Examining thinking skills in the context of large-scale assessments using a validation approach

Krystal Hachey

Thesis submitted to the

Faculty of Graduate and Postdoctoral Studies

in partial fulfillment of the requirements

for the Doctorate in Philosophy degree in Teaching, Learning & Evaluation,

Faculty of Education

University of Ottawa

Table of Contents

List of Tables

Legend

Name	Acronym	First Reference
American Educational Research Association	AERA	2
American Psychological Association	APA	2
The Council of Ministers of Education, Canada	CMEC	1
Classical Test Theory	CTT	33
Differential Item Functioning	DIF	33
Education Quality and Accountability Office	EQAO	1
Exploratory Analysis	EX	74
Goodness-of-Fit	GFI	74
Generation of Starting Values	IV	74
Higher Order Thinking	HOT	2
International Association for the Evaluation of Educational Achievement	IEA	1
Item Response Theory	IRT	34
Large Scale Assessment	LSA	1
Number of Dimensions	M	74
Medical College Admission Test	MCAT	4
Number of Items	N	74
National Assessment of Educational Progress	NAEP	28
National Council in Measurement and Evaluation	NCME	2
No Child Left Behind	NCLB	3
Normal Ogive Harmonic Analysis Robust Method	NOHARM	34

Name	Acronym	First Reference
The Number of Subjects or Observations	NS	74
Organization for Economic Co-Operation and Development	OECD	1
Pan Canadian Assessment Program	PCAP	1
Principle Components Analysis	PCA	72
Productive Thinking	PT	2
Programme for International Student Assessment	PISA	1
Progress in International Reading Literacy Study	PIRLS	1
Printing/no printing of residual matrix	PR	74
Printing/no printing of product-moment matrix	PS	74
Root Mean Square	RMS	74
Trends in International Mathematics and Science Study	TIMSS	1
Type of Data	IN	74

Abstract

Large Scale Assessments (LSAs) of student achievement in education serve a variety of purposes, such as comparing educational programs, providing accountability measures, and assessing achievement on a broad range of curriculum standards. In addition to measuring content-related processes such as mathematics or reading, LSAs also focus on thinking-related skills such as lower level thinking (e.g., understanding concepts) and problem solving. The purpose of the current study was to deconstruct and clarify the mechanisms that make up an LSA, including thinking skills and assessment perspectives, from a validation approach based on the work by Messick (1995) and Kane (1990). Therefore, when examining the design and student data of two LSAs in reading, (a) what common thinking skills are assessed? and (b) what are the LSAs' underlying assessment perspectives? Content analyses were carried out on two LSAs that purported to assess thinking skills in reading: the Pan-Canadian Assessment Program (PCAP) and the Educational Quality and Accountability Office (EQAO). As the two LSAs evaluated reading, the link between reading and thinking was also addressed. Conceptual models were developed and used to examine the assessment framework, test booklets, and scoring guide of the two assessments. In addition, a nonlinear factor analysis was conducted on the EQAO item-level data from the test booklets to examine the dimensionality of the LSA. The most prominent thinking skill referenced after qualitatively analyzing the assessment frameworks, test booklets, and scoring guides was critical thinking, while results from the quantitative analysis revealed that two factors best represented the item-level EQAO data. Overall, the tools provided in the current study can help inform both researchers and practitioners about the interaction between the assessment approach and related thinking skills.

Acknowledgments

I would first like to thank my supervisor, Dr. Marielle Simon, for her support, wisdom, and continued guidance throughout my Ph.D. program. I would also like to thank my committee members, Dr. David Trumpower, Dr. Christine Surrtamm, and Dr. Alain Desrocher, and my external committee member, Dr. Louis Volante, for their vital feedback and perspectives in the area of quantitative and qualitative analysis, large scale assessment and reading. In addition, I would like to thank the Educational Quality and Accountability Office (EQAO) and the Pan-Canadian Assessment Program (PCAP) for providing the data used in this research. Without access to the data, this study would not have been possible. It is important to note that the opinions presented in this paper are solely those of the author and do not necessarily reflect the opinions of EQAO and PCAP.

I would also like to thank my family and friends for their continued support with my endeavour to pursue a PhD. Most importantly, I would like to thank my husband, Adam Arsenault, for his monumental support, patience, encouragement, insight, and optimism.

The Dedication

To Gloria Arsenault, for showing true strength in the face of adversity.

Chapter One: Introduction

Large Scale Assessments (LSAs) of student achievement in education serve a variety of purposes, such as comparing educational programs, providing accountability measures, and assessing achievement on a broad range of curriculum standards (The Council of Ministers of Education, Canada [CMEC], 2008; Taylor & Tubianosa, 2001). They have become a part of the educational system because they enable the documentation of student achievement through provincial, national, and international means (CMEC, 2008; Taylor & Tubianosa, 2001). Examples of key international LSAs include the Programme for International Student Assessment (PISA), which is governed by the Organisation for Economic Co-Operation and Development ([OECD], n. d., 2003, 2008, 2009, 2013), as well as the Progress in International Reading Literacy Study (PIRLS) and the Trends in International Mathematics and Science Study (TIMSS), which are both administered by the International Association for the Evaluation of Educational Achievement (IEA). These international assessments are used to examine educational systems across countries and to determine whether students have the knowledge and skills required to participate in society (IEA, 2007).

There are also national and provincial LSAs. The Pan-Canadian Assessment Program (PCAP), a national assessment, examines cross-provincial and territorial educational system differences (CMEC, 2007). The Educational Quality and Accountability Office (EQAO) assessment, a mandatory assessment in the province of Ontario, Canada, provides yearly comparable results in various disciplines in relation to the knowledge and skills required by the provincial curriculum (EQAO, 2009). Most Canadian educational jurisdictions (i.e. provinces and territories) administer similar assessment programs (CMEC, 2008; Taylor & Tubianosa, 2001; Volante & Jaafar, 2008).

In addition to the assessment of curriculum content (e.g., reading), LSAs also evaluate thinking-related skills, such as critical thinking (Edmonds, 2006; Rijmen, 2011). Thinking is defined as a "…human activity, which involves cognition (knowing), affect (feeling)…", and metacognition (Moseley, Elliot, Gregson, & Higgins, 2005; p. 375). The work by Benjamin Bloom (Bloom, 1956; Bloom, Engelhart, Furst, Hill, & Krathwohl, 1986) became a fundamental part of the educational curriculum for the teaching and assessment of thinking skills (Anderson, 2003; Booker, 2007). His widely used taxonomy, which details the process of building on lower level thinking skills to produce Higher Order Thinking (HOT) skills, includes knowledge, comprehension, application, analysis, synthesis, and evaluation (Booker, 2007). As research continued to develop in the area of HOT, other terms were suggested, such as Productive Thinking (PT; Moseley et al., 2005). Lower level thinking skills have been defined as the ability to organize and form concepts (Moseley et al., 2005) and restate information (King, Goodson, & Rohani, 1998; Lewis & Smith, 1993; Newman, 1990). HOT skills have been described as the "…engagement with ideas, objects, and situations" (Alexander et al., 2011; p. 53), connecting and synthesizing information (Quellmalz, 1985), and finding a solution that provides a productive outcome (Moseley et al., 2005).

Other contextual elements to consider when examining the assessment of thinking skills in the context of LSAs are the theoretical underpinnings underlying the assessment, such as assessment models. Thinking skills and how they are measured can potentially be influenced by the LSAs' theoretical underpinnings; which, as mentioned in the Standards for Educational and Psychological testing, should be explicitly stated by validators (American Educational Research Association [AERA], American Psychological Association [APA], & National Council in Measurement and Evaluation [NCME], 2002). However, based on the lack of clearly defined

assessment models within the literature, test developers of LSAs may be implicitly rather than explicitly utilizing assessment models. As a result, the actual thinking skills assessed by the LSAs may differ from those initially intended.

The Research Problem

LSAs measure not only content-related skills, such as reading, but also thinking-related skills, such as critical thinking (Chiu & Seo, 2009; Tindal & Nolet, 1995). Recently, there has been an increased need to understand the cognitive strengths, including the thinking skills, evaluated by LSAs (Leighton & Gierl, 2007) in order to improve instructional design and student learning (Jang, 2008). Most LSAs, such as PIRLS, only present an overall appraisal of student performance and provide little information as to the specific cognitive strengths or weaknesses of students. Having this information would assist in the meaningful interpretation of test results and could have an impact on improving student learning (Jang, 2008). Therefore, with the increase use in standardized testing and the need for educational systems to be accountable (EQAO, 2012), there is a need for test developers to be explicit as to their goals, content, definitions, development, administration, scoring, and reporting. The lack of LSAs' clearly stated definitions or accompanying information regarding their underlying foundational orientations can lead to the misunderstanding and misuse of their results (Farr & Carey, 1986). Hence, the definitions and foundational bases of an assessment must be explicitly defined (AERA et al., 2002), especially to those individuals who will be affected by the assessment results. For example, if parents and educators are under the impression that the assessment results reflect students' complex thinking skills, but this is not actually true, then this would lead to potential issues when interpreting results. This is especially important in the case of high stakes tests (e.g., No Child Left Behind [NCLB]) and other assessments that evaluate thinking-related skills and abilities, such as the

Medical College Admission Test (MCAT®, 2012). The MCAT® evaluates critical thinking and problem solving skills and applicants must pass the test in order to gain admission to medical school; pointing to the need for test developers to be clear as to the extent with which these assessment actually measure these skills.

An additional example is provided by Luce-Kapler and Klinger (2005). When interviewing students soon after taking a high stakes LSA, they discovered that students' views of the test's purpose differed considerably from the intended purpose of the test. Students misunderstood the purpose of the test and approached the items differently than what was proposed by the test developers. That is, students thought the literacy test was assessing their English skills rather than their cross-curricular literacy. Students' differing item interpretations may, thus, interfere with the assessment's given purpose, such that the items are not accurately testing the construct (e.g., problem solving; Haladyna, 1997), that it is alleged to measure. Taken together, the constructs (e.g., thinking skills) measured by the LSA should be accurately portrayed in the test development documents, within the test results, and within the final report. The key is to develop a detailed test blueprint to assess the different skills of interest (Haladyna, 1997).

The main goal of the present study is therefore to deconstruct, examine, and clarify, using a validation approach, the thinking skills that are assessed by two existing LSAs of reading. In addition, due to the assessment influences that have impacted the development of LSAs over the years (Giordano, 2005), it was also deemed important to examine the assessment perspectives underlying the two LSAs. Because reading is an area of the curriculum that is widely assessed within schools regionally, nationally, and internationally and is a fundamental part of the learning process (Chudowsky, Pellegrino, & Glaser, 2001; McKown & Barnett, 2007), the two

LSAs examined both assess reading. Providing tools to help identify the different thinking skills actually measured by LSAs and their inter-relationships will help inform both researchers and practitioners and aid in developing LSA designs that will assess thinking as intended. The various key definitions, relevant theories and perspectives, including validation, are presented and discussed in detail within the thesis but some background is provided in the following section.

