, in the formal sense that was developed by Ericsson and Simon (1993), is used for strict think-aloud protocols. However, in this study, the steps of analysis have been modified to suit the cognitive interviews. The procedure for analysis is a modification of the seven stages Hughes and Parkes (2003) used in their protocol analysis, based Ericsson and Simon’s 1993 text on protocol analysis. There are other sets of stages or steps that may be used in analysis. For example, Willis (2005) has used the work of Bolton and Bronkhorst (1996), which stops at the coding stage. The Bolton and Bronkhorst procedure is simpler and does not involve an encoding stage as the Ericsson and Simon (1993) procedure. However, encoding is critical to this study, because the analyses are in part a test of relationships hypothesized from the literature. The following stages were used in this thesis and may be called “a modified Ericsson and Simon procedure.”
Stage 1: Recording Verbalization

All interviews were conducted during the spring of 2010 and were recorded using a digital recorder.

Stage 2: Transcribing Interviews

Verbatim transcripts were created during the summer of 2010.

Stage 3: Segmenting Interviews

The segmentation process had two phases. The transcriptions were first segmented into three sections: (1) statistical literacy tasks, (2) experience, and (3) score report interpretation. The three sections were then further segmented based on the tasks and questions within each section. Each participant’s transcript was broken down by question for organizational purposes. This stage is needed to organize the data for the encoding process.

Stage 4: Encoding Episodes

The encoding process also had two phases. In the first phase, the participants’ responses to each segment of the statistical literacy tasks and the score report interpretation tasks were examined for evidence of the four thinking levels (Idiosyncratic, Transitional, Quantitative, Analytical). Codes were assigned using the examples in Tables 4.3 (statistical literacy) and 4.4 (score report interpretation).

Table 4.3

<table>
<thead>
<tr>
<th>Thinking Level</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idiosyncratic (coded 1)</td>
<td>This was demonstrated when an educator showed a unique but inefficient or incorrect solution, or refused to respond to the question.</td>
<td>Question (Representing): Can you use the information in the table to complete the graph? “In the table, okay, I could – silvers... gold is... just a couple…. Bronze is solid, um.” [participant gave up]</td>
</tr>
</tbody>
</table>
Table 4.4

Score Report Interpretation: Thinking Levels, Definitions, and Examples

<table>
<thead>
<tr>
<th>Thinking Level</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idiosyncratic (1)</td>
<td>This was demonstrated when an educator just commented without drawing reference to particular details of any part of the report.</td>
<td>Question: Please explain the report to me.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“I can see those types of kids doing better, athletes, than your average population, umm, your jocks and whatever. That’s what I kept thinking looking at it. Is this actually real?”</td>
</tr>
<tr>
<td>Transitional (2)</td>
<td>This is demonstrated when the educator is referring to the report and is beginning to gain an understanding of the report.</td>
<td>Question: I would like you to focus on the table that shows time, distance and accumulation – what does this mean to you?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“I’m not entirely sure what this is, why there are two colours here. Anyhow, maybe it’s distance and accumulation. Um.”</td>
</tr>
<tr>
<td>Thinking Level</td>
<td>Definition</td>
<td>Example</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Quantitative (coded 3)</td>
<td>This is demonstrated when the educator is able to detail quantitatively the elements of the report but does not draw relationships between elements of the report.</td>
<td>Question: Let’s pretend you’re P. Smith’s trainer. What does this section tell you? “If I were the student’s coach, [I would say,] ‘I’m very pleased that you have met the standard but that doesn’t mean that I’m satisfied completely.’ I would say that it seems quite clear to me that you have the potential of reading at a level 5, like to work on getting you to a level 5. I would say that it seems quite clear to me that you have the potential of reading at a level 5.”</td>
</tr>
<tr>
<td>Analytical (coded 4)</td>
<td>This is demonstrated when an educator draws relationships between the different elements on the report to begin to draw a conclusion.</td>
<td>Question: I would like you to focus on the table that shows time, distance and accumulation – what does this mean to you? “She, for argument’s sake, they timed her and her accumulated distance was 2249 meters and it looks like she spent most of her energy in the first bit of the run and as she was running her accumulated distance was 2249 … it looks like they timed her for 12 minutes as well… What else can I tell you? She must have slowed down near the end because the distance she ran in her 12th minute was like … significantly less than her first minute.”</td>
</tr>
</tbody>
</table>

The second phase of the encoding process was conducted by taking segments from each interview and placing them within a matrix. The statistical literacy matrix is thinking level (Idiosyncratic, Transitional, Quantitative, Analytical) by statistical concept (Describing, Organizing and Reducing, Representing, Analyzing and Interpreting). The score report literacy matrix is thinking level (Idiosyncratic, Transitional, Quantitative, Analytical) by score report interpretation level (Describing, Summarizing, Questioning, Proposed Application).

**Stage 5: Analysis of the Codes**

Once the coding was completed, summary scores were created by assigning numeric values to the thinking level codes (Idiosyncratic = 1, Transitional = 2, Quantitative = 3, Analytical = 4) assigned to each of the four statistical literacy concepts and to each of the four types of interpretation for the two score reports. Overall statistical literacy and score report
interpretation scores (separately for the Cooper’s Fitness Test Score Report and the Jabberwocky Test Score Report) were created by averaging these values across concepts or interpretation type, respectively. Tables were created to show the distributions of these scores. Educators are ordered in each of these tables by their overall statistical literacy score, for ease of comparison across tables. The relationship between statistical literacy and score report interpretation is presented visually in Figure 6.1

**Stage 6: Categorizing Educators**

Educators’ codes were examined using basic descriptive statistics, then the relationships among the codes were analyzed through a series of cross-tabulations. The cross-tabulations were also examined in relation to the educators’ educational history and their level of comfort with information, mathematics, and statistics to find possible relationships.

**Relationship of Analysis to Research Questions**

The results of Stage 5 address Research Questions 1 and 2. The results of Stage 6 address Research Questions 3 and 4. The first two stages of this process are found in many research studies, as it is routine to record interviews (Stage 1) and then transcribe them (Stage 2). Stages 3 and 4, which involve segmentation and encoding elements of transcripts into the predefined framework, are found in qualitative studies that have a predefined coding schematic; these stages are specific to protocol analysis (Ericsson & Simon, 1993). The matrixing and analyses in Stages 4, 5, and 6 are methods described by Miles and Huberman (1994). Throughout the presentation of the results, quotes from the participants are used to illustrate the patterns in the data.
Chapter 5. Results

Background Information

In the sample, there were 6 males and 14 females. Their teaching experience ranged from 1 year to 24 years. All of the participants were from the Greater Toronto Area of Ontario, Canada. Table 5.1 provides a snapshot of these participants’ background information.

Educational History

Participants were asked to describe their education: (a) their undergraduate major, (b) whether they have a graduate degree, and (c) whether they have taken any Additional Qualification (AQ) courses. AQ courses are specialized courses offered by universities for teachers to become more qualified in subjects such as mathematics, arts, special education, and drama. All the participants in the sample have taken AQ courses; some have taken several and have multiple specializations. In many school boards, AQ courses are a way to move from one pay grade to the next, so there is a financial incentive for many teachers to take AQ courses. Table 5.1 also provides the participants’ baccalaureate major and whether they have completed a graduate degree. AQ courses and certifications were not included in this table.

Most of the participants had received a Bachelor of Education (B.Ed.) degree as part of a concurrent (usually five-year) program in which students earn both a B.Ed. and a Bachelor of Arts (B.A.) or Bachelor of Science (B.Sc.) or a consecutive (post-B.A. or B.Sc.) program. Only one participant attended a program which offered a four-year specialized degree in education. Seven of the participants had received a Master’s of Education (M.Ed.) degree, of whom one had recently completed her Doctor of Philosophy (Ph.D.) degree specializing in education and psychology.
Table 5.1

Participants’ Demographics, Experience, and Comfort Levels

<table>
<thead>
<tr>
<th>Participant</th>
<th>Sex</th>
<th>Years of Experience</th>
<th>Role</th>
<th>Undergraduate Major</th>
<th>Highest Degree Earned</th>
<th>Favourite Topic</th>
<th>Experience</th>
<th>Comfort Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>Male</td>
<td>1-5</td>
<td>Teacher</td>
<td>HSS</td>
<td>B.Ed.</td>
<td>Math</td>
<td>General</td>
<td>General</td>
</tr>
<tr>
<td>E2</td>
<td>Female</td>
<td>21+</td>
<td>Board Consultant</td>
<td>HSS</td>
<td>B.Ed.</td>
<td>Math</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>E3</td>
<td>Female</td>
<td>16-20</td>
<td>Teacher</td>
<td>HSS</td>
<td>B.Ed.</td>
<td>Literature</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>E4</td>
<td>Female</td>
<td>1-5</td>
<td>Supply Teacher</td>
<td>HSS</td>
<td>B.Ed.</td>
<td>Mapping Skills</td>
<td>General</td>
<td>High</td>
</tr>
<tr>
<td>E5</td>
<td>Male</td>
<td>1-5</td>
<td>Teacher</td>
<td>HSS</td>
<td>M.Ed.</td>
<td>Oil Painting</td>
<td>Low</td>
<td>None</td>
</tr>
<tr>
<td>E6</td>
<td>Female</td>
<td>1-5</td>
<td>Supply teacher</td>
<td>HSS</td>
<td>M.Ed.</td>
<td>Math and Art</td>
<td>Learning</td>
<td>High</td>
</tr>
<tr>
<td>E7</td>
<td>Female</td>
<td>1-5</td>
<td>Teacher</td>
<td>Science</td>
<td>B.Ed.</td>
<td>Science</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>E8</td>
<td>Female</td>
<td>1-5</td>
<td>Supply teacher</td>
<td>HSS</td>
<td>M.Ed.</td>
<td>Media Literacy</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>E9</td>
<td>Female</td>
<td>1-5</td>
<td>Teacher</td>
<td>HSS</td>
<td>B.Ed.</td>
<td>Life skills</td>
<td>Learning</td>
<td>High</td>
</tr>
<tr>
<td>E10</td>
<td>Male</td>
<td>6-10</td>
<td>Teacher</td>
<td>HSS</td>
<td>B.Ed.</td>
<td>Math</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>E11</td>
<td>Male</td>
<td>6-10</td>
<td>Teacher</td>
<td>HSS</td>
<td>B.Ed.</td>
<td>Literature</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>E12</td>
<td>Male</td>
<td>11-15</td>
<td>Teacher</td>
<td>HSS</td>
<td>B.Ed.</td>
<td>Art</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>E13</td>
<td>Female</td>
<td>6-10</td>
<td>Teacher</td>
<td>HSS</td>
<td>B.Ed.</td>
<td>Literacy</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>E14</td>
<td>Male</td>
<td>11-15</td>
<td>Teacher</td>
<td>HSS</td>
<td>M.Ed.</td>
<td>Math &amp; Science</td>
<td>General</td>
<td>High</td>
</tr>
<tr>
<td>E15</td>
<td>Male</td>
<td>21+</td>
<td>Board Consultant</td>
<td>HSS</td>
<td>B.Ed.</td>
<td>History</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>E16</td>
<td>Female</td>
<td>11-15</td>
<td>Board Consultant</td>
<td>Science</td>
<td>B.Ed.</td>
<td>Science</td>
<td>General</td>
<td>High</td>
</tr>
<tr>
<td>E17</td>
<td>Female</td>
<td>1-5</td>
<td>Teacher</td>
<td>HSS</td>
<td>M.Ed.</td>
<td>Computers</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>E18</td>
<td>Female</td>
<td>6-10</td>
<td>Administrator</td>
<td>HSS</td>
<td>Ph.D.</td>
<td>Math</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>E19</td>
<td>Female</td>
<td>6-10</td>
<td>Teacher</td>
<td>HSS</td>
<td>M.Ed.</td>
<td>Literacy</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>E20</td>
<td>Female</td>
<td>6-10</td>
<td>Teacher</td>
<td>Science</td>
<td>B.Ed.</td>
<td>Science</td>
<td>Learning</td>
<td>Learning</td>
</tr>
</tbody>
</table>

Note: Participant E3 received a 4-year degree in education; HSS = Humanities and Social Sciences; participants responded using a 10-point scale, which was then converted to Low (1-3), Medium (4-6), and High (7-10).
Most of the participants’ undergraduate majors were in the humanities or social sciences; only three had a science-focused degree. None of the participants had a mathematics-focused program as their undergraduate degree.