Theories, Perspectives, and Definitions

LSAs have been shaped by testing and assessment movements of the past hundred years (Abu-Alhija, 2007; Simner, 2000); namely, the scientific paradigm, the mental measurement movement, and the educational assessment movement (Shepard, 2000; Simon, Erickan, & Rousseau, 2012). Early in the 20th century, scientific inquiry was at the forefront in the field of science (Cochrane, 2007; Kuhn, 1961). Scientific experiments followed specific steps: hypothesize, collect and analyze data, and report results (Cochrane, 2007). Psychologists became interested in using such methods of inquiry to quantify intelligence (Giordano, 2005), which led to the mental measurement movement (Simon et al., 2012). The mental measurement movement was defined by the use of intelligence testing. Although intelligence testing was initially used for military recruitment purposes, due to its popularity, it became part of the recruitment process for other occupations (e.g., typewriter repair person; Giordano, 2005). Once educators and educational leaders recognized that intelligence testing could be used in education, assessments were developed to place students into educational programs based on their scores (Giordano, 2005). By the mid 1900's, when education was accessible by all, educators began to use assessments to examine program-level and district-level results (Linn, 2000; Simon et al., 2012). In the early 1980's, Margaret Thatcher, the British Prime Minister, and Ronald Reagan, the

United States President, had similar aims for educational reform. Specifically, they were both interested in the compulsory use of standardized testing in the educational curriculum (Arnove, Altbach, & Kelly 1992; Edwards, 1989). However, by the late 1980's to early 1990's, educators developed performance assessments with the goal to improve instruction and learning (Stecher, 2010). However, due to their low reliability, performance assessments were gradually replaced with multiple choice and short answer items (Abu-Alhija, 2007; Giordano, 2005; Simon et al., 2012; Stecher, 2010; Ungerleider, 2003).

In the current LSA context, items are typically developed to assess thinking. They target either lower level thinking or HOT skills (Haladyna, 1997). As mentioned earlier, lower level thinking skills are used to organize, understand, and develop meaning (Moseley et al., 2005), whereas HOT skills involve the ability to think critically and creatively, reason, and solve problems. Items assessing critical thinking, for example, would therefore encourage students to observe, compare, evaluate, and make a judgment. Similarly, items assessing problem solving would require students to deal with a specific issue using specific steps: identifying, defining, analyzing, and proposing a solution. However, for a given item, depending on previous experience, students may actually activate HOT, while others may use lower level thinking (Lewis & Smith, 1993). This may apply particularly to items assessing creativity because this skill also requires flexibility, discovery, and innovation. Despite the fact that creativity may be difficult to assess in a large-scale context, elements of creativity may still be evaluated (Haladyna, 1997), such as asking students to provide a novel response. In any case, when test developers wish to assess HOT skills, they should provide explicit details as to their intentions.

Following the examination of the assessment movements and the research conducted by Hager and Butler (1996) and Serafini (2001), the current study focused on two potential

perspectives underlying the assessment of thinking skills: (1) Product-based; and (2) Process-based. An assessment perspective (or model in the case of Hager and Butler, 1996 and Serafini, 2001) refers to the dimensions that drive the development of the assessment. It is derived by the level of student and teacher involvement, the way in which student information is collected, the purpose of the assessment, and the way in which the results are reported and to whom (Serafini, 2001). Concerning the product-based perspective, students are considered separate from the assessment; meaning resides in the text; and scoring involves selected-response items. In the context of a LSA, test developers are interested in the product of the assessment; questions are typically developed with fixed responses; and results are used to compare students and schools. On the other hand, according to the process-based perspective, students are considered important contributors to assessment development and test developers work with students and teachers to determine the important elements to be included in a LSA. The assessments derived from a process-based perspective are considered part of the learning experience, such that students have the opportunity to contribute unique information; scoring is open and allows many interpretations; and the main goal of the assessment is to improve learning. There is a time and place for each perspective as each provides test developers a way in which to assess learning. The goal of the assessment is the deciding factor as to which perspective is the most appropriate.

The goal of the current study is not to consider these two perspectives as mutually exclusive or comprehensive, but to use some of the main elements of each perspective to help determine what aspects are present within the two LSAs of reading under focus. The models proposed by Hager & Butler (1996) and Serafini (2001) are explored in more detail in the literature review and the resulting two approaches used in this study are addressed in the conceptual framework.

As the current study examines two LSAs of reading, the content of reading and its link to thinking are also explored. Reading comprehension is a multidimensional process that has many interacting elements (Ruddell & Unrau, 2004). The reading process involves both basic and higher level skills (Alderson, 2000). A reading strategy is how the reader plans to read, including what skills are to be implemented (Power, 2007). Due to the complex nature of reading, models and theories have been proposed over the years to help explain the process of reading (Tracey & Morrow, 2012). The view from the conditioned learning perspective, which includes theories such as Classical Conditioning (Skinner, 1977; Tracey & Morrow, 2012), is that reading is a learned behaviour that can be practiced to achieve a desirable outcome. The cognitive perspective includes models such as the Automatic Information Processing Model (LaBerge & Samuels, 1974) and the Interactive-Compensatory Model (Stanovich, 1980; Tracey & Morrow, 2012). In this perspective, reading is viewed as a progression of stages from visual to episodic memory systems (Tracey & Morrow, 2012). This perspective is ideal for understanding the mechanisms of thinking involved in the process of reading. In the context of LSAs, test developers focus on how different sensory information, such as the reading passages and corresponding questions, affect how students think. For example, test developers may identify specific types of text that elicit different types of thinking skills. Another perspective that is not entirely in line with the cognitive perspective is the transactional theory of reading, proposed by Rosenblatt (1982, 1985, 1986, 1993, 1995), which describes reading as the interaction between the reader and the text.

Although there is a range of theories and models that have influenced the field of reading over the years, the cognitive perspective seems to be at the forefront in describing the reading process. Within this perspective, the current study focuses on the work of Mosenthal and Kirsch

(Mosenthal, 1996), as their work has been a pivotal piece in the development of the PISA reading section (OECD, 2013). The aim of their work is to understand the strategies of document literacy (Mosenthal, 1996). There are two aims of reading: *to comprehend* and *to do* (Mosenthal, 1996; Mosenthal & Kirsch, 1991a). While reading to comprehend deals with the maintenance of knowledge, reading to do is concerned with the purpose of the reading task (Mosenthal, 1996; Mosenthal & Kirsch, 1991a), which is pertinent to the context of LSAs. For each aim, there are several stages to processing a document. When reading to do, the first stage in processing a document is to identify a goal. In the context of LSAs, these goals are determined by the questions derived from the reading passage. The goal of the second stage is to identify which information is available and is missing. Again, in the case of LSAs, this would be determined by examining the reading passage for information that is directly identified in the passage, and whether there is information that is not explicitly available in the text. The third stage is to match the information from the text to the given purpose or questions. That is, individuals would match information in a reading passage to the question in the LSA. If explicit information is not provided, previous knowledge may be required for answering the question. The fourth stage is to fill in the information by way of 'nodes'. Nodes are the elements of the question that relate to information in the text. For example, if the question asks "why", the individual may look for information relating to a goal in the reading passage. The fifth stage is to review how sufficiently the question was answered. At every stage, individuals can go back to a previous stage to complete the question (Mosenthal & Kirsch, 1991b). Their work fits well into the realm of LSAs and the assessment of thinking skills because, depending on the document processing strategy undertaken, different skills are required. For example, if the goal of reading (from a question) is

to locate information that is explicitly stated in the text, then lower level thinking skills will be implemented. This work is further discussed in the literature review.

A validation approach is used in this study to examine the various skills and underlying assessment perspectives of two LSAs. The concept of validity has been debated over the years and there remain differences in opinion as to its definition and process (Lissitz, 2009). The Standards for Educational and Psychological testing define validity as "… the degree to which evidence and theory support the interpretations of test scores…" (AERA et al., 2002; p.9). In general, a validation approach aims to gather evidence regarding the interpretation of test scores and to ensure results are aligned with the intended assessment goals.

To conduct a validation that closely aligns with theory, research has suggested that the work by Messick (1995) and Kane (1992) be implemented (Wolming & Wikström, 2010). Messick (1995) proposed six aspects associated with a construct validity framework: (1) Content, which is the examination of the superficial features; (2) Substantive, and the examination of response processes; (3) Internal, and the examination of the structure of the assessment (e.g., factors); (4) External, which is the comparison of the construct to other similar or dissimilar constructs; (5) Generalizability, which refers to the assessments' consistency in different settings and samples (e.g., low stakes versus high stakes); and (6) Consequential or the intended or unintended impact from the assessment (Lane, 1999; Messick, 1995). Additionally, Kane (1992) proposed an argument approach to validity, in which the validator offers a validity argument and supports it with evidence. Kane's (1992) argument approach helps frame the validation approach with Messick's (1995) aspects of construct validity being used as the sources of evidence. According to Messick (1995) and Kane (2009), multiple sources of evidence are required in order to justify the relevance and appropriateness of inferences made from

assessment results. Validation will be further clarified in the literature review and the validation approach adopted for the current study will be fully described in the conceptual framework.

Organization of the Thesis

The thesis is organized into seven sections, including the introductory chapter: (1) Chapter 1: Introduction; (2) Chapter 2, literature review; (3) Chapter 3, conceptual framework; (4) Chapter 4, methodology; (5) Chapter 5, results; (6) Chapter 6, discussion; and (7) Chapter 7, conclusion. The introduction provides an overview of the study including the research problem. The literature review covers the history of LSAs including the models that have been implemented; the theories of reading and of thinking skills that are relevant to LSAs; and the validation process. In addition, it provides the basis for the conceptual framework. The chapter on the conceptual framework introduces the theoretical foundation for thinking skills, reading, and educational assessment models, in the context of LSAs and test validation. The methodology section presents the research design, the procedure, the coding protocol for the qualitative analysis and the proposed quantitative analysis. The results section offers qualitative and quantitative findings with respect to the topic of thinking skills and assessment perspectives in the context of the two LSAs. The discussion section presents the general interpretation of the results, the study's contribution to the field of education, its strengths and limitations, and suggestions for future studies. Finally, the conclusion summarizes the main ideas and findings of the study.

Although the current study is mainly written in the third person, I feel it is important to provide my own biases and personal views. I have largely been trained within the post-positivist paradigm. However, as part of my current study, and given the nature of the research questions, it became obvious that I had to take on the task of not only conducting quantitative analyses of

assessment results, but also rigorous qualitative analyses of technical documents, via a pragmatic view. My personal views and biases are therefore further described in the conceptual framework including how it impacted the identification of the different thinking skills, and assessment perspectives, cited within the literature, in addition to the use of a validation approach.

Chapter Two: Literature Review

LSAs can measure both content- (e.g., reading) and thinking-related skills (e.g., critical thinking) (Haladyna, 1997). In addition, reading is an area of the curriculum that is widely assessed within schools regionally, nationally, and internationally, is cross-curricular, and is a fundamental part of the learning process (Chudowsky, Pellegrino, & Glaser, 2001; McKown & Barnett, 2007). Therefore, the current study examines the thinking skills of two LSAs of reading. In order to examine the construct of thinking, a validation approach was conducted. Therefore, the purpose of this chapter is to present the literature relating to the development of LSAs, the thinking skills and different perspectives of assessment assumed or intended within the context of LSA in reading, as well as the research on validation. First, the history of LSAs, including the influences, purposes, and models of assessment, is introduced. The second section details the different thinking skills typically assessed, including those that are considered lower and higher level. As the current study examines two LSAs that measure reading, the goal of the third section is to review the many perspectives of reading, including the link between thinking and reading. Assessment results must lead to inferences associated with the intended object of assessment (i.e. targeted thinking skills). Therefore, the final section of the literature review includes an overview of the concept of validity, including the work by Messick (1995) and Kane (1992).

LSAs

History of LSAs. In the early 20th century, scientific inquiry was an important part of the field of science (Cochrane, 2007; Kuhn, 1961) and experiments were designed to follow specific steps to ensure that results were unbiased. Scientific inquiry influenced the emergence of the mental measurement movement, during which psychologists began to consider methods to quantify intelligence (Giordano, 2005). The work of Francis Galton, Charles Spearman, Alfred

Binet, Lewis Terman, and Arthur Otis were all associated with this movement (Simon et al., 2012). After the First World War, there was an increase in the use of these intelligence tests (Giordano, 2005; Simon et al., 2012).

The educational assessment movement was first initiated in the early to mid-twentieth century, when teachers would use standardized tests to select students into programs or schools, based on their scores (Giordano, 2005). Educators and educational leaders began to recognize that intelligence testing, being applied by psychologists, could also be incorporated into the educational system. Tests were adapted from intelligence tests, with questions having fixed responses, to group students into those who were "gifted", "developing normally", had "learning disabilities", and were "juvenile delinquents" (Giordano, 2005). As a result, students' scores on these tests would determine where they were grouped and if they were selected into a specific program or school. Achievement results were eventually used to examine program-level and district-level measures (Linn, 2000; Simon et al., 2012). That is, LSAs were employed at the level of educational systems to ensure programs were meeting curricular expectations (Giordano, 2005) by measuring group performance (Linn, 1986). Decisions could then be made regarding how programs were developed, where funding would be focused, and how schools were doing. The aim was to improve educational systems (Linn, 2000). Finally, by the early 1980's the United States was using standardized testing as a necessary part of the education system (Arnove et al., 1992)

In the late 1980's and early 1990's, there was interest in using large-scale performance assessments to improve instruction and learning. During this time, several U.S. states, including Vermont (Koretz, McCaffrey, Klein, Bell, & Stecher, 1992), Kentucky, Maryland, Washington, and California (Stecher, 2010) adopted performance assessments to their national testing

programs. The implementation of these assessments revealed that it was possible to measure higher-level cognitive skills in a large-scale format. However, these type of assessments required extra attention to ensure psychometric quality and rating consistency due to their use of open forms of data collection (e.g., portfolios) and views of learning (e.g., students are active creators of knowledge; Stecher, 2010). Due mainly to financial costs and politics (e.g., NCLB), states began to return to the traditional testing programs (Stecher, 2010).

Assessment Models. Based on the many assessment movements and influences of the past 100 years (Giordano, 2005; Heald-Taylor, 1998; Short & Burke, 1994), different models of assessments have emerged. Researchers, such as Hager & Butler (1996) and Serafini (2001), proposed models that describe the different paradigms of educational research.

Hager and Butler (1996) discussed how the scientific assessment model, which focuses on providing an objective measure of student achievement, has been the strongest force driving educational assessment (Hager & Butler, 1996). However, with the advancements in cognitive science, assessment models have shifted to account for changes in the study of intelligence and cognition (Hager & Butler, 1996). Therefore, Hager and Butler (1996) suggest that there are two models of assessment: (1) Scientific measurement model, and (2) Judgmental model. They break down both models by the metaphysical assumptions about humans, or how thinking is viewed; the epistemological assumptions, or the perspective of assessment; the assumptions about the measurement process, or what is measured (e.g., scientific measurement model: measure 'g' factor); the assessment principles, or how the tasks are assessed (e.g., scientific measurement model: one right response); and assessment practices, or the type of conditions surrounding the assessment (e.g., scientific measurement model: focus on test instruments). While the scientific measurement model is focused on the intelligence approach (i.e., quantifying) to achievement,

the judgment model is focused on the cognition approach (i.e., thinking) to achievement. In addition, the scientific measurement model includes structured problems with defined answers, the assessment of simple discrete tasks, the emphasis of one correct response, the focus on controlled test conditions (e.g., 3 hour paper and pencil test), and the focus on objectivity. On the other hand, the judgement model includes unstructured problems, the assessment of the processes of achievement, responses with multiple interpretations, various types of work (e.g., group work, portfolios), and the collaboration of students and teachers in the design of their assessment tasks (Hager & Butler, 1996).

The work by Hager & Butler (1996) arose after the emergence of the use of performance assessments and the interest in assessing complex learning. Their model was an initial examination of a new perspective on the development, administration, scoring, and reporting of assessments. Their work helps to understand the different mechanisms underlying the two perspectives or models of assessment and provides tangible characteristics of each. However, their work does not account for more recent assessment reforms and work that has been done after the rise and fall of performances assessments in the 1990's.

More recently, Serafini (2001) suggests that there are three assessment models: (1) 'Assessment as measurement'; (2) 'Assessment as procedure'; and (3) 'Assessment as inquiry'. Measurement is defined as "…the assignment of numerals to behaviour according to rules" (Brown, 1983; p. 11). For this model, Serafini (2001) states that "… meaning is believed to reside within the text, and only one interpretation or judgement is accepted in the standardized tests" (p. 385). As described by Serafini (2001), an assessment developed based on this model would focus on goals, such as judging the quality of overall programs, improving educational systems, and comparing students. A fixed scoring approach and multiple choice items are

favoured. The aim of the measurement-based model is to assess the product (i.e., the student understands what is read) of achievement rather than the processes used to achieve (e.g., the thinking skills implemented by the students). The procedure model can be considered an extension of the measurement model, as the only difference is the inclusion of multiple methods of data collection to assess learning, such as the inclusion of open-ended questions rather than just multiple choice formats.

The 'assessment as inquiry' or inquiry-based model is closely linked to performance assessments, as these assessments include a wide range of activities to evaluate student achievement. An assessment based on this model is developed with the goal to improve learning at the individual level and assess processes. Students and teachers are actively involved in the assessment development. For example, students can be assessed on higher-order learning outcomes and the results can then be used to facilitate learning (Serafini, 2001). However, such tests are more costly and time consuming than those based on the measurement model due to the extra attention and resources required for their development, administration, and scoring (Serafini, 2001).

Although the models by Serafini (2001) help identify tangible elements of assessments, they describe only a limited account of student learning; offer an incomplete description of the difference between the measurement and procedure model; explain little in the way of how the models are grounded in the history of assessment; and provide limited guidance as to how they can be applied. It must be noted that Serafini's (2001) initial goal was to document the different assessment models in the context of classroom assessment; however, the elements described also bear weight in the examination of LSAs. Though limited, Serafini's (2001) work introduced the

idea of different perspectives/models of assessment. Therefore, the idea of potential differences in underlying perspectives of assessment provides a focus for the current study.

Upon closer examination of the assessment movements and taking into account the categorization of assessment models proposed by Hager & Butler (1996) and Serafini (2001), the current study it may be more appropriate to identify two perspectives: (1) Product-based, which is similar to elements of the measurement and procedure model; and (2) Process-based, which approaches the inquiry model. These two perspectives are not mutually exclusive. Test developers may include different elements that tap into both perspectives. These two perspectives are complex and describe differences in relation to student and teacher involvement, development, administration, and reporting of LSAs. Due to the many elements of each perspective and how they correspond to specific goals, it is, therefore, important for test developers to be explicit as to which perspective elements underlie the test and to show how these elements support the assessment goals. There is a place for each perspective; however, it depends on the goal of the assessment. The conceptual framework will further define the two assessment perspectives adopted for the current study.

It is apparent that, based on the different movements, the goals of the LSA will ultimately influence the orientation and development of the assessment framework (i.e., the blueprint for how the assessment is developed, administered, and reported; Abu-Alhija, 2007). Therefore, it is important that there are common practices to which test developers can adhere to. The *Principles for Fair Student Assessment Practices for Education in Canada* (1993) was developed by a working group to provide guidelines for both internal (i.e., classroom) and external (i.e., LSA) assessment practices. One recommendation is that test developers be clear about their purpose.

For example, it is suggested to "[p]rovide evidence that the assessment yield results that satisfy its intended purpose" (p. 15) and that test developers should

> [d]escribe the process by which the method was developed. Include a description of the theoretical basis, rationale for selection of content and procedure, and derivation of scores (p. 15).

In other words, to provide a fair assessment, test developers should be making clear links between their purpose, the way in which the test is to be developed, including underlying theories and models, and the way in which the results are to be interpreted in light of the given purpose (Principles for Fair Student Assessment Practices for Education in Canada, 1993).

Thinking Skills

After the introduction of Bloom's cognitive taxonomy (Bloom, 1956), many researchers developed definitions of HOT. Greeno (1989) suggested that HOT is concerned with two aspects: one is involved with the skills required to solve new problems and the other is to learn new domains. Carnine and Kameenui (1992) proposed that HOT involves the integration of different elements including ideas and concepts, while Ivie (1998) indicated that HOT contains three criteria: the use of abstract structures when thinking, the organization of information, and the use of rules and judgement. Lewis and Smith (1993), after reviewing previous literature on HOT, suggested individuals take in new information to rearrange it for a purpose. They also indicated that HOT skills include decisions on what to believe and what to do, as well as aspects of prediction and creation Additionally, Underbakke, Borg, and Peterson (1993) defined HOT as the combination of old and new information to achieve a purpose. Quellmalz (1985) also suggested that HOT is the purposeful engagement with thoughts and include processes of identifying, defining, clarifying, judging, and connecting information, while King et al. (1998)

stated that HOT includes critical, creative, and reflective skills, but also metacognitive thinking. More recently, Alexander et al. (2011) suggested that HOT is a mental engagement with ideas, objects, and situations. HOT has also included analyzing, synthesizing, applying, and evaluating skills (Paul & Nosich, 1992). Another recent definition, put forth by Leighton (2011), describes HOT as involving four components of knowledge and skills: (1) inquiring or identifying; (2) applying multiple and appropriate criteria; (3) developing and/or generating; and (4) regulating and appreciating the cognitive effort. On the other end of the scale, lower level thinking is the routine application of information including restating, observing, and classifying information (King et al., 1998; Lewis & Smith, 1993; Newman, 1990). Overall, HOT is complex. HOT is the purposeful engagement with many informational elements, such as ideas, rules, concepts, and situations, using aspects of knowledge and skills that results in a productive outcome.

Some researchers have replaced HOT with the term Productive Thinking (PT) to avoid labelling one thinking skill as lower or higher level than the other. Moseley et al. (2005) proposed a framework that took into account 35 previous frameworks and taxonomies. Essentially, they wanted to avoid the hierarchical nature of HOT and emphasized that PT results in a productive outcome, such as a decision or judgment. Lower level thinking was substituted with 'building understanding', and included organizing, forming concepts, and developing meaning. Recent research has yet to implement the framework proposed by Moseley et al. (2005), including the term PT, and continues to use HOT and lower level thinking (e.g., Leighton, 2011). However, their work provides a comprehensive account of the different frameworks of thinking within education.