**Favourite Topics and Comfort Levels**

Participants were asked to describe their favourite topic, lesson or activity. This question was intended to identify participants’ subject area strengths. They were also asked about their experiences with assessment and data and to rate their comfort levels with information, mathematics and statistics. Table 5.1 includes a summary of these responses.

Some \((n = 6)\) of the participants said mathematics was one of their favourite topics, although none of them majored in mathematics in university. Other popular favourites were literature or literacy \((n = 5)\), science \((n = 4)\), arts or maps \((n = 3)\), and some had individual interests, such as computers, life skills, and history.

Participants’ experiences with assessment varied. Three participants said they are constantly learning. These three participants expressed that they are open to new learning even though they have had a great deal of professional development on the topic of assessments. Half the participants believed they have had a lot of experience with assessment, based on how much professional development they have received, if they have been involved in administering the provincial assessments at Grades 3, 6, 9, or 10, or worked with standardized tests for special populations. Some \((n = 4)\) said they had a general or average amount of experience with assessment. Three said they have little experience with assessment; two of these three said they also have little experience with data, while the third said he has high experience with data. This third teacher had taken a statistics course in university, in which he analyzed data.
There were four other participants who drew a distinction between assessment and data by stating that they have high levels of experience with data and were learning to work with assessments. Another two participants believed they had a great deal of experience with assessment but little experience with data. As Table 5.1 shows, almost half the participants (9 out of 20) rated their experience with assessment the same as their experience with data.

Participants were asked to rate on a scale of 1 (low) to 10 (high) their comfort levels in reading information from media such as newspapers or a school report, mathematics and statistics. For Table 5.1, these ratings were divided into low (1 to 3), medium (4 to 6), and high (7 to 10). All participants rated themselves as high in reading information that may be graphically displayed and in which numerical and written information is combined. Participants’ comfort levels in mathematics varied. All participants believed there are two different types of math: (1) the math they teach and (2) higher-level mathematics, including statistics. Eleven participants had a high level of comfort with the mathematics they teach in elementary school, 9 had a moderate level of comfort with the mathematics they teach. The variability in responses was larger when asking about statistics; of the 11 teachers who had a high comfort in mathematics, one stated she has a low comfort level with statistics, the other 10 remained high. However, almost half ($n = 4$) of the 9 participants who said they have a moderate level of comfort with math said they have a low level of comfort with statistics; the rest stayed at the same moderate level.
Statistical Literacy: Research Question 1

Describing

Participants had minimal difficulty in describing the tasks, except the task related to the graph in Figure 5.1. The graph is a scatter plot with the number of area codes plotted against a state’s population.

![Graph showing state population and area codes](image)

*Figure 5.1. Part of the state population and area codes task (a statistical literacy task).*

Many participants had an “internal dialogue,” captured in the cognitive interview, before coming to a correct solution. The following is an example of this dialogue:

Participant E20: Okay, these are the number of area codes and it tells me by the dot how many states … What’s the question again?

Interviewer: How many states have five area codes?
Participant E20: Okay, let me read this … okay … How many states have 5 area codes? … These are the codes here and each dot is a state? 1,2,3,4,5 … but this is the population in millions. I’m not sure I have the information you want. I’m not sure I get this.

Interviewer: You’re on the right track. Each dot is a state.

Participant E20: Okay, 5…

Once participants understood that each dot represents a state they did not have difficulty completing the task at a high thinking level.

**Organizing and Reducing**

The first two tasks ask participants to find a typical number in a set of data. In the Olympic medal task, the dataset is so small that most participants were able to average the numbers in their head. For example, many participants had thinking similar to the following:

Participant E3: Well, I’ll take typical to mean average, so, well, China got 0, 0-6, so, you go, 6 + 5 is 11, 13, and 15 ok, 1, 2, 3, 4, 5, so, I would say, 3.5 countries, the average amount would be 3.

However, for the second task when they were asked for the typical number for 15 actresses, there were more complications. Many participants were like Participant E3, eventually choosing a median, or mode to represent the data instead of an average.

Interviewer: What is the typical salary for actresses?

Participant E3: Well … you would add all of these up, … but you could kind of go to the median, too…. so that’s 1, 2, 3, 4, 5, 6, 7…. Well, the median would be about 5, but the average would be about …. Hmm, I don’t know – 10, 20, 30, 40, 70 … I don’t know.

Interviewer: Which one do you think would be more representative?

Participant E3: You’d have to compare them …. Yeah, I’d have to do both.

From this task and the combination of the first two tasks, many of the participants were able to switch the way they work with data. When faced with a novel concept of typical, which is
supposed to encompass the idea of central tendency, all participants were able to translate the concept of represent a set of data by one number. However, all were initially perplexed by the use of the word *typical*.

**Representing**

Participants had difficulty in representing information or data that required anything other than a bar graph. This was most evidently found when participants were asked to make a graph that would compare the salaries of actors to actresses, provided in tabular form, as shown in Figure 5.2. There were three main strategies used when participants were able to complete this task. The most popular method was to use a bar graph, where actors were ranked from highest to lowest and actresses were also ranked and compared by ranking.

**Figure 5.2.** Part of the movie stars salary task (a statistical literacy task).

The thought process by which the bar graph emerged was fairly common across participants:

<table>
<thead>
<tr>
<th>Salaries of 15 Top Actors and Actresses (in Millions of Dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Actors</strong></td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>$17.5</td>
</tr>
<tr>
<td>15.0</td>
</tr>
<tr>
<td>20.0</td>
</tr>
<tr>
<td>20.0</td>
</tr>
<tr>
<td>20.0</td>
</tr>
<tr>
<td>19.0</td>
</tr>
<tr>
<td>20.0</td>
</tr>
<tr>
<td>18.0</td>
</tr>
<tr>
<td>5.5</td>
</tr>
<tr>
<td>6.0</td>
</tr>
<tr>
<td>10.0</td>
</tr>
<tr>
<td>16.5</td>
</tr>
<tr>
<td>12.5</td>
</tr>
<tr>
<td>10.0</td>
</tr>
<tr>
<td>7.0</td>
</tr>
</tbody>
</table>
Participant E7: So, maybe should I do a bar graph?

Interviewer: Up to you.

Participant E7: So, actors … so, how do I do this? … 20 million up here and zero million … yeah … dark… actresses 10 million… is that gonna work? …

Interviewer: Can you please explain your thinking?

Participant E7: So, um …. I was thinking that I was just going to pair up the actors and actresses from most money to least money and draw a bar graph for each couple or pair to show that the actor making the most money who is like a man and the woman beside it who is making the most money to compare the two together.

Many participants described and drew a graph that looked like Figure 5.3. Others attempted to create the same graph but using lines rather than bars, and another took a while but she averaged the salary for men and salary for women and just drew two bars, one for men and one for women. In general, participants displayed information and data in horizontal or vertical bar graphs. The greatest range in thinking levels was found in this skill.

Figure 5.3. A common graph that participants tried to draw when asked to draw a graph to represent the table in Figure 5.2.
Analyzing and Interpreting

Part of the statistical literacy tasks asked educators to analyze and interpret data. These were often more complex tasks, as educators were asked to compare and contrast different types of graphs or to figure out, for example, if Canada was to become a U.S. state, how many area codes it would need if its population was 31 million. These tasks required participants to predict or hypothesize about a situation. In the third task, participants were asked to draw a pictograph, then a bar graph and compare how they would be similar or different. One participant who was struggling with many of the tasks responded by stating:

Participant E11: Just visually it’s nicer, easier to read, use colours, whatever to represent.

Interviewer: Well, how do the two graphs compare?

Participant E11: They would …. The pictograph is easier in giving you the hard numbers.

Interviewer: How would they be alike?

Participant E11: They’re both representing similar kinds of data.

It was difficult to obtain a response from the participants on any task that required them to extend beyond the simple descriptive. Most commonly, participants would state that both the bar graph and pictograph are visual and the bar graph is more accurate. With this participant, it was difficult to determine if he actually did not know how they were similar or different, if the task was too simple, or if he was looking for a trick in the questioning.

Distributions

As described earlier, each participant’s thinking level was coded (Idiosyncratic = 1, Transitional = 2, Quantitative = 3, Analytical = 4) for the four statistical literacy skills (Describing, Organizing and Reducing, Representing, Analyzing and Interpreting). As thinking level is assumed to be on an ordinal scale, these codes were then averaged across the skills to
obtain an overall statistical literacy score. Table 5.2 provides the participants’ codes and the overall score. The higher the overall score, the higher the thinking level. For example, Participant E20 was coded as 3.7 when Describing, 2.5 when Organizing and Reducing, 3.7 when Representing, and 3.5 when Analyzing and Interpreting. This means Participant E20 has a high level of thinking (at the Quantitative or Analytical level) when it comes to Describing, Representing, and Analyzing and Interpreting data, but seems to be more Transitional when it comes to Organizing and Reducing data. The participants are ordered from highest performance (E18) to lowest performance (E19) (this is the order for both Tables 5.3 and 5.4, so the reader can compare the tables).

Table 5.2

*Participants’ Statistical Literacy Skills*

<table>
<thead>
<tr>
<th>Educator ID</th>
<th>Describing</th>
<th>Organizing and Reducing</th>
<th>Representing</th>
<th>Analyzing and Interpreting</th>
<th>Overall Statistical Literacy Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>E18</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
</tr>
<tr>
<td>E9</td>
<td>4.0</td>
<td>4.0</td>
<td>4.0</td>
<td>3.8</td>
<td>3.9</td>
</tr>
<tr>
<td>E10</td>
<td>3.7</td>
<td>3.5</td>
<td>4.0</td>
<td>4.0</td>
<td>3.8</td>
</tr>
<tr>
<td>E1</td>
<td>3.7</td>
<td>3.5</td>
<td>3.3</td>
<td>4.0</td>
<td>3.6</td>
</tr>
<tr>
<td>E8</td>
<td>4.0</td>
<td>3.0</td>
<td>4.0</td>
<td>3.5</td>
<td>3.6</td>
</tr>
<tr>
<td>E13</td>
<td>4.0</td>
<td>3.0</td>
<td>4.0</td>
<td>3.5</td>
<td>3.6</td>
</tr>
<tr>
<td>E14</td>
<td>4.0</td>
<td>2.5</td>
<td>4.0</td>
<td>3.8</td>
<td>3.6</td>
</tr>
<tr>
<td>E6</td>
<td>3.7</td>
<td>3.0</td>
<td>3.7</td>
<td>3.8</td>
<td>3.5</td>
</tr>
<tr>
<td>E17</td>
<td>4.0</td>
<td>2.5</td>
<td>3.3</td>
<td>4.0</td>
<td>3.5</td>
</tr>
<tr>
<td>E3</td>
<td>3.7</td>
<td>4.0</td>
<td>3.0</td>
<td>3.0</td>
<td>3.4</td>
</tr>
<tr>
<td>E12</td>
<td>4.0</td>
<td>2.5</td>
<td>3.7</td>
<td>3.3</td>
<td>3.4</td>
</tr>
<tr>
<td>E20</td>
<td>3.7</td>
<td>2.5</td>
<td>3.7</td>
<td>3.5</td>
<td>3.3</td>
</tr>
<tr>
<td>E4</td>
<td>3.7</td>
<td>3.0</td>
<td>3.0</td>
<td>3.3</td>
<td>3.2</td>
</tr>
<tr>
<td>E5</td>
<td>3.7</td>
<td>2.5</td>
<td>3.0</td>
<td>3.3</td>
<td>3.1</td>
</tr>
<tr>
<td>E7</td>
<td>3.7</td>
<td>2.5</td>
<td>3.3</td>
<td>2.8</td>
<td>3.1</td>
</tr>
<tr>
<td>E15</td>
<td>4.0</td>
<td>2.0</td>
<td>2.7</td>
<td>3.3</td>
<td>3.0</td>
</tr>
</tbody>
</table>
Most educators had statistical literacy thinking levels of Quantitative or Analytical; however, three (E2, E11, and E19) were in the Transitional range, although they did well in Describing. Within the Organizing and Reducing data skill, three participants (E15, E16, and E19) are at the Transitional level. The majority of teachers \( (n = 7) \) are between the Transitional and Quantitative levels. Some participants \( (n = 5) \) are at the Quantitative level, and an additional 2 are between the Quantitative and Analytical levels. Three participants (E3, E9, and E18) were coded Analytical. A similar breakdown showed that, when participants were asked to represent information, most participants’ \( (n = 16) \) thinking levels ranged from Quantitative to Analytical. Fewer participants (E2, E11, and E15) were between the Transitional and Quantitative levels. One participant was coded low in comparison to others (E19); this participant was between the Idiosyncratic and Transitional levels.

According to the literature (Jones et al., 2000, Mooney 2002), the most complex skill is Analyzing and Interpreting data. None of the participants was at the lowest level; most \( (n = 15) \) were between the Quantitative and Analytical levels. The remainder were split between the Transitional level (E11) and the Transitional and Quantitative levels (E7, E19, E2, and E3). No educator was coded as having an Idiosyncratic thinking process in Describing, Organizing and Reducing, Representing, or Analyzing and Interpreting data.
Score Report Interpretation: Research Question 2

As described in the Methods chapter, participants were given an opportunity to examine each report in its entirety and were then asked to explain the report. Some participants’ explanations were elaborate, including summaries and conjectures across the different elements and even describing what they would do with this student next. Others were much more descriptive without a link to how the information may inform the participant. Participants’ responses varied across the score report interpretation skills of Describing, Summarizing, Questioning, and Proposed Application.

Describing and Summarizing

Educators can accurately and astutely describe and summarize a report. All participants were coded as Analytical when it comes to Describing a score report. Almost all the participants were also coded Analytical when Summarizing a score report. When explaining the Cooper’s Fitness Test Score Report, one participant said:

Participant E13: I remember doing the 12-minute run …. Um, okay, it’s telling us how much like the distance this kid has run after each minute and giving us the accumulated and tracking how much he does each minute and adding it together to tell us how much in total, um … hmm …. This is harder to understand. I don’t know if I understand this. I know it’s telling us he is of average fitness. It looks to me this would be the average of the province because the school seems to be lower or he did better than the school …. So, yeah, this I understand better than I understand this.

Interviewer: Would you like to add anything else?

Participant E13: Hmm … well, this is telling us what his average was and where he hit his maximum, which is at the beginning and it averaged everything out and where his slower times were and slowed towards the end and dropped somewhere in the middle.
The participant showed she could describe multiple parts of the report and she was able to make links across different elements in the report. She showed a high level of thinking when explaining the report. However, as we will see later in this chapter, Participant E13 had difficulty in the more advanced skill of Proposed Application of what was on the report.

**Questioning**

As illustrated in the above interview excerpt, Participant E13 questioned an element of the report. In both reports, there were elements included for participants to question. The Cooper’s Fitness Test Score Report had a large graphic (see Figure 5.4) on the second page, which was often questioned, as Robin’s bar was disproportionately larger than any other part of the graph in an attempt to represent standard error and was misplaced. As well, more subtly in that report, Robin was reported as Average in the title, low average for her age, and visually on the comparative summary looked as if she was above the regional averages. In the Jabberwocky Test Score Report, the rings (see Figure 5.5), which were intended to represent standard error, were not explained and it was expected that participants would question them.

![Figure 5.4. An element from the Cooper’s Fitness Test Score Report (the thick bar is intended to represent measurement error).](image-url)
Figure 5.5. An element from the Jabberwocky Test Score Report (the rings are intended to represent measurement error).

Proposed Application

There were some differences when it came to participants trying to figure out how they might use the information in the score reports:

Participant E8: The max speed will tell me how fast you can go. I might look at it and say, “You might want to be a sprint runner – if I was your coach, I was looking at where I should place you – I would like to know how fast you can run and how long you can sustain it and if your average speed is much less than your max speed.” Then, the student that I’m looking at wouldn’t be fit for… then they’re not fit to be a long distance runner and they may be faster in the sprinting or short track type of running. Um, I think that basically would be what I look at.

Participant E7: I – that section tells me she is running her maximum speed right very early on and maybe too quickly. Um, because if I was her coach I would probably tell her to pace herself, but I don’t know and it also shows me – let’s look down – her distances get a little bit smaller as the time goes on so maybe she is very tired near the end.

Distributions: Cooper’s Fitness Test Score Report

As described earlier, each participant’s responses were coded (Idiosyncratic = 1, Transitional = 2, Quantitative = 3, Analytical = 4) on four different score report interpretation
skills (Describing, Summarizing, Questioning, Proposed Application). The score report interpretation analysis differs from the statistical literacy analysis in that there was not average across different tasks; educators’ responses were coded once based on a coding scheme provided in Appendix E. Table 5.3 displays participants’ codes on the skills demonstrated when responding to the Cooper’s Fitness Test Score Report. Each row represents a different participant (ranked by their statistical literacy); each participant has four codes, representing Describing, Summarizing, Questioning, and Proposed Application. For example, Participant E20 was coded to be Analytical when Describing and Summarizing the score report, Idiosyncratic when Questioning the score report, and Quantitative for Proposed Application for the report.

Table 5.3

<table>
<thead>
<tr>
<th>Educator ID</th>
<th>Describing</th>
<th>Summarizing</th>
<th>Questioning</th>
<th>Proposed Application</th>
<th>Overall Score Report Interpretation Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>E18</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3.25</td>
</tr>
<tr>
<td>E9</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>3.25</td>
</tr>
<tr>
<td>E10</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3.50</td>
</tr>
<tr>
<td>E1</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3.75</td>
</tr>
<tr>
<td>E8</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3.50</td>
</tr>
<tr>
<td>E13</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2.50</td>
</tr>
<tr>
<td>E14</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3.00</td>
</tr>
<tr>
<td>E6</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3.75</td>
</tr>
<tr>
<td>E17</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3.50</td>
</tr>
<tr>
<td>E3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3.00</td>
</tr>
<tr>
<td>E12</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3.25</td>
</tr>
<tr>
<td>E20</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2.75</td>
</tr>
<tr>
<td>E4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2.75</td>
</tr>
<tr>
<td>E5</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2.75</td>
</tr>
<tr>
<td>E7</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2.25</td>
</tr>
<tr>
<td>Educator ID</td>
<td>Describing</td>
<td>Summarizing</td>
<td>Questioning</td>
<td>Proposed Application</td>
<td>Overall Score Interpretation Score</td>
</tr>
<tr>
<td>------------</td>
<td>------------</td>
<td>-------------</td>
<td>-------------</td>
<td>----------------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>E15</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3.25</td>
</tr>
<tr>
<td>E16</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2.75</td>
</tr>
<tr>
<td>E2</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3.25</td>
</tr>
<tr>
<td>E11</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2.50</td>
</tr>
<tr>
<td>E19</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3.00</td>
</tr>
</tbody>
</table>

All participants scored high on Describing and can be thought of as being able to be Analytical in describing the reports and information on the reports. Educators are able to describe many elements of the report, often going from one part of the report to the next. Additionally, the majority of participants \((n = 15)\) are Analytical when Summarizing the report. However, five participants may be considered Quantitative; that is, they only summarize elements one at a time, but do not extend and connect with other elements on the report.

A greater variability exists when Questioning the report. A few participants \((E2, E16, \text{and} E18)\) ranged from Quantitative to Analytical; that is, they at least questioned more than one element in the report. Most educators \((n = 11)\) only questioned one element of the report and the remainder \((n = 6)\) did not question any part of the report.

For Proposed Application based on the report, many \((n = 7)\) participants were able to describe ways they can use multiple elements of the report. One participant \((E8)\) was able to make connections across different elements of the report. However, the majority of participants \((n = 13)\) did not present any way they might apply the information gained, or were able to apply only one element of the report.
Distributions: Jabberwocky Test Score Report

The results from the Jabberwocky Test Score Report skill codes, shown in Table 5.4, were very similar to those for the Cooper’s Fitness Test Score Report. The same coding scheme was used as described earlier, each participant’s response were coded (Idiosyncratic = 1, Transitional = 2, Quantitative = 3, Analytical = 4) on four different score report interpretation skills (Describing, Summarizing, Questioning, Proposed Application). All were coded Analytical when Describing the Jabberwocky Test Score Report. Most of the participants (n = 15) were coded Analytical when Summarizing the reports. There were fewer participants (n = 3) who asked multiple questions about the report. Most (n = 11) had only one question and the remaining participants did not question anything in the report. The majority of participants (n = 8) did not describe a way in which they might apply the information, and a few (n = 5) described a way they could apply one element. Seven out of the 20 participants Proposed Applications based on more than one element, and only one of those participants made references across different elements in the report.

Table 5.4
Participants’ Score Report Interpretation Skills Demonstrated on the Jabberwocky Test Score Report

<table>
<thead>
<tr>
<th>Educator ID</th>
<th>Describing</th>
<th>Summarizing</th>
<th>Questioning</th>
<th>Proposed Application</th>
<th>Overall Score Report Interpretation Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>E18</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3.75</td>
</tr>
<tr>
<td>E9</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>3.00</td>
</tr>
<tr>
<td>E10</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>3.00</td>
</tr>
<tr>
<td>E1</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3.25</td>
</tr>
<tr>
<td>E8</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3.50</td>
</tr>
<tr>
<td>E13</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2.50</td>
</tr>
<tr>
<td>E14</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2.75</td>
</tr>
<tr>
<td>E6</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3.25</td>
</tr>
<tr>
<td>E17</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3.00</td>
</tr>
<tr>
<td>E3</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3.00</td>
</tr>
<tr>
<td>E12</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2.75</td>
</tr>
<tr>
<td>E20</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3.00</td>
</tr>
<tr>
<td>E4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2.75</td>
</tr>
<tr>
<td>Educator ID</td>
<td>Describing</td>
<td>Summarizing</td>
<td>Questioning</td>
<td>Proposed Application</td>
<td>Interpretation Score</td>
</tr>
<tr>
<td>-------------</td>
<td>------------</td>
<td>-------------</td>
<td>--------------</td>
<td>-----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>E5</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2.25</td>
</tr>
<tr>
<td>E7</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2.25</td>
</tr>
<tr>
<td>E15</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3.00</td>
</tr>
<tr>
<td>E16</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2.75</td>
</tr>
<tr>
<td>E2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>3.50</td>
</tr>
<tr>
<td>E11</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2.50</td>
</tr>
<tr>
<td>E19</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2.75</td>
</tr>
</tbody>
</table>
Chapter 6. Patterns

Data Habit of Mind is a combination of statistical literacy and score report interpretation.

The following analysis will aid in identifying patterns of proficiencies and weakness in participants’ use of data. The analysis also focused on finding discrepancies, such as someone who may be very statistically literate but have lower skills when interpreting score reports. Table 6.1 shows some descriptive statistics for the codes used in this study.