Based on the literature, the common terms associated with thinking include lower level thinking skills (e.g., building understanding, Moseley et al., 2005), critical thinking (e.g., using a

set of rules to judge; Ivie, 1998), creativity (creating a new idea; Lewis & Smith, 1993), and problem solving (e.g., finding solutions; Greeno, 1989). The following section reviews the definitions associated with the different thinking skills that are most commonly referenced within the literature, while the conceptual framework will further clarify the skills pertinent to the current study.

Lower level skills have also been termed core skills (Marzano et al., 1989) and building understanding skills (Moseley et al., 2005). Lower level skills include restating ideas, developing meaning, working with patterns and rules, forming concepts, and classifying ideas. Therefore, lower level skills are needed to organize, understand, develop meaning, and conceptualize information (Moseley et al., 2005).

Critical thinking skills have been featured prominently in the literature (Lewis & Smith, 1993). As a result, many definitions of critical thinking have been put forth. Greeno (1989) suggested that, in order for one to think critically one needs to have reflective thinking. Critical thinking has also been viewed as the evaluation of thinking with the goal to improve it (Paul & Elder, 2005). It is also considered the process of analyzing, synthesizing, or evaluating information that is generated from multiple areas such as observations and experiences (National Council for Excellence in Critical Thinking, 1987). Moreover, critical thinking skills have been associated with attributing, comparing, contrasting, classifying, sequencing, prioritizing, drawing conclusions, determining cause/effect, analysing for bias, analysing for assumptions, solving for analogies, and evaluating (Fogarty & McTighe, 1993). Overall, the definitions of critical thinking indicate that it is a process of observing, comparing, evaluating, and making a judgement.

Problem solving has been predominantly associated with the fields of science (e.g., mathematics) and psychology (Lewis & Smith, 1993; CMEC, 2007). Problem solving involves the process of identifying, defining, and analyzing a problem, proposing solutions, experimenting, and drawing conclusions (Haladyna, 1997). Problem solving entails recognizing and identifying something to be solved, and defining, analyzing, and proposing a solution.

Creativity is a HOT skill (Lewis & Smith, 1993). Creative skills can include brainstorming, visualizing, personifying, and inventing (Fogarty & McTighe, 1993). Greeno (1989) suggested that creativity involves the "…flexibility of thinking and restructuring of understanding in innovative ways" (Greeno, 1989; p. 140). Creativity, as stated by King et al. (1998), involves, but is not limited to, the discovery of new problems and the use of principles in new situations. Additionally, Marzano et al. (1989) proposed that creative thinking is the use of free flowing thoughts (Marzano et al., 1988). According to these definitions, creativity generally involves flexibility, discovery, and innovation (Craft, Jeffrey, & Leibling, 2001); yet, it is not known to what extent this type of skill should be evaluated in the context of LSAs. In general, creativity can include multiple interpretations and, therefore, the specific definition of creativity used will help identify what evidence is needed to evaluate this skill (Treffinger, Young, Selby, & Shepardson, 2002).

Thinking and Reading

Similarly to the history of assessment, reading research over the years has gone through several shifts in perspective and, as a result, has been influenced by many models and theories (Ruddell & Unrau, 2004; Tracey & Morrow, 2012). Reading models, based on theories and research, provide a concrete way to understand the components of the reading process. Particular areas that have guided the many theories and models include the influence of conditioned

learning (e.g., behaviourism), natural learning (i.e., the brain), information processing (i.e., cognitive perspectives), sociocultural learning (i.e., elements of the social environment), and engaged learning perspectives including the importance of the text and motivational elements (Alexander & Fox, 2004). Models of reading ultimately, even implicitly, influence the instruction and assessment of reading (Ruddell & Unrau, 2004).

The era of conditioned learning (e.g., Operant Conditioning Theory; Skinner, 1977; Alexander & Fox, 2004), led researchers to identify reading as a conditioned process that could be programmed. Reading, thus, contained isolated skills, that could be broken down and practiced until a desirable outcome was achieved. Interest lay in the observable behaviours elicited during reading and led those in education to focus on visual signals and sounds. As a result, phonics instruction became an integral part of reading during this era (Alexander & Fox, 2004).

As educators and the reading community became dissatisfied with the limited view of thinking and learning associated with conditioned learning, the focus of reading shifted to elements of the mind (Alexander & Fox, 2004). With increased interest in the study of cognition, different waves of information processing models began to emerge (Ruddell & Unrau, 2004), such as Gough's Model and the Automatic Information-Processing Model (Gough & Turner, 1986; LaBerge & Samuels, 1974; Tracey & Morrow, 2012). Bottom-level models describe reading as the input of sensory information to deeper levels of lexical and semantic structures (e.g., Gough's model; Ruddell & Unrau, 2004). Letter and word input are processed in a linear fashion, such that higher levels within the process cannot change or modify lower levels. As a result, these models do not account for higher level processes (Ruddell & Unrau, 2004; Rumelhart, 2004; Tracey & Morrow, 2012). On the other hand, top-down models take into

account the internal mechanisms of memory and the impact on the nature of reading (e.g., Schema theory; Ruddell & Unrau, 2004).

In contrast to bottom-up or linear descriptions of the reading process, theories such as the Automatic Information Processing Model (LaBerge & Samuels, 1974) were put forth to include both of these aspects. The model posits that the automatic information processing in reading is a progression of stages that begin when visual information is transformed into visual, phonological, and episodic memory systems. The degree to which information passes through these stages depends on learning and whether it is accurate and automatic. Additional models include the Interactive Model and the Interactive-Compensatory Model (Stanovich, 1980), which suggest a nonlinear process of reading; enabling researchers to understand higher level thinking in the reading process (Tracey & Morrow, 2012). The Interactive Model postulates that, during processing, there are many modalities in which perceptual input is analysed (e.g., visual and auditory), visual and auditory -input are processed in parallel, and the act of perceiving is an interactive process (Tracey & Morrow, 2012). The Interactive-Compensatory model describes the interactive nature of both vital top-down and bottom-up processes, but also helps explain the reading process of poor and good readers. That is, readers can compensate at different levels to understand what is read. For example, a reader may not be able to identify a word in a sentence, but they may be able to figure it out based on elements in the sentence (e.g., context; Stanovich, 1980).

Apart from cognitive theories and models, the Transactional Theory of Reading, which has been proposed by Rosenblatt (1982, 1985, 1986, 1993), is a model that emphasizes the interaction between the reader and the text. Essentially, efferent- and aesthetic-type reading, based on the selective attention of an individual, will influence the different experiences of the

individual (Rosenblatt, 1985; Tracey & Morrow, 2012). Efferent reading refers to an individual focusing on the public meaning of what is being read to analyze, restate, etc., while aesthetic reading is when the selective attention of the individual is focused on the lived experience during the reading process. Based on the purpose of the text, both efferent and aesthetic reading can take place (Rosenblatt, 1993). A purely efferent text would involve the reading of a newspaper article, directions, etc., in which meaning is abstracted after the event. In comparison, a purely aesthetic text would take into account past experiences of the reader, as the reader would live through what is being read (e.g., a story, a play, etc.). If, for example, a Shakespearean poem is reviewed by an individual, he or she would take an aesthetic stance to review the poem because of the elements, ideas, sounds, scents of the text. On the other hand, if the purpose is to read the poem and find metaphors, analyze, etc., then the individual must take an efferent approach in order to uncover the public attributes of the text. The aesthetic transaction is a purely personal experience; no one can read an aesthetic text for someone else (Rosenblatt, 1995). Therefore, from a pedagogical perspective, providing connections between a student's personal experiences and the material to be learned can enable students to deepen their understanding through an aesthetic response.

These are just some of the different perspectives of research that have been proposed to help explain reading and the reading process. The range of reading models is diverse. However, cognitive models of reading have been important in describing the role of thinking in the reading process. One such cognitive model concerns the work of Mosenthal and Kirsch (Mosenthal, 1996; Mosenthal & Kirsch, 1991b), which has been central to the development of the reading section of the PISA (OECD, 2013). As previously mentioned in the introduction, there is reading to comprehend and reading to do (Mosenthal, 1996; Mosenthal & Kirsch, 1991b). Reading to do

is pertinent to the current study as it relates to reading for a purpose and, in the context of an LSA, answering assessment questions. There are several stages of document processing: (1) Stage 1, identifying a goal; (2) Stage 2, identifying the information that is specified and what is requested; (3) Stage 3, matching the information to the goal; (4) Stage 4, completing the goal; and (5) Stage 5, verifying how well the information matches the initial goal. In addition to the stages of document processing, different assessment questions may require students to implement several types of document processing strategies.

Given that the LSAs to be examined in this study are in reading, it is also important to present the link between thinking skills and document processing strategies. Document processing strategies occur at stage 3 and include specific strategies: *locate, cycle, integrate*, and *generate*. Locate strategies refer to the retrieval of explicitly stated information within in the text. For example, if an assessment included the question "Where is the main character located?" the student would look in the reading passages for references to location. This type of strategy is closely linked to lower level thinking skills, as the information is explicitly stated in the text (Mosenthal & Kirsch, 1992). Cycle strategies refer to the use of numerous locate matches. That is, the assessment question contains more than one element that must be located in the reading selection. For example, if a reading selection includes a list of names and class schedules, a question may ask "On what day did Tanya and John have a mathematics class?" Students would be required to first look through a list of student names to identify Tanya and John, and then they would have to go through each day to determine what day Tanya and John had math. Again, information would be explicitly stated in the text, but students would have to examine several parts of the reading selection to find the response. Cycle strategies still make use of explicitly stated information, but in a complex way. Integrate strategies are the processes by which students

compare and contrast the information retrieved when conducting locate and cycle strategies. For example, for a question asking "On what days do Jeremy and Sandra have similar class schedules?" a student would be required to go through the student list to find both names and then compare their schedules on each day of the week. This type of strategy involves students' HOT skills, as they need to take information in the text and produce new information (i.e., reporting the days they have similar schedules). That is, the actual response to the questions is not explicitly stated in the text. Finally, generate strategies involve the use of previous knowledge, as the response for the question cannot be gleaned from the text. For example, an item may state "Name another type of extinct animal not identified in the list". Again, this type of strategy involves students' HOT skills, as it requires them to use a number of informational elements (e.g., situations, experience, etc.) to produce an outcome (Mosenthal, 1996). This strategy also overlaps with elements of creativity, as the use of this strategy may enable students to provide a novel response. Overall, the different document processing strategies are closely linked to the types of thinking skills, which will be explored in the current study of two LSAs that measure reading.

Assessing Thinking in the Context of LSAs

LSAs typically assess two dimensions: content (e.g., mathematics; Chiu & Seo, 2009) and cognitive processes or thinking-related skills (Edmunds, 2006; Rijmen, 2011). Reading is a process involving thinking skills (CMEC, 2012) and assessment questions can be developed to assess these skills (Haladyna, 1997). LSAs provide a definition of reading and what skills it contains. For example, the draft PISA 2015 reading framework (OECD, 2013) states that

> Reading literacy is understanding, using, reflecting on and engaging with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society. (p. 9)

while, Mullis, Martin, and Sainsbury, from PIRLS, state that

> Reading literacy is the ability to understand and use those written language forms required by society and/or valued by the individual. Readers can construct meaning from texts in a variety of forms. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment.