Table 6.1
Descriptive Statistics of Skill Codes

<table>
<thead>
<tr>
<th>Skill</th>
<th>Minimum</th>
<th>Maximum</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Statistical Literacy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Describing</td>
<td>3.00</td>
<td>4.00</td>
<td>3.77</td>
<td>0.24</td>
</tr>
<tr>
<td>Organizing and Reducing</td>
<td>2.00</td>
<td>4.00</td>
<td>2.89</td>
<td>0.65</td>
</tr>
<tr>
<td>Representing</td>
<td>1.67</td>
<td>4.00</td>
<td>3.33</td>
<td>0.65</td>
</tr>
<tr>
<td>Analyzing and Interpreting</td>
<td>2.00</td>
<td>4.00</td>
<td>3.38</td>
<td>0.53</td>
</tr>
<tr>
<td><strong>Score Report Interpretation: Cooper’s Fitness Test Score Report</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Describing</td>
<td>4.00</td>
<td>4.00</td>
<td>4.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Summarizing</td>
<td>3.00</td>
<td>4.00</td>
<td>3.75</td>
<td>0.44</td>
</tr>
<tr>
<td>Questioning</td>
<td>1.00</td>
<td>4.00</td>
<td>2.10</td>
<td>0.091</td>
</tr>
<tr>
<td>Proposed Application</td>
<td>1.00</td>
<td>4.00</td>
<td>2.45</td>
<td>1.10</td>
</tr>
<tr>
<td><strong>Score Report Interpretation: Jabberwocky Test Score Report</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Describing</td>
<td>4.00</td>
<td>4.00</td>
<td>4.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Summarizing</td>
<td>3.00</td>
<td>4.00</td>
<td>3.75</td>
<td>0.44</td>
</tr>
<tr>
<td>Questioning</td>
<td>1.00</td>
<td>4.00</td>
<td>1.95</td>
<td>0.89</td>
</tr>
<tr>
<td>Proposed Application</td>
<td>1.00</td>
<td>4.00</td>
<td>2.00</td>
<td>0.97</td>
</tr>
</tbody>
</table>
The codes for statistical literacy skills of Representing and Analyzing and Interpreting have similar means. The Describing skill has the smallest standard deviation. This was also true for the skills in score report interpretation. For Describing, all educators were coded with the highest thinking level (Analytic). For Summarizing, there was a little more variability; however, for both score reports, educators received a code of 3 (Quantitative) or 4 (Analytical). Of the score report interpretation skills, Questioning and Proposed Application have the greatest variability in codes, with educators ranging from 1(Idiosyncratic) to 4 (Analytical).

Based on Table 6.1, there is less variability in statistical literacy skill level than in the skills required for score report interpretation. This variability may be due to educators’ proficiency when it comes to Questioning and Proposed Application, but it also may be affected by the styles of score reports. The comparisons in this section are intended to shed light on why variability exists among these educators when it comes to higher order skills of questioning and proposing an application from a score report.

**Relationship of Statistical Literacy with Interpretations of the Cooper’s Fitness Test Score Report and the Jabberwocky Test Score Report: Research Question 3**

To investigate the relationship between the participants’ overall statistical literacy and their score report interpretation skills, the average thinking levels across the four statistical literacy skills (Describing, Organizing and Reducing, Representing, Analyzing and Interpreting), and across the interpretation skills (Describing, Summarizing, Questioning, Proposed Application) demonstrated on the Cooper’s Fitness Test Score Report and the Jabberwocky Test Score Report were computed. It is important to note that, although the thinking levels are ordinal, they may not
be strictly interval. Figure 6.1 shows the relationship between participants’ overall statistical literacy level and their overall score report interpretation level.

*Figure 6.1.* Overall score report interpretation level by overall statistical literacy level.

Figure 6.1 shows a positive relationship \((r = .49)\) between the overall score report interpretation level and the overall statistical literacy level. However, as shown in Table 6.1, there
is a great deal of variability in score report interpretation skills, especially when it comes to Questioning and Proposed Application. The following analyses, therefore, focus on the Questioning and Proposed Application score report interpretation skills.

**Four Data Habits of Mind**

Educators vary in their skills and in their values and beliefs about data, statistics, score reports and testing in general. However, analyzing the patterns of educators’ skills may help the education community produce more useful score reports and supporting materials to improve the quality of interpretations and potential uses of tests. Figure 6.2 is an attempt to divide the conceptual model of Data Habit of Mind into four quadrants.

<table>
<thead>
<tr>
<th>Score Report Interpretation</th>
<th>Statistical Literacy</th>
<th>Analytical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idiosyncratic</td>
<td>Lower Statistical Literacy &amp; Lower Score Report Interpretation</td>
<td>Higher Statistical Literacy &amp; Lower Score Report Interpretation</td>
</tr>
<tr>
<td>Transitional</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantitative</td>
<td>Lower Statistical Literacy &amp; Higher Score Report Interpretation</td>
<td>Higher Statistical Literacy &amp; Higher Score Report Interpretation</td>
</tr>
<tr>
<td>Analytical</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 6.2. Categories based on Data Habit of Mind skills.*

Ideally, all educators would be categorized as having higher statistical literacy and score report interpretation skills. To populate this figure with educators, multiple series of cross tabulations were conducted. Statistical literacy was averaged across its skill codes and cross tabulations were performed with Questioning and Proposed Application skills for each score report individually. The score report interpretation skills of Describing and Summarizing are not
included in the figures below, because there was little variability in the codes and thus they act as constants in the analysis.

**Questioning Patterns**

Figures 6.3 and 6.4 summarize the cross-tabulations of the overall statistical literacy score with the Questioning skill of score report interpretation. Educators with overall statistical literacy scores below 2 were placed in the Idiosyncratic category, between 2 and 3 were placed in the Transitional category, between 3 and 3.5 were placed in the Quantitative category, and above 3.5 were placed in the Analytical category. Judgments of who was higher or lower were based on the combination of Idiosyncratic and Transitional categories (lower) and the Quantitative and Analytical (higher) categories.

<table>
<thead>
<tr>
<th>Score Report Interpretation</th>
<th>Statistical Literacy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Idiosyncratic</td>
</tr>
<tr>
<td>Idiosyncratic</td>
<td>E11</td>
</tr>
<tr>
<td>Transitional</td>
<td>E2, E15, E16</td>
</tr>
<tr>
<td>Quantitative</td>
<td>E19</td>
</tr>
<tr>
<td>Analytical</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 6.3. Categorization of participants based on Questioning score report interpretation skill demonstrated on the Cooper’s Fitness Test Score Report by overall statistical literacy score.*
### Table: Categorization of Participants Based on Questioning Score Report Interpretation

<table>
<thead>
<tr>
<th>Score Report Interpretation</th>
<th>Idiosyncratic</th>
<th>Transitional</th>
<th>Quantitative</th>
<th>Analytical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idiosyncratic</td>
<td>E11</td>
<td>E5, E7</td>
<td>E9, E10, E13</td>
<td></td>
</tr>
<tr>
<td>Transitional</td>
<td>E15, E19,</td>
<td>E3, E4, E12,</td>
<td>E1, E6, E8,</td>
<td></td>
</tr>
<tr>
<td>Quantitative</td>
<td>E16</td>
<td>E17, E20</td>
<td>E14,</td>
<td></td>
</tr>
<tr>
<td>Analytical</td>
<td>E2</td>
<td></td>
<td>E18</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6.4.** Categorization of participants based on Questioning score report interpretation skill demonstrated on the Jabberwocky Test Score Report by overall statistical literacy score.

Comparing Figures 6.3 and 6.4 shows how educators’ demonstrate different skills in response to the different score reports. For example, Participant E17 was coded as Analytical for Questioning in response to the Cooper’s Fitness Test Score Report, but was coded as Transitional for Questioning the Jabberwocky Test Score Report. Many educators, however, did not shift their level of Questioning based on the reports (E4, E5, E7, E8, E9, E11, E12, E13, E14, E15, and E18). Two educators (E11, E15) were consistent in their low level of Questioning and a low overall statistical literacy score. None of the educators with lower overall statistical literacy scores had high levels of Questioning to either score report. The majority of educators (E4, E5, E7, E8, E9, E12, E13, E14, and E20) who did not shift based on the type of report had higher overall statistical literacy scores and lower Questioning score report interpretation skills. Only one educator (E18) consistently had a higher statistical literacy score and a higher level of Questioning.

Fewer than half ($n = 7$) of the educators were in different categories for the two score reports. Two educators (E2 and E16) demonstrated higher Questioning on the Jabberwocky Test Score Report than on the Cooper’s Fitness Test Score Report. Five educators (E1, E3, E6, E10,
and E19) demonstrated higher Questioning on the Cooper’s Fitness Test Score Report than on the Jabberwocky Test Score Report.

**Proposed Application Patterns**

Figures 6.5 and 6.6 summarize the relationship between the Proposed Application skills in score report interpretation and the overall statistical literacy score.

<table>
<thead>
<tr>
<th>Statistical Literacy (Overall Score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idiosyncratic</td>
</tr>
<tr>
<td>Idiosyncratic</td>
</tr>
<tr>
<td>Transitional</td>
</tr>
<tr>
<td>Quantitative</td>
</tr>
<tr>
<td>Analytical</td>
</tr>
</tbody>
</table>

*Figure 6.5. Categorization of participants based on Proposed Application score report interpretation skill demonstrated on the Cooper’s Fitness Test Score Report by overall statistical literacy score.*
Figure 6.6. Categorization of participants based on Proposed Application score report interpretation skill demonstrated on the Jabberwocky Test Score Report by overall statistical literacy score.

Examination of Figures 6.5 and 6.6 shows that Participant E18, who was consistent in their level of Questioning, on Proposed Application shifted from being Transitional in the Cooper’s Fitness Test Score Report to Quantitative when they described what they would do with the information in the Jabberwocky Test Score Report. Three-quarters of the educators (E1, E3, E4, E5, E6, E7, E8, E9, E10, E11, E13, E15, E16, E17, and E19) did not shift their level of Proposed Application between score reports.

Three educators (E11, E16, and E19) were consistently in the category of lower (fewer) Proposed Applications and lower statistical literacy. One educator (E15) was categorized as having lower statistical literacy and higher Proposed Applications. Similar to the comparisons in Figures 6.3 and 6.4, the majority of educators (E3, E4, E5, E7, E13, and E17) did not shift based on the type of report; they had higher statistical literacy and lower Proposed Application skills on score report interpretation. However, in comparison to Questioning there were more educators (E1, E6, E8, E9, and E10) with both higher statistical literacy and higher Proposed Application.
There were five educators whose categorization changed across score reports. One educator (E18) had higher Proposed Application skills on the Jabberwocky Test Score Report than on the Cooper’s Fitness Test Score Report. Four educators (E2, E12, E14, and E20) had higher Proposed Application skills on the Cooper’s Fitness Test Score Report than on the Jabberwocky Test Score Report.

**Comparing Questioning and Proposed Application Patterns**

Across Questioning and Proposed Application, there was one educator (E11) who had lower statistical literacy, did little Questioning, and had few Proposed Applications of the reports. The other consistent group is made up of 4 educators (E4, E5, E7, and E13) who had higher overall statistical literacy scores and did little Questioning and had few Proposed Applications of the reports.

The educators who had lower statistical literacy and did more Questioning were not necessarily the same educators who had higher Proposed Application. Similarly, educators who had higher statistical literacy and did more Questioning did not also have more Proposed Applications. This suggests that Questioning, which may be seen as interpretation, and Proposed Application, which may be seen as proposed use, are two very different elements. Thus, when researchers design validity studies, they need to consider that interpretation may not be related to how an educator would use data.

**Relationship of Data Habit of Mind to Background Information: Research Question 4**

There is an assumption that knowledge of statistics is a moderator for educators to use score reports from large-scale testing. While knowledge of statistics is important, what was discovered in this analysis is that it does not predict use of score reports. Often, educators with higher statistical literacy were not able to transfer that knowledge in questioning or proposing an
application for the score reports. It is possible that the lack of transfer may be in part due to the way that the reports were designed. For example, the Cooper’s Fitness Test Score Report provided opportunities for Summarizing across time points, while the Jabberwocky Test Score Report did not. Those participants who had higher levels of Summarizing on the Cooper’s Fitness Test Score Report than on the Jabberwocky Test Score Report typically focused on the time dimension.

Overall, this sample was made up of very well-educated individuals, all having undergraduate degrees plus an additional degree in education. Five out of the 20 have graduate degrees. What is also curious about this sample is that all educators felt highly comfortable reading statistics and information when presented in articles and journals, which may have a direct relationship to all the educators’ codes being Quantitative or Analytical when Describing and Summarizing score reports.

The background information for the 5 educators (E11, E4, E5, E7, E13) who were consistently categorized in Figures 6.3, 6.4, 6.5 and 6.6 was examined, but there was little commonality in the background information. The only common background element is that these educators are all current teachers, and not administrators, board consultants, or another designation. These educators all have Bachelor of Education degrees and one (E5) has a Master of Education degree. There is little consistency between the background information collected in this study (gender, years of teaching experience, position, undergraduate degree, highest degree earned, favourite topic to teach, experience with assessment, experience with data, comfort with information, math and statistics) and the educators’ Data Habit of Mind categories. However, during the interviews, educators did shed light on their values, which may be useful in understanding the Data Habit of Mind categories.
Educators who tended to have higher statistical literacy and higher Questioning and Proposed Application skills often demonstrated self-reflection during the interview. For example, one participant described the following insights:

Participant E18: Assessment is probably, for me, has always been in my annual learning goals ... my life-long learning goals in my formal years of education. It has been a constant pursuit of mine to understand, advance and implement assessment. We are a very data-driven school and I didn’t have to come to a data-driven school to be hungry for data. It already was in anything that I did …. You keep yourself alive by having data. It just goes hand in hand. We are a very … we are a school that is constantly pushing for data.

This participant was coded high in statistical literacy and score report interpretation, except when proposing an application for the Cooper’s Fitness Test Score Report: In that task, she was coded as Transitional. This participant was an interesting case, as she emphasized that she valued assessment and made it her “annual learning goal.” This educator had only been teaching for 4 years, although teaching is a second career for her; she had spent over 20 years in youth services and working with not-for-profit programs, where they often needed to use data to support decisions.

The major group of consistent educators (E4, E5, E7, and E13) who were higher on statistical literacy and lower when it came to Questioning and Proposed Application for the score reports had a more interesting take on assessment. Generally, when the researcher asked about their experience with assessment and data, the responses were limited to one or two sentences, and they had similar responses when the researcher asked them about their comfort with information and statistics. However, one participant (E13) had a more extended response. He described the plethora of workshops and courses he had taken, but he viewed assessment and evaluation as simply a method of gathering marks:
Participant E1 I’ve done a few workshops on DRA and reading, in math, like effective instruction. I’ve done a lot of reading myself there and we have had a lot of in-service training. That being said, it is an area that is still a little bit hazy because we are going along in our lessons and I don’t necessarily assess them. We do the evaluation side to see what we can get in marks. That’s why I’m interested in the whole literacy thing because we assess the kids in the reading and see where we need to go from there.

This quote eloquently describes how many educators view the use of assessment: “We do the evaluation side to see what we can get in marks.” There has been a great deal of effort in Ontario in the area of assessment for learning, to promote educators’ use of assessment results to support children’s learning. This group of educators, however, views score reports as merely reporting, without further applicability.

Although, in the current sample there were no educators who had lower statistical literacy and higher report interpretation skills, this does not mean they do not exist. Participant E2 was coded as having Transitional statistical literacy, but would switch back and forth between Transitional and Quantitative when it came to Questioning the different score reports and between Transitional and Analytical for Proposed Applications for the different score reports. Unlike Participant E13, described above, this educator drew more on experience than on a set of rules.

When asked about her level of comfort, Participant E2 responded:

Participant E2: Ha ha, very high … Well, I think it’s kind of the package, frankly. I have my reading specialist, I have my religion specialist, which is essential for the Catholic school board. I do have a Master’s degree. I also think the fact that I’ve taught all three divisions helps a lot. Um, I am, in fact, I was an adjunct professor, and you have to mentor the young teachers… I’ve gone anywhere from basal readers to whole language to comprehensive literacy. I’ve kind of seen everything. That goes for math education, as well.

The participant had over 24 years in the education system, first as a teacher and recently as a consultant with the school board. She was able to describe a school’s performance, and what
educators and the school needed to do next to improve their performance. This participant was able to pick out which parts of the report would be relevant to teachers as opposed to principals and school board personnel. She was also much more comfortable with the Jabberwocky Test Score Report than the Cooper’s Fitness Test Score Report, because for her it resembled the local provincial testing report, with which she was familiar.

There was only one participant who stood out in both the interviews and the analysis in this chapter because he was intentionally indifferent about statistics and testing. Throughout the interview, he described how things can be looked at from an assessment point of view and from a non-assessment point of view. He often avoided questions and required numerous probes. This educator was focused on showing his indifference when it comes to statistics and assessment. As he expressed it:

Participant E11: Assessment to me ... it’s not always like paper and a test, a hard assessment that way. Sometimes I think you can get enough of an assessment by watching, observing and talking to the students. Some might not be able to articulate on paper and pencil what they do by talking.

This educator’s focus was on playing down the role of assessment. He was intentionally not being against assessment or for assessment; he also did not see any benefit in being either way. He was much livelier when talking about his favourite subjects to teach. When it came to questions about score report interpretation or statistical literacy, he was doing his best to avoid the questions and to be indifferent in his responses.

The Role of Statistical Literacy and Score Report Interpretation

One might ask, What is the basic level of statistics required to read and interpret score reports? This question is based on the idea that understanding the report requires university-level statistics. The type and presentation of data on these reports, and on several others that were
presented by Goodman and Hambleton (2004), is consistent with the data management skills taught in Ontario’s elementary school curriculum (Ontario Ministry of Education, 2005). Thus, educators and adults who were educated in Ontario should have the basic skills required to read and interpret the statistics on these reports. Although there was no clear conclusion that can be drawn about the background information of educators, the interviews shows that educators’ values, beliefs, ideals and needs affect both the interpretation and use of test results. Thus, if validity is based on the interpretation and intended use of reports, then validation processes must consider that different teachers will have different values that guide their practices.
Chapter 7. Discussion, Limitations and Implications

This study proposes Data Habit of Mind as a way to understand educators’ interpretation and potential use of score reports. Habits of mind, as described by Keating and Sternberg, are more than just mental states or behaviours, but rather patterns in how people apply their knowledge. Many educators rely, when interpreting score reports, not just on their statistical literacy, but also on their intuitions about testing and how students will respond to test results. Data Habit of Mind may help us understand what happens when an educator receives a score report.

Validity and Data Habit of Mind

Understanding what educators plan to do with test results is important for validity theory in educational measurement. Over several decades, the educational measurement community has shifted in its conceptualizations of validity (Cronbach 1971, 1982, 1989; Cronbach & Meehl, 1955; Kane, 1992, 2006; Messick, 1975, 1980, 1989; Moss, 1994; Shepard, 1993, 1997). The most recent attempt at a collective shift in understanding was Educational Researcher’s Volume 36, Issue 8 (2007). In this issue, Lissitz and Samuelsen (2007) argue that validity should be measured by not taking into account the ways in which a test may be interpreted or used, but rather by making sure the test content is being measured; they argue that this is much more useful and “doable” than trying to understand how users would interpret the test. However, other researchers (Embreton, 2007; Gorin, 2007; Mislevy, 2007; Moss, 2007; Sireci, 2007) in the same issue try to convince the reader that there is more to validation than just a measure of how well a test does what it is supposed to do, and provide a convincing argument as to why we need to include how test scores are interpreted and used.
The confusion around validation is sometimes traced to an article Messick wrote in 1975, in which he cites the work of Cronbach on how our interpretations of data and information are not value free. The findings of this study also suggest that educators will interpret results and potentially use them in different ways based on their values and belief systems.

There is a “haziness” in the validation process; validity is supposed to be a match between the use envisioned by the test designer and the actual use. The measurement community is struggling to measure this match or the extent of a mismatch. The most recent and influential approach is to conduct validity studies through the use of argumentation. Kane (1992) suggested that validity ought to be seen as a logical sequence of arguments. More recently, other researchers (Bachman, 2005; Bachman & Palmer, 2010; Mislevy, 2003, 2005, 2007, 2009) have been using adapted versions of Toulmin’s (1958) structured arguments as a framework to build a validity argument. Each researcher has his or her own terminology and there are competing frameworks for how to build a validity argument. However, these systems are still in their infancy and we do not have “best practice” when it comes to validation. Sireci (2007) eloquently reflects on his role as an educational measurement specialist. He writes, “…it is essentially impossible not only to prove that a test is valid but to prove even that we are measuring what we think we are measuring. Therefore, our job in supporting the use of a test for a particular purpose involves presenting sufficient evidence to defend such use” (p. 477).

Conceptualizing the Data Habit of Mind Framework

This thesis has presented the beginning of a framework to understand how educators are interpreting and proposing to use score reports. The conceptual model presented in Figure 2.3 is an initial model, which needs to be further refined through research. However, using this initial model, the researcher was able to investigate educators’ level of statistical literacy, examining
the skills of Describing, Organizing and Reducing, Representing, and Analyzing and Interpreting data. Educators for the most part did not find these tasks difficult. Though the SOLO taxonomy that this model is built on is intended to be sequential (that is, more complex skills would subsume the simpler skills), this was often not the case. Educators showed some difficulty in representing data. However, many educators commented on the fact that they have not drawn a graph since they were in grade school. Thus, the model of statistical literacy may need further adjustments to fully capture the statistical skills educators need when interpreting and using score reports.

The second element of Data Habit of Mind is score report interpretation. All of the educators performed well in Describing and Summarizing the score reports. However, Questioning, which requires educators to reflect on why the data are presented in the way they are and what the report is trying to tell the reader, was more difficult for the participants. While Describing and Summarizing may be thought of as surface level skills, high level Questioning requires an educator to fully comprehend all the elements within a score report.

Proposed Application of the score report is a skill that differs from the other skills in the model. This is the only skill in the Data Habit of Mind model that asks educators, How would you use this information? As shown in the results, if educators are planning on using results, what they would do differs depending on the type of report and their level of statistical literacy. Some educators in this study had a high level of statistical literacy and low levels of Questioning and Proposed Application on both score reports. These are educators who are competent enough, but essentially did not value the information that provided from test score reports. A future research question might be, To what extent can we build a validity argument, if we know that the level of interpretation and use of test score reports is value based?
Limitations

This study has several limitations. Data Habit of Mind has not been previously studied, so in many ways this research is exploratory. In any exploratory research, elements may not fit together quite as well as initially conceived. For example, the study would have been more informative had it included a measure of school culture, as described in Figure 3.1.

The statistical literacy component of the Data Habit of Mind model divided educators into very rough higher and lower categories. The tasks used to measure statistical literacy have been previously used to study children’s and adolescents’ proficiency with statistics, thus application to educators was not a far stretch. However, many educators struggled in representing data, but were much more comfortable in comparing different graphs for similarities and difference, which makes the researcher wonder whether the skills of statistical literacy may need to be refined to capture the full extent of educators’ proficiency with statistics.

The other part of the Data Habit of Mind model is score report interpretation, on which there is limited research. The creation of this part of the model is also exploratory, and it requires furthered refinement. There was a distinct separation when coding educators’ responses between Describing and Summarizing, which all educators did well, and Questioning, on which educators varied. Questioning seems to draw on more varied approaches. Additionally, Proposed Application of score reports does not seem to fit with the other skills of interpretation.

The score reports themselves are novel, though based on a literature review. There is major question whether this is the best method of displaying information. Are there typographies and graphics that are typically used in score reporting? Should the reports have been viewed on digital screens, as this is the trend in communications in general? There are numerous limitations in the design of reports, which are mainly due to a lack of literature on how best to display
information from tests. There are few studies (Jang, 2009; Wolfe, Childs & Elgie, 2004) that integrate how data is displayed and interpreted by educators as part of a validation framework.