PIRLS (IEA, 2012) identifies four reading skills: (1) retrieving explicitly stated information; (2) making straightforward inferences; (3) interpreting and integrating ideas and information; and (4) examining and evaluating content, language, and textual elements. The first skill is similar to the use of lower level skills, such as restating ideas. The second, third, and fourth skills relate to the use of critical thinking (e.g., making a judgement), and are also found within the PISA assessment framework (OECD, 2009). Similarities in both definitions include understanding and using. However, the focus of PISA in on the engagement of texts, while PIRLS is concerned with the construction of meaning from a variety of text forms. In addition to reading skills, the assessment framework of the National Assessment of Educational Progress (NAEP, 2008) refers to the importance of problem solving skills when students are reading literary texts. That being said, it appears that lower level thinking, critical thinking and problem solving skills are either explicitly or implicitly addressed in the LSA development frameworks (e.g., item specification table). As depicted, LSAs in reading generally share the common aim of evaluating reading skills, but these definitions are often expressed differently. That is, the language and terms used may differ between LSAs.

In the context of reading, Haladyna (1997) has suggested that questions assessing thinking skills can be developed to tap into lower level thinking, critical thinking, problem solving and creativity. Table 1 presents the suggested item stems for the different HOT skills proposed by Haladyna (1997). Although not exhaustive, the item stems do present some potential phrases to examine these skills. For critical thinking, students may be asked to predict the outcome of a specific scenario (e.g., what would happen if…?) and asked to make a judgment (Haladyna, 1997). To assess problem solving, students may be asked to identify a problem plagued by the main character in a text (Haladyna, 1997) and to provide a possible solution (Haladyna, 1997). Finally, for creativity, the focus is on mental flexibility, thinking outside the box, and producing unique perspectives. For example, students may be asked "How might this story apply to your own life?" (Topping & Bryce, 2004, p. 621), "Come up with a scenario…". "Build your own….", etc. These types of questions enable a student to go beyond the question and provide a completely unique response (Haladyna, 1997). However, the need for partial scoring for items tapping into creative thinking is very important. The specific definition for creativity that is adopted will best help determine what evidence is required. To assess creativity, it is also important to use multiple sources of data to gain an understanding of the student's previous experiences using creativity and whether their responses can be deemed creative (Treffinger et al., 2002).

Overall, developing an assessment to evaluate thinking skills can be difficult, especially HOT, due to the many aspects that must be considered. For example, HOT is generally difficult to assess in general, and test developers must agree on which skills are designated as complex (i.e., HOT; e.g., critical thinking) and also measureable within a large-scale context. This is especially important as some questions developed to assess HOT may, in reality, only

Table 1

Item shells proposed by Haladyna (1997)

Thinking Skill	Item stem	Key Verbs
Critical Thinking	What is the most effective (appropriate) for...? Which is better (worse)...? What is the most effective method for...? What is the most critical step in this procedure? Which is (un)necessary in a procedure? What would happen if...? If this happens, what should you do? On the basis of..., what should you do? Given..., what is the primary cause of...?	Anticipate, appraise, analyze, classify, compare, contrast, critique, defend, distinguish, evaluate, hypothesize, infer, judge, predict, relate
Problem Solving	What do you need to solve the problem? What is a possible solution? Which is a solution? Which is the most effective (efficient)	Answer, compute, conclude, determine, find, figure out, solve
Creativity	Write a persuasive response… Design your own… Plan a… In your own words…	Build, construct, create, design, invent, make, perform, plan, redesign, write

require students to implement lower level thinking, depending on their previous experience (Lewis & Smith, 1993). For example, if a student is asked to define a word in the context of the reading selection and has not been exposed to the word previously, then he or she would need to synthesize (i.e., HOT) the information in the text to determine its meaning. On the other hand, if the student had previously been exposed to the word then he or she would not need to rely on the text to determine its meaning and would thus rely on lower level thinking. As the current study examines the construct of thinking and associated underlying perspectives, it was deemed important to examine these aspects via the study of the literature on validation.

Validation

Validity has paramount importance for researchers and test developers alike. AERA et al. (2002) propose a set of standards for evaluating tests and a framework for ensuring that validity issues are dealt with in a rigorous and thorough manner. The standards stress the importance of using multiple lines of evidence to support a validity argument and offer suggestions for dealing with content, response processes, internal structure, criteria, generalizability, and social consequences. Some of the requirements proposed for frameworks include the full presentation of the theoretical construct and careful attention paid to the test format (AERA et al., 2002; Jonson & Plake, 1998). While the standards provide criteria for a holistic approach to validity, they fail to completely incorporate and detail specific developments and issues in the area of validity and provide in-depth examples for applying their criteria. In light of these limitations, it has been suggested (Wolming & Wikström, 2010) that researchers combine both Kane's (1990, 1992, 2006) view and Messick's (1995) validity framework in order to provide a validation process that closely aligns with theory (Wolming & Wikström, 2010). As a result, the following section will review the work of Messick (1995) and Kane (1990, 1992, 2009).

Samuel J. Messick. Messick (1975, 1980, 1989, 1990, 1994a, 1994b, 1995, 1998) was one of the first to propose that validity should be seen as a unified concept, with *construct validity* representing the term for the overarching concept of validity. He claimed that the traditional view of validity—which included content (i.e., established by expert judges), criteria (e.g., correlating test scores with another test), and construct—was insufficient, because it did not address the social consequences of tests. As a result, Messick (1995) proposed six specific elements of construct validity that must be addressed: (1) content, (2) substantive, (3) structural or internal, (4) external, (5) generalizability, and (6) consequential.

The content element of validation is the most basic aspect of Messick's (1995) construct validity framework. It refers to how well the various subject matter dimensions are represented (Miller & Linn, 2000), including the relevance of the content and its technical quality (Messick, 1995). Content validity includes the scoring procedures, rubrics, guidelines, wording, and the items themselves. In order to provide test content validity evidence, individuals or experts can examine: (1) how well the test specifications parallel the construct domain; (2) whether the items represent the construct domain; (3) whether the content of all related test documents reflects the given blueprint; and (4) whether the scoring procedures (including the rubrics) are relevant to what is being measured. For the most part, an examination of the test content is an expert judgment on how well the content of the test represents the construct domain (Lane, 1999). Accordingly, content validity is test-based, observable and addresses the specifications of the test (Sireci, 1998a). Review subject matter by experts is a vital step to a content study that examines validity (Sireci, 1998a; Sireci, 1998b). In short, content addresses how well the test items, scoring procedures, etc. adhere to the test specifications.

The substantive element involves examining the response processes or the thinking skills that individuals actually implement when answering test items (Lane, 1999). Various methods have been used in this kind of analysis: think-aloud and interview procedures with test takers (Hamilton, Nussbaum, & Snow, 1997); internal reviews, in which pilot tests are conducted with students during the development of an instrument; and external reviews by subject matter experts (Magone, Cai, Silver, & Wang, 1994). Studies that incorporate interviews and think-aloud procedures examine these processes in more detail; nonetheless, some researchers recommend using in-depth descriptions of the processes that are supposed to be elicited from students who complete the test items (Lane, 1999). In cases of secondary data analyses, therefore, these descriptions can provide the best evidence for the substantive element in Messick's (1995) construct validity framework.

Factor structure is commonly used to address the internal or structural element of Messick's (1995) construct validity framework (Kupermintz, Ennis, Hamilton, Talbert, & Snow, 1995; Kupermintz & Snow, 1997; Nussbaum, Hamilton, & Snow, 1997). The internal element refers to the study of how the items, taken together, address the intended construct (Messick, 1995). Methods can include inter-correlations among test items, the use of principal components analysis, factor analysis (Hamilton, Nussbaum, Kupermintz, Kerkhoven, & Snow, 1995; Kupermintz & Snow, 1997; Nussbaum et al., 1997), as well as differential item functioning (DIF; Lane, 1999). Statistical methods are commonly used in examining the construct.

Care should be taken when examining the dimensionality or internal structure of LSAs. Studies that examine the structure of assessments, specifically achievement, tend to conduct linear factor analyses, or use Classical Test Theory (CTT) measurement approaches (Woods, 2002; Moss & Koziol, 1991), rather than more appropriate measures such as Item Response

Theory (IRT) measurement approaches or nonlinear factor analyses. Since it cannot be assumed that binary data are normally distributed, CTT approaches—such as factor or principal components analyses—are not necessarily appropriate (Woods, 2002). In other words, traditional factor analysis methods are more appropriate for continuous data rather than ordinal (e.g., Likert items) or dichotomous data (e.g., true/false; Woods, 2002). As also discussed by Dunn (2002), Tate (2003) found that there were parametric methods that could be used with binary data, such as tetrachoric correlations. Tetrachoric correlations are used when it is assumed that there are continuous latent responses underlying the dichotomous variables (Muthén & Hofacker, 1988). However, tetrachoric correlations and nonparametric methods testing essential dimensionality are not supported in SPSS, which is one of the most commonly used statistical software programs (Guido & Rains, n. d.). One method, to examine factor structure that was also suggested for nonlinear dichotomous data, was the Normal Ogive Harmonic Analysis Robust Method (NOHARM[1]; Fraser & McDonald, 1988). NOHARM is a program that conducts a nonlinear factor analysis, but takes advantage of the normal ogive model, which is also used with IRT (de Ayala, 2009). This program was also suggested and used by Dunn (2002), indicating the potential significance of conducting a nonlinear factor analysis using NOHARM for binary data.

The fourth element in Messick's construct validity framework is the external element, which involves the examination of the relationship between test scores and other measures (Lane, 1999), as well as the examination of concurrent (i.e., level of agreement between two assessments measuring the same criterion) and predictive validity (i.e., the extent to which the assessment results predict scores on a criterion measure) (Messick 1994). For example, Gong (1999) examined how student scores on the Massachusetts Comprehensive Assessment System related to other indicators of student achievement using correlations. Correlations between two

[1] NOHARM 4 can be downloaded at http://noharm.niagararesearch.ca/

assessments (Lane, 1999), which measure the same or different constructs, tend to be the most commonly used procedures for external validity (Bacci, Koger, Hoffman, & Thacker, 2003). In addition, predictive procedures document how well test scores can predict similar outcomes (e.g., employment test and performance), while discriminant procedures compare test scores to dissimilar outcomes (e.g., self-esteem and mathematical ability).

Generalizability refers to the consistency in student achievement across different populations and settings (Messick, 1995). Researchers can address this element by changing the consequences of the test (e.g., high-stakes versus low-stakes) and examining its impact on score interpretations (DeMars, 2000). Unfortunately, this type of evidence may be difficult to acquire unless sample-specific characteristics are released to the researcher, the test can be re-administered with different consequences of use, or the tasks involved can be re-defined.