There are many more limitations to this study; the two largest being time and funding. The researcher documented the method of coding the transcribed interviews, such that the study maybe modeled in future with refinements, but funds were not available to support an independent coder.

**Furthering Validity Research**

This study is an attempt to better issues affecting test score use validation by interviewing educators on how they interpret and propose to use test scores. The major finding of this study is that interpretation and proposed use of a score report are two different elements of what happens post-testing. All of the educators in this study could describe and summarize a score report. However, many had difficulty questioning the relationship of the score report to their own practice and what they might do with the results. Educators’ statistical literacy is related to how educators can describe and summarize a test score report. Thus, if interpretation is defined as being able to describe and summarize a score report, then we can be confident that if we put a great deal of explanation on a score report, that educators will be able to describe and summarize a report in the ways in which we intended. However, if interpretation is defined as an educator relating the score report to their own classroom, then our validity arguments will fall apart, because educators interpret reports in varied ways.

Kane (2010), building on the work of Cronbach (1971), Messick (1989), and Kane (2006), writes that “Validation involves an evaluation of the credibility, or plausibility, of the proposed interpretations and uses of test scores” (p. 177). Our validity arguments will falter as educators’ proposed use of test score reports differ from their interpretations, and are not
strongly related to their statistical literacy or the ways they can describe and summarize a score report. In many ways, the differences in interpretation and proposed use are more related to educators’ individual values and belief systems than to their statistical literacy.

The Data Habit of Mind conceptual model allows for an investigation of how educators interpret and propose a use for test score reports. The findings from this study add to our knowledge by providing an in-depth look at both interpretation and proposed use and how difficult these are to measure. Essentially, a validity argument, without examining the level of test interpretation or test use, only provides validity evidence for the construct (Messick 1980). Many validation studies become narrowly focused on the construct because understanding how educators may interpret and potentially use the results is too complex. In future research, there is a potential for this Data Habit of Mind model to act as a guide in studying test score report interpretation and use in the validation process.

**Links to Other Practices and Potential for Future Research**

The above results provide four categories of Data Habits of Mind that educators may have developed over time. To complement these focuses, Langlois (2002) describes four methods a physician can make in diagnosing a patient: (1) Exhaustive, (2) Algorithmic, (3) Pattern Recognition, and (4) Hypothetical-deductive. The four methods require the physician to ask the patient a series of questions, run multiple tests and interpret reports. The differences among the four methods are based on the physician’s experience and beliefs about how to diagnose a series of symptoms. The Exhaustive method, as described by Langlois, is the preferred method in medical school, but is rarely practiced in real life situations due to the lack of time available to make diagnoses. The Algorithmic method is practiced more often, and is based on documented symptoms that relate to a specific issue; for example, when a patient has
difficulty lying down, one part of the algorithm suggests diagnosing a leaky valve in the heart. Pattern recognition is a method used by more experienced physicians. Patterns are based on previous cases they have come across, for example, when a new patent shows the same signs as a previous patent the doctor has had. Langlois (2002) wrote that the method most often used is the Hypothetical-deductive, where “clues and hunches are used in a systematic way to guide a focused inquiry and the development of a rank order list of hypotheses” (p. 198). In medicine, the consequences of accurate or inaccurate diagnosis are extremely high. In education, consequences also exist in understanding what students are struggling with; however, the diagnostic process is not clear cut.

There may be links between Langlois’s medical diagnosis methods and the four Data Habit of Mind categories presented in this study. An educator how may be proficient at statistical literacy and have high levels of Questioning and Proposed Application may have an Exhaustive or Algorithmic practice when it comes to diagnosing a student’s progress. Educators who have these tendencies may exhaust a score report in relation to their own collection of students’ work or may look for patterns or even have a hypothesis about what a child’s performance means. They might examine students’ score reports in relation to students’ performance in class, leading them to question the results and reports. Other educators who were highly statistical but did not value the score reports may also have practices that are Algorithmic, use Pattern Recognition or have a Hypothetical-deductive approach, but they would do this only by looking at the reports, not always taking into account class performance. Thus, their level of questioning would be more basic, and may not be tied to how they may apply the information they have gained. The same might apply to educators who had higher skills when questioning and proposing a use for the score reports. However, educators who seem to be indifferent, putting little value on statistics or
test score reports, may still have predictions about their students, although these are not backed by methods of collecting and organizing information. Confirming these possible parallels between Langlois’s medical diagnosis methods and the four Data Habit of Mind categories will require additional research.

**Future Directions**

Rosenthal (2006) wrote *Struck by Lightning*, a national best-selling book on applications of probability to everyday situations. In the book, Rosenthal describes how people often try to draw connections to data based on what they have experienced. While sometimes this is useful, sometimes the connections made are just by chance. For example, when students’ performance on a school-based test is similar to their results on a large-scale test, educators need to dig deeper to ensure that the connection is not by chance, but rather that children are performing well on complementary expectations and skills.

In this study, educators were asked some basic questions about statistics, at the level of the Grade 7 and 8 curriculum in Ontario. The tasks they were asked to do were more complex than the types of statistical information presented on score reports. While almost all participants performed well on the statistical literacy tasks, many struggled when the statistics related to student performance. This may suggest that, as part of pre-service training, educators need to have some experience in working with achievement statistics.

Alison Gopnik (2010) has discovered that babies, even at 20 months, can begin to recognize similar pattern distributions. From her research, we may begin to believe that people in general have some predisposition to make connections between similar patterns. Thus, we generally see similar things, but it is only when we begin to develop the idea that we may be thinking something differently than someone else that our interpretations become uniquely ours.
There are two questions on which we should focus in the future: (1) How can we support educator training to make accurate interpretations of data? (2) How can measurement specialists design score reports to facilitate the interpretation of scores in the ways intended?
References


Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies.


### Appendix A
#### Standards Literature Analysis

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Suggestions on how to interpret results while taking into account test properties</td>
<td>Standard 13.14. In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores.</td>
<td>C-1. Provide information to support recommended interpretations of the results, including the nature of the content, norms or comparison groups, and other technical evidence. Advise test users of the benefits and limitations of test results and their interpretation. Warn against assigning greater precision than is warranted.</td>
<td>6.2. Provide to those who receive assessment results information about the assessment, its purposes, its limitations, and its uses necessary for the proper interpretation of the results.</td>
<td>H. Explain to users how reports may be useful to them, why and how the evaluation was conducted, and how the data support the interpretations and recommendations.</td>
</tr>
<tr>
<td>(2) Interpretation guides</td>
<td>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 scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.</td>
<td>C-2. Provide guidance regarding the interpretations of results for tests administered with modifications. Inform test users of potential problems in interpreting test results when tests or test administration procedures are modified.</td>
<td>6.3. Provide to those who receive score reports an understandable written description of all reported scores, including proper interpretations and likely misinterpretations.</td>
<td>C. Indicate to users how the report may be useful to them, how the evaluation was conducted, how the data support the interpretations and recommendations (See A5. Defensible Information), and how the findings may be used to further the educational development of the students (see U7. Follow-UP).</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>----------------------------------</td>
<td></td>
</tr>
<tr>
<td>(3) Appropriate uses and non-appropriate uses of the scores and the report</td>
<td>Standard 13.1. When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user. Standard 13.13. Those responsible for educational testing programs should ensure that the individuals who interpret the test results to make decisions within the school context are qualified to do so or are assisted by and consult with persons who are so qualified.</td>
<td>C-3. Specify appropriate uses of test results and warn test users of potential misuses. 6.8. Avoid making, and actively discourage others from making, inaccurate reports, unsubstantiated claims, inappropriate interpretations, or otherwise false and misleading statements about assessment results. 6.10. Report any apparent misuses of assessment information to those responsible for the assessment process.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>(4) How standards and/or norms were generated</td>
<td>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 scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.</td>
<td>C-4. When test developers set standards, provide the rationale, procedures, and evidence for setting performance standards or passing scores. Avoid using stigmatizing labels.</td>
<td>6.5. Evaluate and communicate the adequacy and appropriateness of any norms or standards used in the interpretation of assessment results.</td>
<td></td>
</tr>
<tr>
<td>(5) Multiple measures to report on overall students’ progress</td>
<td>Standard 13.7. In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision.</td>
<td>C-5. Encourage test users to base decisions about test takers on multiple sources of appropriate information, not on a single test score.</td>
<td>6.7. Use multiple sources and types of relevant information about persons or programs whenever possible in making educational decisions.</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>-------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>(6) Information to allow for group comparisons</td>
<td>Standard 5.12. When group-level information is obtained by aggregating the results of partial tests taken by individuals (e.g., as is the case with matrix sampling), validity and reliability should be reported for the level of aggregation at which results are reported. Scores should not be reported for individuals unless the validity, comparability, and reliability of such scores have been established.</td>
<td>C-6. Provide information to enable test users to accurately interpret and report test results for groups of test takers, including information about who were and who were not included in the different groups being compared, and information about factors that might influence the interpretation of results.</td>
<td>C-7. Provide test results in a timely fashion and in a manner that is understood by the test taker.</td>
<td>C-8. Provide guidance to test users about how to monitor the extent to which the test is fulfilling its intended purposes.</td>
</tr>
<tr>
<td>(7) A rapid return time</td>
<td>Standard 11.6. Unless the circumstances clearly require that the results be withheld, the test user is obligated to provide a timely report of the results that is understandable to the test taker and others entitled to receive this information</td>
<td>C-7. Provide test results in a timely fashion and in a manner that is understood by the test taker.</td>
<td>6.4. Communicate to appropriate audiences the results of the assessment in an understandable and timely manner, including proper interpretations and likely misinterpretations.</td>
<td>G. Produce timely reports, so that information provided will be of maximum use.</td>
</tr>
<tr>
<td>(8) Methods for monitoring progress and change</td>
<td></td>
<td></td>
<td>B4. Allow enough time to prepare and check the completeness and accuracy of formal reports before they are shared with users.</td>
<td>I. Obtain feedback from users on the usefulness, clarity, timeliness, and accuracy of reports they have received (see A11, Metaevaluation).</td>
</tr>
<tr>
<td>-------------------------</td>
<td>---------------------------------------------------------------</td>
<td>-----------------------------------------------------------------</td>
<td>-----------------------------------------------------------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>(9) Information on students’ performance for parents and guardians</td>
<td>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 scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.</td>
<td>6.6. Inform parties involved in the assessment process how assessment results may affect them.</td>
<td></td>
<td>B6. Provide for conferences between teachers and parents. Whenever it is appropriate, students should participate in the conferences.</td>
</tr>
<tr>
<td>(10) Confidence that information about a student is confidential</td>
<td>Standard 5.15. When test data about a person are retained, both the test protocol and any written report should also be reserved in some form. Test users should adhere to the policies and record-keeping practice of their professional organizations. Standard 5.13. Transmission of individually identified test scores to authorized individuals or institutions should be done in a manner that protects the confidential nature of the scores.</td>
<td>6.9. Disclose to examinees and others whether and how long the results of the assessment will be kept on file, procedures for appeal and rescoring, rights examinees and others have to the assessment information, and how those rights may be exercised.</td>
<td>6.11. Protect the rights to privacy of individuals and institutions involved in the assessment process.</td>
<td>D. Determine if the reports should be retained in the student's file for future use (e.g., to assist in measuring growth over time) and whether the student's report in whole or in part should contribute to other decision situation (e.g., at the school or district level). E. Maintain confidentiality in reporting (see P3, Access to Evaluation Information).</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>---------------------------------------</td>
</tr>
<tr>
<td>(11) Multiple methods of reporting</td>
<td></td>
<td></td>
<td></td>
<td>F. Use a variety of approaches to reporting (e.g., report cards with a portfolio and a student-led conference.)</td>
</tr>
<tr>
<td>(12) Guidance in understanding the consequential use of scores</td>
<td>Standard 5.16. Organizations that maintain test scores on individuals in data files or in an individual’s records should develop a clear set of policy guidelines on the duration of retention of an individual’s records, and on the availability, and use over time, of such data.</td>
<td>6.1. Conduct these activities in an informed, objective, and fair manner within the context of the assessment’s limitations and with an understanding of the potential consequences of use.</td>
<td>A. Prepare a written policy to guide the reporting system for a school or jurisdiction.</td>
<td></td>
</tr>
<tr>
<td>(13) Methods to facilitate learning</td>
<td></td>
<td></td>
<td></td>
<td>B2. Provide clear, concrete, and accurate feedback, including specific strengths and weaknesses with suggestions for improvement in problem areas (see U7, Follow-Up).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B3. Avoid the use of unfamiliar terms and jargon. Explain what letter grades and other symbols mean and how they should be interpreted and used.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>B5. Design reports to enhance clarity, readability, and understandability.</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------------------------------------------</td>
<td>-------------------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>(14) Explanatory information on the reports</td>
<td></td>
<td></td>
<td></td>
<td>B.1. Include a description of the goals and objectives of instruction to which the evaluation is referenced.</td>
</tr>
</tbody>
</table>
### Appendix B
Guidelines from Measurement Research