The consequential element of Messick's (1995) construct validity framework refers to the intended and unintended results of the assessment (Messick, 1995), including the consideration of the value implications of score meanings and the social consequences related to the assessment (Messick, 1995, 1998). It is not a case that there is or is not consequential validity; instead it is the examination of the intended and unintended consequences of test usage.

Potential queries can include the impact on current teaching and learning in schools, on students, in society, and on future learning (Downing, 2004; Neble & Jaeger, 1983). For example, an assessment which is believed to provide an account of a student's HOT skills, but does not actually assess any of these skills can impact how those results are interpreted to the schools. If a school fares poorly on the assessment, it may be viewed as a school that does not support and promote HOT skills. As a result, the school may get less funding, the teachers may be viewed as inept to teach these skills, and the students may be viewed as deficient in using

these skills. There may also be unintended consequences even if the test measures what it is purported to measure. For example, if a school does well future programs may focus less on these skills, despite being an important aspect to student development.

Although an assessment can be developed with the utmost quality, the interpretation of the results can greatly impact how they are used. The results have the potential to help teachers identify strengths and weaknesses in student learning and in the curriculum, as well as supply information about the individual needs of students. Administrators and policymakers can use the assessment results to examine school policies and make judgements about the quality of the school's performance (Abu-Alhija, 2007). Accordingly, the consequences related to assessment results can be both positive and negative (Abu-Alhija, 2007). Positive consequences can include students being driven to achieve high scores because of the resulting impact on their achievement scores (Volante, 2007), increased funding, and the identification of well-developed programs. Negative consequences relate to the added stress on students, and the abuse by administrators or policy makers on how test results are reported to ensure that funding is not removed (Abu-Alhija, 2007). Teachers teaching to the test can be considered both a negative and positive consequence. For example, it can mean that teachers are only focused on a finite set of skills that are measured by the assessment (negative) or the finite set of skills are accurately reported by the assessment and important for student development (positive). As a result, the proper interpretation of the achievement results is important for ensuring the appropriate decisions are made (e.g., funding) (Abu-Alhija, 2007).

It is important to ensure all elements of reporting are well explained by those involved with the development, administration, and reporting of the test. For example, assessment results presented only as numerical values and in isolation can lead to multiple interpretations (Simnel,

2000). Misinterpretation of the achievement scores may result in the misallocation of funding, the misunderstanding of teachers' and students' capacities, and the false impression that a school may be failing to meet educational standards (Mazzeo, 2001). The public may also view a school negatively based on its results (Mehrens, 1998). The consequences associated with achievement results, no matter the perspective (e.g., product or process) or goals (e.g., selection or improving learning) should, thus, take into account the contextual elements of student achievement as well as other elements of the assessment including the purpose, content, item and test format, and procedures (e.g., reporting) (Mehrens, 1998).

Even though Messick (1995) expressed the importance of examining consequential validity, opposing views have been present (e.g., Popham, 1997). Some researchers find it difficult to actually collect the consequences related to test use (e.g., Reckase, 1998) and some are not sure as to who should be responsible for collecting and reporting this evidence (i.e., test developers or test users; Nichols & Williams, 2009).

Michael T. Kane. Kane (1990, 1992, 2001, 2002, 2004, 2006, & 2009) proposed an argument-based approach to validation. He recommends that an argument be made to defend the use of the test and that different approaches should be applied to support the use and interpretations of the resulting test scores. The structure of the argument is made up of two parts: validity and interpretive. The validity argument refers to the overall analysis of the evidence to support both the uses and intended interpretations. The interpretive argument includes the specific questions or assumptions that will be explored during the validation. Several types of evidence may be required to evaluate some assumptions. Therefore, "[i]n general, then, the interpretive arguments are artefacts, they change with time, they may need to be modified for particular examinees or circumstances, and they are more-or less plausible" (Kane, 1992; p.

533). Although all the assumptions underlying an interpretive argument cannot be proven, the goal is to show that the assumptions are plausible given the evidence collected. The interpretive argument is not absolute and can also be interpreted as a hypothesis (Downing, 2004). Kane's (1990, 1992, 2009) approach to validity relies less on consequences, but he does provide practical guidelines for the validation process (Wolming & Wikström, 2010).

In order to apply Kane's (1992) argument-based approach to the validation, test developers and users must clarify their argument and present the plausibility of the assumptions; after which, they must provide evidence to support the interpretive argument. Therefore, a validator needs to: (1) propose the interpretive argument including the inferences and assumptions; (2) assemble all the available evidence for the interpretive argument, by also including the ones that may be questionable; (3) assess the questionable assumptions in more detail; and (4) change the interpretive argument based on the overall evaluation. The third and fourth steps are iterative until all arguments and assumptions are addressed.

Depending on the goals of the validation (e.g., to examine whether an Algebra test is successful at placing students in a calculus class), different assumptions can be made (e.g., students who lack the skills on the test would have a difficult time in the calculus class), and evidence for those assumptions can be collected (e.g., examine the content and the methods of instruction, compute a correlation between test scores and scores on a subsequent calculus class). Finally, all assumptions can be assessed and the interpretive argument can be re-defined. Overall, the process of assumptions and of gathering evidence is iterative. Thus, "[i]ndividual studies in a validity argument may focus on statistical analysis, content analyses, or relationships to criteria, but the validity argument as a whole requires the integration of different kinds of evidence from different sources" (Kane, 2006; p. 23).

Views concerning validity and the validation process are not consistent among all researchers. For example, Lissitz and Samuelsen (2007) and Borseboom, Cramer, Kievet, Scholten, & Franić (2009) have deviated from the idea of a unitary concept of construct validity. Lissitz and Samuelsen (2007) proposed a method of validity that is more appropriately titled "content validity" and de-emphasized referring to it as a unitary concept. They believe that there are two parts in determining the validity of a test: (1) the test itself, which includes the internal structure; and (2) the testing of other criteria or measures. The internal structure of the test is addressed by examining the practical content, reliability, and the theoretical latent processes. Practical content involves the process of evaluating the match between the test items and the table of specifications. Reliability includes internal consistency analyses, such as confirmatory factor analyses and analyses of variance, to examine the theoretical nomological network. Lissitz and Samuelsen (2007) stress that both internal and external factors are important to the validation process, but the focus should be on the examination of content and reliability measures. They do suggest that validity is inherent in the test, which can be considered a narrow view of validity (Chalhoub-Deville, 2009). Additionally, Borseboom et al. (2009) proposed that validity should not be labelled as construct validity as it has relied heavily on the logical post-positivist philosophical framework. They proposed that, at any point, evidence can be found to support validity. Considering validity as a form of evidence is problematic, as validators will find supportive rather than unsupportive evidence. They believe that construct validity depends on how the researcher interprets validity and, thus, can cause different score interpretations. However, as validity is not an exact science, the validation process can take any shape depending on the goals of the validators. Kane (2008) suggested that it is up to the validator to propose the interpretations and uses of the test scores. As a result, the validator can then take up a narrow

(i.e., label validity as content) or open view (i.e., including other aspects of validity) when conducting a validation.

These examples further denote the significance of clarifying the goal of a validation by a validator and the provision of clear and concise arguments. Researchers can get inundated with the many methods and approaches to validity suggested in the literature. The work of Messick (1993) and Kane (1992) were chosen for the current study because the combination provided a clear and efficient approach to examining the underlying constructs and theoretical underpinnings of the two LSAs.

Summary. Messick (1993) and Kane (1992) have similar ideas as to validity and validation including the importance of using multiple sources of data. While Messick (1993) views validity as a unitary concept, Kane (1992) views validity as the process of supporting arguments with evidence. The argument-based approach to validation is important because it not only documents successes and failures of a test, but it also provides the inherent argument in the test, which can help improve test design and use. Kane's (2009) argument-based approach to validity offers a realistic approach to the process of validation and the resulting arguments are relatively easy to interpret. Elements of Messick's (1993) construct validity framework serve as basis for examining constructs and also provide evidence to the arguments proposed by the validator.

Literature Review Summary

The chapter reviewed the literature on the history of LSAs, the many goals of assessment (e.g., regulate learning), the different assessment perspectives (i.e., product- and process-based), and the link between reading and thinking. The goals of the assessment will ultimately drive the implementation of an assessment model. The significance associated with identifying the models

and theories of assessment and reading lie in the potential impact on achievement results. Differences in the way in which the test is developed and interpreted can have detrimental effects on many aspects of the educational system (e.g., funding). As a result, providing clear goals and connections to what is being assessed and the way in which it is being interpreted is important. The models and theories outlined in the literature review will provide a basis for developing criteria to examine LSAs that evaluate reading.

The literature also revealed the research on thinking and the link between thinking and reading. The most commonly cited thinking skills include lower level thinking, critical thinking, problem solving, and creativity. Although thinking skills are complex, assessment items can be developed to evaluate these skills if they are clearly defined (Haladyna, 1997). With regards to reading, locate strategies are closely linked to lower level thinking, while cycle, integrate and generate strategies are closely linked to HOT. Overall, test developers must provide comprehensive descriptions of the ways in which these types of skills will be assessed, such that there is consistency in the goals of the assessment, the framework implemented, and the item formats used.

The literature review on validity and the validation process indicated that it would be optimal to incorporate elements of Messick's (1995) and Kane's (1992) approaches to validity. This includes the formulation of arguments based on Kane's (1990, 1992) approach to the validation process and specific evidence that can be collected based on Messick's (1995) construct validity framework. Therefore, as described by Sireci (2009), "[i]f the use of this test for the purpose I am using it for were challenged in court, do I have sufficient evidence to persuade the judge or jury and win the case?" (p. 31). Researchers must detail an argument that supports their view, as well as the sources of validity evidence. For the current study, the

research on validity will help provide a framework to examine the underlying assessment perspectives and skills, which are present within LSAs that evaluate reading—a task which has yet to be accomplished.

Chapter Three: Conceptual Framework

The literature review considered different concepts of thinking skills, assessment models, and validation frameworks. The goal of the current chapter is to identify the thinking skills and assessment perspectives, arising from the literature review, which will be used for the current study. Additionally, the importance of epistemology, the influence of personal biases and views, as well as the validation approach devised for the current study, will also be presented.

Research Questions

As mentioned in the introduction, the goal of the current study is to examine the conceptual understanding of thinking and the underlying assessment perspectives of two existing LSAs. The goal of the current study is not to develop any new models or theories or to actually validate new LSAs, but instead to use existing theories and models present within the literature to examine how thinking skills are assessed within the context of two current LSAs that measure reading. Thus, the research questions of interest are: When examining the design and student data of two LSAs in reading, (a) what common thinking skills are assessed? And (b) what are the LSAs' underlying assessment perspectives? The purpose is to examine the measurement of thinking skills in the two LSAs in reading, using a validation approach based on the work of Messick (1995) and Kane (1992, 2001). It was also important to examine the underlying assessment perspective(s) being used by each of the two LSAs, as they also have the potential to impact what skills are being assessed. Therefore, the perspectives of assessment and definitions of thinking skills reported in the literature review will be used to investigate the two chosen LSAs.

Epistemology

Epistemology plays a large role in not only the development of the research questions, but also in the application of a validation approach on existing data, including what evidence is to be collected (Koch & DeLuca, 2012). Validators can take any view of validity; but that view will inevitably influence what evidence is collected (Kane, 2008). The current study is centered on the principles of pragmatism due to its practical approach to research. Pragmatism offers social scientists the chance to incorporate pluralist methods, while attending to the many discourses of the field (Maxcy, 2003). Within the pragmatic approach, the ontological assumption asserts that the practical determines reality, while the epistemological assumption stresses that a practical approach will determine the relationship between the researcher and the participant and the way in which knowledge is viewed (Mertens, 2005). A pragmatist views knowledge as "…being both constructed *and* based on the reality of the world we experience and live in" (Johnson & Onwuegbuzie, 2004, p. 18).

There are several research designs that can be implemented as a mixed methods approach, but the appropriateness of the method depends on the research question of interest (Creswell & Plano-Clark, 2007). For instance, sequential or parallel designs allow the researcher to choose whether both approaches will be completed at the same time or one after the other (Creswell, Plano Clark & Garrett, 2007). However, using both quantitative and qualitative approaches in one research design can be demanding and time consuming (Creswell & Plano-Clark, 2007). Overall, this perspective supports the understanding of a phenomenon from both qualitative and quantitative perspectives (Morgan, 2007).

Pragmatism, as an epistemological stance, provides an ideal approach for answering the current study's research questions, because of its inclusion of multiple sources of data and

perspectives as well as its mixed method approach. To address both thinking skills and underlying assessment perspectives within the assessment, both the content and structural makeup of the assessment must be examined, such that qualitative and quantitative perspectives are included. Additionally, the qualitative perspective will serve as a guide for the quantitative perspective.

Personal Views and Assumptions

In conducting the present study, a number of key assumptions were made based on personal views and biases, as follows:

a) In addition to assessing subject matter content, LSAs typically target thinking skills. Some thinking skills overlap or are common across LSAs. Sometimes these thinking skills may be presented and labelled as subject-matter-related skills.

b) Higher order thinking skills are seen as equal as opposed to having a hierarchical relationship among them (e.g., considering problem solving skills as a basis for critical thinking skills). Additionally, thinking skills are grouped into two categories: Lower and HOT.

c) The large scale assessment of thinking skills is affected by the adopted (consciously or not) underlying assessment model guiding the development, administration, and score interpretation processes.

d) Although difficult to fit theories, models, and perspectives into distinct categories, it is believed that the assessment perspectives can be situated on either ends of a continuum.

e) The review of the literature produced no clear and concrete relevant theories, models or perspectives related to the issues of thinking skills, assessment, and reading in the

context of LSAs. Thus, the proposal of respective taxonomies for the purpose of the study is required.

f) A LSA may tap onto more than one assessment model, but it is likely that one model will predominate in each case.

g) A validation process inspired by aspects of Messick's (1993) and Kane's (1990) work can serve as a way to systematically examine key LSA documents and assessment results.

Given the pragmatic stance taken in the current study, despite the large presence of qualitative analyses, the thesis is written in a style largely influenced by the post-positivist view.

Thinking, Assessment, and Reading in the Context of LSAs of Reading

Based on the literature review, it was established that there are many definitions and terms associated with thinking skills. The current study was able to offer a list of thinking skills that are potentially assessed by LSAs, including lower level thinking skills, critical thinking, problem solving, and creativity. Table 2 provides the definitions and key information to be used as the basis for the current study, specifically with regards the coding of the assessment documents. As mentioned previously, lower level thinking skills are described as the process of defining, selecting, and generating the concepts, principles, facts and procedures from previously exposed information. For example, students may be asked to restate information from a reading selection or be asked about the definition of a word. Critical thinking is classified as the process of observing, comparing, evaluating, and making a judgement. For example, students might be asked to determine the outcome of a scenario based on the information that is given. Students may also be asked why the main character made certain decisions. Problem solving is the process of identifying, defining, and analyzing a given complex issue, and proposing solutions. Most

Table 2

Elements of thinking skills

Thinking Skill	Definition	Determining if an item is assessing this skill.
Lower level thinking	Lower level skills are needed to organize, understand, develop meaning, and conceptualize information.	• The information is explicitly stated in the text. • The student has likely been previously exposed to the information.
Critical thinking	Critical Thinking is a process from observing, comparing, evaluating and making a judgement.	• The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response. • For example, the student may be asked to determine why a character in the story made a specific decision, why the story took a certain direction, how elements of the story can be used in other scenarios etc. • Items requiring critical thinking skills would enable students to make a judgment about elements in the text.

Table 2

Continued

Thinking Skill	Definition	Determining if an item is assessing this skill.
Problem Solving	Problem solving is a process from recognising, identifying, analyzing and proposing a solution.	• The student is given an issue that must be resolved. • For example, a student can be asked to make a specific decision about elements in the text, what other ways the character in the text could have resolved an issue, etc. • Items requiring problem solving skills would enable students to make decisions based on specific elements in the text.
Creativity	Creativity involves flexibility, discovery, and innovation.	• The student is asked to provide a novel response, such as their own innovative ideas that are independent from what is presented in the text. • For example, a student could be asked their unique personal response to elements in the text, to reflect on their own personal experience, etc. • Items requiring creativity would enable students to use their own ideas to reflect on the information provided.

importantly, some test developers of LSAs in reading refer to problem definition and solution in their assessment frameworks (e.g., PIRLS, 2011; Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009; Councils of Ministers of Education, Canada [CMEC], 2007). Students may be asked to determine the best solution to the problem facing the main character. Although creativity does not seem to be as commonly cited in the literature review as the other skills, it still is an important skill on which students should be assessed in a large-scale format (Bracey, 2001). For example, students may be asked to provide a novel response or to "think outside the box". Within the Canadian context, some of these skills are apparent in the EQAO (EQAO, 2007) and PCAP (CMEC, 2007), in which they identify the use of lower level thinking and critical thinking skills.

Despite the intended aim given to each individual item, in terms of the targeted thinking skills, questions can still tap into more than one skill. For example, a student may have been previously exposed to the information that is asked for in a question and would not need the text to determine the response. Thus, such a question would call for a lower level thinking (e.g., the meaning of a word). On the other hand, if a student had not been previously exposed to the information, he or she would require the information presented in the text to determine the response, thus applying HOT. It is likely that questions that tap into both lower level and HOT skills concern information not unique to the reading text, such as questions regarding word meanings and the use of language rules.

Overall, common thinking skills cited within the literature include lower level thinking, critical thinking, problem solving, and creativity, all of which can be assessed within a large-scale format (Bracey, 2001). Thinking is cross-curricular and occurs whether a student is completing mathematical problems or reading a text. Due to the complex nature of thinking,

some questions may tap into more than one thinking skill. As the goal of the current study is to examine the assessment of common thinking skills, it is important to examine these four skills in the context of an LSA in reading.

The focus of the current study is the examination of common thinking skills in two LSAs that assess reading. As reading is the content measured, it is important to identify the match between different reading strategies and thinking skills, particularly in LSA contexts. As identified and summarized in the literature review, the work of Mosenthal and Kirsch (Mosenthal, 1996; Mosenthal & Kirsch, 1991a) has typically been used for LSA development (OECD, 2013). As proposed, there are four types of document processing strategies applied by individuals for reading to do. Table 3 presents the different document processing strategies and their relationship to the common thinking skills. Locate strategies are similar to lower level thinking skills as they relate to information that is explicitly stated in the reading selection. Cycle strategies have elements of both lower level thinking and HOT. Assessment questions that are developed to tap into cycle strategies have the reader finding more than one piece of explicitly stated information in the text. There are both independent and dependent strategies. For independent cycle strategies, a question may ask "How many wolves and dogs did the park ranger see?" A student would be able to figure out how many dogs there were without having to know how many wolves the park ranger saw. On the other hand, dependent cycle strategies refer to an assessment questions for which the different elements of the questions are dependent on each other. For example, with regards to information that lists class schedules for several students, a question may ask "If Olivia and Liam decided to work on their math work at lunch together, which day are they both free?" Students would be required to examine the student list to find Olivia and Liam and then see what day of the week both of them are free at lunch. A

Table 3

Document Processing Strategies

Document Processing Strategy	Aspects	Match to Thinking Skills
Locate	• Retrieve explicitly stated information. • Match one or more elements from a question to the information in the reading passage.	• Lower level thinking skills
Cycle	• Retrieve explicitly stated information in more than one part of the text. • Complete more than one locate strategy.	• Lower level thinking skills. • Potential HOT skills based on dependent and independent cycle strategies.
Integrate	• Compare and contrast numerous cycle strategies.	• Problem solving and critical thinking.
Generate	• Use prior knowledge to answer the question	• Problem solving and critical thinking.

student may need to use HOT skills, as this would require them make a decision based on the explicitly stated information. Integrate strategies would require students to compare and contrast information in the text to produce new information. They may need to propose a solution or make a judgement, thus, involving HOT skills. Finally, generate strategies require students to use previous knowledge to expand on the information in the text with the ability; again, with the use of HOT skills.

Based on the literature review, two overarching assessment perspectives have been proposed: (1) Product-based; and (2) Process-based perspectives (Hager & Butler, 1996; Serafini, 2001). Implementing a purely product-based perspective would mean that meaning resides in the text, the goal would be to potentially diagnose learning, scoring would be rigid, and the results would be used to compare other students or schools. Conversely, implementing a purely process-based perspective would mean that learning would be viewed as a process involving students and the use of an open-ended type of scoring.

It is possible for a LSA to incorporate elements of each perspective. Awareness of what elements underlie a LSA is important for determining the goals and skills to be assessed. Therefore, listing specific criteria associated with each perspective is critical. Table 4 provides an overview of the aspects of each of the two perspectives, which will be used for the coding of the two LSAs. The criteria displayed in the table will also be used for the analyses of the current study to determine the elements of each perspective present within the LSAs. The criteria comparing each perspective include the associated goals, object of assessment, assumptions, item format, mode of scoring, and scoring interpretation. The associated goals refer to the overall purpose of developing the assessment; for example, regulating learning is a goal for the product-

Table 4

Assessment Perspectives

Perspective	Goals	Object of Assessment	Assumptions	Item Format	Mode of Scoring	Scoring Interpretation
Product-based	Regulate learning (e.g., assess how programs are performing) Diagnose (e.g., assess students' level of ability)	To measure the product of achievement, such as the level of knowledge acquired over time.	Meaning resides in the text	Multiple choice and short open-ended questions (selected-response items).	Fixed response	Students are compared to some type of standard or other schools and students.

Table 4

Continued

Model	Goals	Object of Assessment	Assumptions	Item Format	Mode of Scoring	Scoring Interpretation
Process-based	Improve learning (e.g., student level learning)	To understand the learning process.	Knowledge is viewed as being constructed. Students are active creators of knowledge. Assessment is part of the learning process.	Multiple choice and open-ended questions including portfolios, demonstrations, and unique individual responses.	Partial scoring (i.e., need some criteria to help score the responses)	Assessment results are used as part of the learning process.

based perspective. The object of assessment directly relates to what is being measured, such as the measurement of the learning process. Various assumptions are associated with each perspective. For example, one assumption of the process-based perspective is that students are active creators of knowledge. That is, students are part of the learning process and are actively involved in what they learn. The item format, mode of scoring, and the scoring interpretation depend on the underlying assumptions of the model, which; as a result, will impact how learning is evaluated. The process-based perspective relies heavily on open-ended formats, while the product-based perspective relies heavily on multiple choice type questions. There is a place for each perspective or elements; however, the implementation of either perspective depends on the goals of the assessment.