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Relevant information</td>
<td>1. Include all information essential to proper interpretation of assessment results in student score reports (e.g., statements explaining the purpose of the assessment, the meaning of performance levels and test scores, and how the test results should be used, and examples of how to interpret confidence bands). Consider creating larger reports that can accommodate this information (a pamphlet design created by folding an 11&quot; × 17&quot; sheet in half appears particularly promising).</td>
<td>Report Audience: Reports are prepared for various audiences and what is contained in the report and how the information is presented generally varies depending on the audience and users of the reports.</td>
</tr>
<tr>
<td>(2) How standards and/or norms were generated</td>
<td>2. Include detailed information about the assessment and score results in a separate interpretive guide, ideally one in which the student score report can be inserted.</td>
<td>Scale or metric for reporting: The scale or scales in which scores are reported can add clarity or confusion to the score report. It is often simpler to report raw scores or percent scores, but these scales do not provide comparability across strands on a single test or between two different tests.</td>
</tr>
<tr>
<td>(3) Customized information based on user or group</td>
<td>3. Personalize the student score reports and interpretive guides.</td>
<td>Reporting Unit: Score reports are routinely provided for individual students and for different aggregations of students from classrooms to the entire nation. Certain features of all reports are the same, but each level of reporting requires different information approaches.</td>
</tr>
<tr>
<td>(4) Interpretation Guides</td>
<td>4. Include an easy-to-read narrative summary of the student’s results at the beginning of the student score report.</td>
<td>Reference for interpretation: Test results can be interpreted in reference to some normative information such as percentiles or by reporting how students in the school, district and state perform on the test. In most states, test scores are reported in terms of content and/or performance standards. Reporting students’ test scores in terms of performance achievement levels is becoming a popular approach.</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>------------------------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>(5) Information on students’ performance for parents and guardians</td>
<td>5. Identify some things parents can do to help their child improve. Ideally, these suggestions would be included in a separate section near the end of the score report and would be tailored to the student’s performance. Advise parents and guardians to talk with their child’s teacher about other ways to improve performance.</td>
<td>Assessment Unit: Educators are interested in instructionally useful information about how students perform on individual items or subsets of items such as content standards. Strand-level information is commonly reported but has technical limitations</td>
</tr>
<tr>
<td>(6) Methods to facilitate learning</td>
<td>6. Include sample questions in the interpretive guides that illustrate the types of achievement represented by each performance level.</td>
<td></td>
</tr>
<tr>
<td>(7) Explanatory information on the reports</td>
<td>7. Include a reproduction of student score reports in the interpretive guides to clearly explain the various elements of the reports.</td>
<td></td>
</tr>
<tr>
<td>(8) Precision of measurement at the subscale level</td>
<td>Error of Measurement: The precision with which test scores are measured is often reported for performance at the total test level. However, the standard error of measurement is not always presented when strand-level achievement is reported. Charts and tables that invite comparing students’ achievement across strands should report the standard errors of the differences.</td>
<td></td>
</tr>
<tr>
<td>(9) Multiple methods of reporting</td>
<td>Mode of Presentation: test results can be presented numerically, graphically or in descriptive narrative form. The best approach for different audiences is not always the same and use of multiple modes of presentation with some built in redundancy is often seen in score reports.</td>
<td></td>
</tr>
<tr>
<td>(10) Accessibility to users on multiple platforms</td>
<td>Reporting Medium: Test results have been traditionally presented in printed hard copy form. This practice will likely continue for some time but electronic versions supplied via Internet or on CDs are quite common.</td>
<td></td>
</tr>
</tbody>
</table>
Appendix C
Letter of Consent

Investigating Data habit of mind

What the study is about: The purpose of this study is to improve reports from large-scale assessments by learning about how educators understand and use these reports.

What we will ask you to do: If you agree to be a part of the study, you will take part in an hour-long interview. The interview questions focus on elements of assessment reports. The questions are general enough that you will be able to answer from memory, without needing to refer to any materials.

Risks and benefits: We do not anticipate any risk or benefits for your participation in this study. We hope that what we learn in the study will help us enhance reporting to benefit future students and teachers.

Taking part is voluntary: Taking part in this study is completely voluntary. If you decide to take part, you are free to withdraw at any time. If you decide to withdraw, the data collected from you up to the time you withdraw will be destroyed.

Your answers will be confidential: The records of this study will be kept private. In any report about the study we will not include any information that could be used to identify you. Research records will be kept in a locked file; only the researchers will have access to the records.

For more information Contact Saad at schahine@oise.utoronto.ca or Professor Ruth Childs at ruth.childs@utoronto.ca

Name_____________________
Date: ____________
Signature: ____________________
Do you find test results enlightening or confusing?

If you are an elementary teacher or have been within the last three years and are willing to spend an hour talking about how test results can be more useful for teachers, please contact us.

Saad at schahine@oise.utoronto.ca or Professor Ruth Childs at ruth.childs@utoronto.ca

Participants receive a $35 gift card.
Appendix D
Interview Protocol

Interviewer: Welcome. <Introduction>
Interviewer: We are interested in understanding your thinking, so please don’t worry about getting the right answer.
Interviewer: You don’t have to complete the question, just start and I will stop you and ask you what you are thinking.
Interviewer: This is not a test to see how well you do, but to understand your thought process to solve these questions.
Part 1. Statistical Literacy

Olympic Medal Task

The graph shows the medals won at the 1998 Winter Olympics for five countries. The graph is not finished. The table shows the medals won by Italy and China.

![Graph showing medals won at the 1998 Winter Olympics for United States, Netherlands, Finland, Italy, and China.]

<table>
<thead>
<tr>
<th>Country</th>
<th>Gold</th>
<th>Silver</th>
<th>Bronze</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>China</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

Source of data: CB5 Sportline (http://www.sportsline.com).

Interviewer: The first one is called the Olympic medal task, the graph shows the medals won at the 1998 Olympics for 5 countries the graph is not finished the table shows the medals won by Italy and China. So can you use the information in the tables to complete the graph?

Interviewer: Explain how you completed the graph.

Interviewer: How do the countries compare in medals won?

Interviewer: Which country won the most gold medals?

Interviewer: How can you tell?

Interviewer: And which country won the fewest silver?

Interviewer: How can you tell?

Interviewer: What would you say is the typical number of gold medals won by the 5 countries?

Interviewer: Can you tell me what your thinking was?
Movie Star Salaries

Here is a table showing the salaries of 15 top actors and actresses.

<table>
<thead>
<tr>
<th>Actors</th>
<th>Actresses</th>
</tr>
</thead>
<tbody>
<tr>
<td>$17.5</td>
<td>$12.5</td>
</tr>
<tr>
<td>15.0</td>
<td>9.0</td>
</tr>
<tr>
<td>20.0</td>
<td>11.0</td>
</tr>
<tr>
<td>20.0</td>
<td>9.5</td>
</tr>
<tr>
<td>20.0</td>
<td>2.5</td>
</tr>
<tr>
<td>19.0</td>
<td>12.0</td>
</tr>
<tr>
<td>20.0</td>
<td>3.0</td>
</tr>
<tr>
<td>18.0</td>
<td>4.0</td>
</tr>
<tr>
<td>5.5</td>
<td>4.0</td>
</tr>
<tr>
<td>6.0</td>
<td>2.5</td>
</tr>
<tr>
<td>10.0</td>
<td>6.0</td>
</tr>
<tr>
<td>16.5</td>
<td>8.5</td>
</tr>
<tr>
<td>12.5</td>
<td>4.5</td>
</tr>
<tr>
<td>10.0</td>
<td>3.0</td>
</tr>
<tr>
<td>7.0</td>
<td>10.0</td>
</tr>
</tbody>
</table>

AII:
How do the actors’ salaries compare to the actresses’ salaries?

DD4/DD1:
What is the highest salary for actors?

OR3:
What is the range of salaries for actresses?
OR1:
Can you organize the data in a different way? Explain what you did.

AI1:
What information about actors’ and actresses’ salaries can you get from the organized information?

OR2:
What is the typical salary for the actresses? How did you determine the typical salary?

RD1:
Can you make a graph that will allow you to compare the salaries of actors and actresses? Explain what you did.

AI1:
What information does the graph tell you?

Dog Breeds

The table shows the number of registered dog breeds in the American Kennel Club in 1996 for four breeds of dogs. A pictograph was started to represent the information in the table.

<table>
<thead>
<tr>
<th>Number of Registered Dogs in the American Kennel Club in 1996</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Breed</strong></td>
</tr>
<tr>
<td>Beagle</td>
</tr>
<tr>
<td>Dalmatian</td>
</tr>
<tr>
<td>German Shepherd</td>
</tr>
<tr>
<td>Rottweiler</td>
</tr>
</tbody>
</table>


Number of Registered Dogs in the American Kennel Club in 1996
OR2:
What is the typical number of pets sold each day? How can you tell?

OR3:
What is the range of pets sold? How can you tell?

State Population and Area Codes
This graph shows the population and the number of area codes for each state in the United States.


DD4/DD1:
How many states have 5 area codes? How did you decide this?

DD4/DD1:
What information does the point labeled A represent? How did you decide this?

AI3:
How does the population of a state compare to the number of area codes the state has? How did you decide this?
Part 2. Educational History and Comfort Level

Interviewer: Can you please tell me a little about your history in the education field? Probes:
- Undergraduate degree
- Years of experience
- Topics taught (What’s your favourite topic to teach?)
- Experience in assessment
- Experience with data
- Additional Qualification courses
- Grade teaching now

Interviewer: On a scale of 1 to 10 (with 10 being the highest), can you please describe your comfort level with the following?
- Newspaper
- Newspaper with data/numbers in the report
- School-based report (such as a reading diagnostic)
- Graphs
- Tables with numbers
- Data integrated with text (when there are numbers in the text)
- Mathematics
- Statistics
Part 3. Score Report Interpretation

Cooper’s Fitness Test Score Report

Interviewer: Take 5 minutes to read the report and please tell me when you are ready to be able to explain the report to me.

Interviewer: Please explain this report to me

Probe: Is there anything else you would like to add?

Probe: Is there anything else you would like to add?

Interviewer: I’d like you to focus on where it says maximum speed and average speed – what does that mean to you?

Probe: Is there anything else you would like to add?

Probe: Let’s pretend you’re Robin’s teacher/coach. What would that tell you?

Interviewer: I’d like you to focus on that table that shows time, distance and accumulation – what does that mean to you?

Probe: Is there anything else you would like to add?

Probe: Let’s pretend you’re Robin’s teacher/coach. What would that tell you?

Interviewer: I’d like you to pay attention to the three bullet points at the bottom, for you as a coach or Robin. Are these important to you? Please explain.

Interviewer: I would like you to look at those bars of distributions. What do you think those represent?

Probe: What do you think of Robin’s bar?

Interviewer: When you first saw the report, what was the first thing that you looked at?

Interviewer: What was the last thing that you looked at?

Interviewer: What concepts or ideas were meaningful to you when reading the report?

Interviewer: What concepts or ideas were least meaningful to you when reading the report?

Interviewer: What additional information do you need to help support Robin’s endurance?
Cooper Test of Endurance
Individual Student Summary
November 13, 2013

<table>
<thead>
<tr>
<th>Time (min)</th>
<th>Distance (m)</th>
<th>Accumulation (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>337</td>
<td>337</td>
</tr>
<tr>
<td>2</td>
<td>322</td>
<td>659</td>
</tr>
<tr>
<td>3</td>
<td>207</td>
<td>866</td>
</tr>
<tr>
<td>4</td>
<td>228</td>
<td>1094</td>
</tr>
<tr>
<td>5</td>
<td>163</td>
<td>1257</td>
</tr>
<tr>
<td>6</td>
<td>139</td>
<td>1396</td>
</tr>
<tr>
<td>7</td>
<td>124</td>
<td>1520</td>
</tr>
<tr>
<td>8</td>
<td>218</td>
<td>1738</td>
</tr>
<tr>
<td>9</td>
<td>203</td>
<td>1941</td>
</tr>
<tr>
<td>10</td>
<td>116</td>
<td>2057</td>
</tr>
<tr>
<td>11</td>
<td>89</td>
<td>2146</td>
</tr>
<tr>
<td>12</td>
<td>103</td>
<td>2249</td>
</tr>
</tbody>
</table>

FINAL DISTANCE 2249

Average Speed: 187m/min
Maximum Speed: 337m/min

Ways to Improve

- Review your Time/Distance chart. You ran too fast to begin with and expended all of your energy. We suggest that you ask your coach for an aerobic program that will allow you to achieve to your maximum ability.
- Review your Range. It is very large. This suggests that you are not expending an equal amount of energy while you are running. We suggest that you ask your teacher to help you run at a uniform pace.
- Review the standard for a student your age. You are in the low average range. We suggest that you discuss with your coach how you can improve your fitness.

Congratulations, Robin, you are of **Average** fitness

2249
Coopers Test of Endurance
Comparative Summary
November 13, 2013

Robin 2248m

Please note: Scores in the graph are based on all students and outliers have been removed.

About the Cooper Test of Endurance

Kenneth H. Cooper, developed the 12 minute run in 1968 for the US Army. The test was developed on the evidence of a strong correlation between the distance a person can run and their ability to use oxygen efficiently. The Coopers test of endurance is an estimate of VO₂ MAX. For more information on health and endurance please visit the Statistics Canada Health website:

http://www4.statcan.gc.ca/health-sante/index-eng.htm

Jabberwocky Score Report
Interviewer: Take 5 minutes to read the report and please tell me when you are ready to be able to explain the report to me.
Interviewer: Please explain this report to me
   Probe: Is there anything else you would like to add?
   Probe: Is there anything else you would like to add?

Interviewer: I’d like you to focus on where it says “your assessment results.” What does that mean to you?
   Probe: Is there anything else you would like to add?
   Probe: Let’s pretend you’re P. Smith’s trainer. What would that tell you?

Interviewer: I’d like you to focus on where it says “understanding results.” What does that mean to you?
   Probe: Is there anything else you would like to add?
   Probe: Let’s pretend you’re P. Smith’s trainer. What would that tell you?

Interviewer: I’d like you to pay attention to where it says “Summary of Region.” For you as a trainer, are these important to you?

Interviewer: I would like you to look at those circles around the dots. What do these mean to you?

Interviewer: When you first saw the report, what was the first thing that you looked at?

Interviewer: What was the last thing that you looked at?

Interviewer: What concepts or ideas were meaningful to you when reading the report?

Interviewer: What concepts or ideas were least meaningful to you when reading the report?

Interviewer: What additional information do you need to help support this student’s learning?
Individual Student Report

Assessment of Jabberwockyness
Junior Division

STUDENT INFORMATION

Student name: P. Smith
Student Number: 987-654-321

YOUR RESULTS ON THE ASSESSMENT OF JABBERWOCKYNESS

<table>
<thead>
<tr>
<th></th>
<th>Level 0</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th>Level 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not enough Evidence</td>
<td>Extremely Below Standard</td>
<td>Below Standard</td>
<td>Approaching Standard</td>
<td>At standard</td>
<td>Above Standard</td>
</tr>
</tbody>
</table>

Translating
answered 45 of 45

Combining
answered 35 of 45

● symbol represents the student's level on this assessment.

*Answered questions* is the total number of questions answered on each section

*Jabberwockyness* is a term used to describe a person’s level of nonsensicalness. The term Jabberwocky was first used in a poem written by Lewis Carroll in 1872, in which he blends together nonsensical terms to describe how a boy slays a beast.

Understanding your results:

This report indicates that you have met the standard of Jabberwockyness. The standard has been determined by the Jabberwocky society of Canada. This society has evidence that you are able to create and combine nonsensical words in a way that gives an illusion of sense. You are able to adapt the English language to bring whimsy and intrigue without over-confusing the reader.
Summary of Regional, Provincial and Canadian Results

Results of your region, province and country are provided to assist you in understanding assessment results.

Translating

Combining

Percentages in the above graphs are based on all students of the Junior level. Percentages may not add up to 100% due to rounding.

For more information please contact the association at jb@jab.com
Final Part

Interviewer lays out both reports on the table and asks:
1. Which report are you most comfortable with? Please explain.
2. If you were to design your own report, what elements would you combine from these reports?
3. Do you have an additional comments or questions for me?
## Appendix E

**Relationship between Research Questions, Questions in the Interview, and the Conceptual Framework**

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Interview Task</th>
<th>Questions</th>
<th>Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are educators’ levels of statistical literacy?</td>
<td>Olympic Medal Task</td>
<td>Can you use the information in the table to complete the graph?</td>
<td>1-4 Representing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How do the countries compare in the medals won?</td>
<td>1-4 Analyze/Interpret</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Which country won the most gold medals? How can you tell?</td>
<td>1-4 Describing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is the typical number of gold medals won by the five countries? How did you determine the typical number?</td>
<td>1-4 Organizing</td>
</tr>
<tr>
<td>Movie Star Salaries</td>
<td>How do the actors’ salaries compare to the actresses’ salaries?</td>
<td>1-4 Analyze/Interpret</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>What information about actors’ and actresses’ salaries can you get from the organized information?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>What information does the graph tell you?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is the highest salary for actors?</td>
<td>1-4 Describing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is the range of salaries for actresses?</td>
<td>1-4 Organizing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Can you organize the data in a different way?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is the typical salary for the actresses? How did you determine the typical salary?</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Can you make a graph that will allow you to compare the salaries of actors and actresses? Explain what you did.</td>
<td>1-4 Representing</td>
</tr>
<tr>
<td>Dog Breeds</td>
<td>Can you use the information in the table to complete the graph? Explain what you did.</td>
<td>1-4 Representing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Can you make a different graph to represent the information in the table?</td>
<td>1-4 Representing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>How does your graph compare to the pictograph? How are the graphs alike? How are they different?</td>
<td>1-4 Analyze/Interpret</td>
<td></td>
</tr>
<tr>
<td><strong>State Population and Codes</strong></td>
<td><strong>How many states have 5 area codes? How did you decide this?</strong></td>
<td>1-4 Describing</td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>What information does the point labelled A represent? How did you decide this?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>How does the population of a state compare to the number of area codes the state has? How did you decide this?</td>
<td>1-4 Analyze/Interpret</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Suppose the United States decides to make Canada a state. How many area codes would Canada have if the population of Canada is 29,100,000? How did you decide this?</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>How do educators read and interpret reports</strong></td>
<td><strong>How does the population of a state compare to the number of area codes the state has? How did you decide this?</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cooper’s Fitness Test Report</strong></td>
<td>Take 5 minutes to read and examine the report and please tell me when you’re ready to explain the report to me.</td>
<td>1-4 Describing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Please explain the report to me.</td>
<td>1-4 Questioning</td>
<td></td>
</tr>
<tr>
<td></td>
<td>I’d like you to focus on where it says maximum speed and average speed. What does this mean to you?</td>
<td>1-4 Summarizing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>I would like you to focus on the table that shows time, distance and accumulation – what does this mean to you?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Let’s pretend you’re Robin’s coach. What does this section tell you?</td>
<td>1-4 Proposed Application</td>
<td></td>
</tr>
<tr>
<td></td>
<td>I would like you to pay attention to the three bullet points at the bottom. For you as a coach, are these important?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>What additional information do you need in this report to help support students’ learning?</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>I’d like you to pay attention to those bars of distribution. What do you think those represent?</td>
<td>1-4 Questioning</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Jabberwocky Score Report</strong></th>
<th><strong>Take 5 minutes to read and examine the report and please tell me when you’re ready to explain the report to me.</strong></th>
<th>1-4 Describing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Please explain the report to me.</td>
<td>1-4 Questioning</td>
</tr>
<tr>
<td></td>
<td>I’d like you to focus on where it says “your assessment results.” What does this mean to you?</td>
<td>1-4 Summarizing</td>
</tr>
<tr>
<td></td>
<td>I would like you to focus on where it says “understanding results.” What does this mean to you?</td>
<td></td>
</tr>
</tbody>
</table>
Let’s pretend you’re P. Smith’s trainer. What does this section tell you?

I’d like you to pay attention to the “Summary of Region.” For you as a coach, are these important to you?

What additional information do you need in this report to help support students’ learning?

I’d like you to pay attention to those circles around the dots. What do you think those represent?

<table>
<thead>
<tr>
<th>How can an educator’s history and confidence contribute to Data Habit of Mind?</th>
<th>Can you please tell me a little about your history in the education field?</th>
<th>Descriptively coded for:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>• Years of Experience</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Gender</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Location</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Undergraduate Degree</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Favourite Topic/Subject</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Assessment</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Additional Qualifications</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Current teaching grade</td>
</tr>
</tbody>
</table>

On a scale of 1 to 10, can you please describe your comfort level with …?

Descriptively coded for:

- Information (text/newspaper)
- Mathematics
- Statistics
Appendix E
Letter of Approval from Dr. Mooney

SAAD CHAHINE <saad.chahine@gmail.com>

inquiry about the statistical thinking framework
5 messages

SAAD CHAHINE <saad.chahine@gmail.com>                           Tue, Dec 15, 2009 at 4:01 PM
To: mooney@ilstu.edu

Hi Dr. Mooney,

I am a doctoral student at the OISE/University of Toronto. My dissertation examines the ways in which educators understand data from large scale assessment reports.

I would like to use the framework that you have in your 2002 paper “A Framework for Characterizing Middle School Students’ Statistical Thinking” as well as the items types and codes you used.

I’m wondering if this would be okay with you, as well if you have any more research that you have conducted since that you think maybe useful.

thank you for your time,

Saad Chahine

PhD Candidate
Human Development and Applied Psychology,
OISE/UT

Edward Mooney <mooney@ilstu.edu>                                     Thu, Dec 17, 2009 at 1:03 PM
To: SAAD CHAHINE <saad.chahine@gmail.com>

Saad,

Thank you for the interest in my work. You are welcome to use the coding and tasks if you like. I would suggest looking at Watson (2006) for a more comprehensive framework. She also has several good tasks that cover a wide range of statistical content (Watson, Kelly, Callingham & Shaughnessy, 2003; Watson & Callingham 2003). I don’t think there is any current research that will relate. I, along with other colleagues, have been looking at the role of context on students’ statistical thinking. Here are a few references to those.