Validation

The literature review clearly showed that the concept of validity has been thoroughly discussed and many different interpretations of validity have been suggested. Yet, basic assumptions hold. Validity is not a property of a test, a property of inferences, or the scores of the instrument. Instead, it refers to the inferences about individuals from the interpretation of the scores and the use of these scores (Kelly O'Malley, Kallen, & Ford, 2005).

Given the numerous researchers who recommend the use of Messick's (1995) validity framework (Kane, 1990, 2009; Miller & Linn, 2000; Zumbo, 2009) and Kane's argument approach to validity (Wolming & Wikström, 2010), both frameworks are important to validation. Kane's (1990) approach to validity allows researchers to identify the arguments and support them with evidence. As a result, researchers can develop research questions rooted within a validity argument and provide evidence to support their claims. The evidence can be in the form of the elements of Messick's (1995) construct validity framework. As such, the current study

adopts the argument approach to validity (Kane, 1990) and incorporates some elements of Messick's (1995) validity framework. This is accomplished by detailing several interpretive arguments and assumptions, based on the work of Kane (1990), and using supporting evidence from Messick (1995). Content, substantive, and internal elements of Messick's validity framework are examined for the current study. However, the external, generalizability, and consequential elements are not examined, first, because they do not directly address the research questions and, second, because they do not easily lend themselves to secondary data analysis. With regards to generalizability, a researcher would have to examine student achievement across different settings, which is difficult when using secondary data. An examination of the external element could have been possible with access to item-level data for both assessments, but as will be explained later, this was not the case. Finally, consequential elements were not addressed because it would require collecting new data, which went beyond the scope of this study.

The proposed interpretive arguments are presented in Table 5. These arguments show that, in any LSA: (1) students are assessed on thinking skills and student results reflect this; and (2) There is an underlying perspective framing the development and scoring of the assessment. The first argument posits that a LSA with the goal to evaluate thinking will be explicit in defining not only the skills in the assessment framework, but also the way in which the skills connect to the items themselves. The second argument refers to the presence of underlying assessment perspectives that impact the types of and the ways in which skills are assessed. If the assessment framework claims to measure the development of explicit skills, then it should be clearly stated as to how the skills will be assessed. Each argument is associated with explicit assumptions and means to establish evidence. These means include content, substantive, and internal/structural elements of Messick's (1989, 1995) construct validity framework.

Table 5

Interpretive Arguments and Means of Evidence for the Current Study

Interpretive Arguments	Assumptions	Means of Evidence (based on Messick's (1995) construct validity framework)
Students are assessed on thinking skills and student results reflect this.	The assessment is multi-dimensional, such that there is the presence of more than one thinking skill.	A nonlinear factor analysis (achievement data tends to be binary; right/wrong) examining the factors that make up the assessment.
	Thinking skills are explicitly detailed in the assessment framework and in the items themselves.	Content analysis using a thinking skills coding scheme
There is an underlying perspective framing the development and scoring of the assessment.	The assessment framework, scoring guide, and items take into account an assessment perspective that is aligned with the assessment of thinking skills.	Content analysis using an assessment perspective coding scheme

Conceptual Framework Summary

The conceptual framework brought out vital aspects of assessment perspectives, thinking skills, and the validation approach in order to bring them all together. The validation approach was based on pragmatism and the research questions were developed to reflect this. As such, Kane's (1990) argument approach to validity framed this approach, with evidence provided in the context of Messick's (1995) construct validity framework. With regards to the theoretical and conceptual models, assessment perspectives included those derived for product and for process. The area of thinking has included diverse interpretations, as well as definitions (Lewis & Smith, 1993). The most commonly cited thinking skills included lower level thinking, critical thinking, problem solving and creativity, the latter three of which are considered HOT skills. That being said, the research questions concern the examination of thinking skills in two LSAs that assess reading, as well as the underlying assessment perspectives used to measure these thinking skills. The tables presented in the conceptual framework on the different thinking skills, document processing strategies, and assessment perspectives will serve as the coding protocol for the analyses described in the methodology chapter.

Chapter Four: Methodology

The conceptual framework presented the research questions and reviewed the various perspectives and models adopted for the current study. The goal of the following chapter is to review the two assessments used for the current study. The chapter also details the procedure followed as well as the qualitative and quantitative analyses applied to the secondary data. In order to address the research questions, an exploratory (Creswell, 2003, 2007; Creswell & Plano-Clark, 2007; Creswell et al., 2008) and sequential mixed-method design (Johnson & Onwuegbuzie, 2004; Johnson, Onwuegbuzie, & Turner, 2007; Maxcy, 2003; Morgan, 2007) was conducted. A sequential research design is a method in which results from the first step help inform what is to be accomplished in the second step and so forth. Appendix A provides the full ethics approval for the current study.

LSAs Contexts and the Nature of Data Collected

In Canada, each province is responsible for their educational programs (Taylor & Tubianosa, 2001), unlike some other countries that have national curriculums, such as the United Kingdom (Department for Education, 2012). As the current study focuses on assessments in the Canadian context, in which there are both provincial LSAs and a national Canadian LSA with different aims, both were reviewed.

The two assessments of interest in this study included the PCAP, a national test, and EQAO, a provincial assessment. Both LSAs assess reading and are used to compare achievement levels across schools (EQAO) or provinces (PCAP). They also have an impact on teaching and learning by directly or indirectly influencing assessment practices in the classroom (Gustafsson & Rosén, 2005) and policy. Although the results from the two assessments cannot be generalized to other assessments, there are some parallels to other LSAs. For example, like PCAP, PISA is

not a high-stakes test and the assessment is derived from a general design in order to include diverse provinces, states, and countries. On the other hand, EQAO is similar to other provincial assessments across Canada (e.g., New Brunswick provincial assessment; Taylor & Tubianosa, 2001), such that all students are required to complete the assessment as part of their educational program. Provincial assessments are based on the curriculum derived from that province. In addition, both LSAs assess reading; they both feature thinking skills; they both make a great deal of their assessment public via their website; and they both claim that their assessments yield valid and reliable data. EQAO data are relatively easy to obtain online and administrators of PCAP are easy to contact. As a result, these assessments were chosen because of their ease of access, their parallels to other assessments, and because EQAO was a provincial assessment based on a provincial curriculum and PCAP was a national assessment based on curriculums from each of the provinces. Key components of each LSA are presented in Table 6 and are further explained in the following sections.

EQAO. The first assessment examined in the current study was the 2006-2007 junior division (Grades 4-6) EQAO assessment in reading, writing, and mathematics. It evaluates reading expectations based on the Ontario curriculum. All students in the province must participate. The junior division EQAO assessment is administered at the end of the school year (i.e., end of May, early June) (EQAO, 2012b); is developed, reviewed, and scored by Ontario educators; and is field tested with Ontario students (EQAO, 2010). The goal of the assessment is to provide a comprehensive and reliable snapshot of student achievement. The results are used to indicate how students are meeting the provincial standard; to identify strengths and areas of

Table 6

Components of EQAO LSA and PCAP on Reading

Assessment	Reading Definition	Stated Skills to be Assessed
EQAO	"Reading is defined as the process of actively making meaning across a variety of fiction and non-fiction written texts that students are expected to understand based on the expectations in *The Ontario Curriculum* across all subjects up to the end of Grade 3 and Grade 6." (EQAO, 2007)	(1) "Understanding explicitly stated information and ideas" * (2) "Understanding implicitly stated information and ideas…" * (3) "Responding to reading by making connections between information and ideas in a reading selection and the reader's personal knowledge and experience (interpreting a reading selection by integrating its information and ideas with personal knowledge and experience)" *
PCAP	"Reading is defined as a dynamic, interactive process whereby the reader constructs meaning from texts. The process of reading effectively involves the interaction of reader, text, purpose,	"Comprehension: Readers construct meaning using information provided explicitly and implicitly by the text." (CMEC, 2007; p. 4)

Table 6

Continued

Assessment	Reading Definition	Stated Skills to be Assessed
	and context before, during, and after reading." (CMEC, 2007; p. 2)	"Interpretation: Readers examine the text to develop an understanding of the relationships of discrete elements to the whole" (CMEC, 2007; p. 4). "Personal Response: Readers step out of the text to reflect on their own experience, identify themselves with aspects of the text, establish connections with their own experience, and/or react emotionally to the text" (CMEC, 2007; p. 4) "Critical Response: Readers stand apart from the text, considering it objectively and/or evaluating its quality and/or appropriateness to the world at large." (CMEC, 2007; p. 4)

* (EQAO, 2007; p. 11)

improvement; to address where resources are needed; to maintain ideal teaching practices; and to report the results and discuss with parents (EQAO, 2010). On the EQAO website, there are a multitude of resources for parents, students, and educators. These include scoring guides, sample questions, assessment frameworks, and technical reports. Furthermore, the test developers of the EQAO distinguish between LSAs and classroom assessment to ensure there is no confusion as to their respective goals, and highlight similarities between the EQAO with other assessment programs such as PCAP and PIRLS. The goal is to situate the EQAO assessments and to clarify how the results are used and interpreted (EQAO, 2007).

The junior division EQAO assessment includes three booklets; one on mathematics, one on reading, and one on writing. As displayed in Table 6, the test developers of EQAO consider reading as a process of meaning making, in which students are required to demonstrate three reading skills: (1) understand explicitly stated ideas; (2) understand implicitly stated ideas; and (3) respond to reading (EQAO, 2007). These three skills can also be considered thinking skills (Haladyna, 1997), although they are described very broadly. For example, the first skill could be considered a lower level thinking skill and a locate strategy because students will have previously been exposed to the material. Similarly, the second and third skill could be related to HOT skills including critical thinking, as well as cycle, integrate, and generate strategies.

There are three booklets (i.e., mathematics, reading, and writing) and two sections to each test booklet, each of which are suggested to take one hour. However, the assessment itself is considered untimed. The reading selections for the two language booklets, which contain the reading and writing portions, include narrative text, non-narrative informational text, poetry, and graphic text. The reading selections are either short, ranging from 300 to 350 words, or long, ranging from 650 to 700 words (EQAO, 2012b). The first booklet contained 12 questions

dedicated to reading, with four questions defined as open-ended and eight multiple choice questions. The second booklet contained 23 questions dedicated to reading, with six open-ended and 17 multiple choice items. The two language booklets were scored using a rubric and students' responses are coded based on eight categories: (1) "Blank", when nothing is written; (2) "Illegible", when the content cannot be read; (3) "Irrelevant", when the question is not addressed; (4) "Off topic", when there is no relationship between what is written and the question; (5) "10", when the response is showing misunderstanding; (6) "20", if the response only addresses part of the question; (7) "30", when the response addresses the complete question, but only some specific information is used; and (8) "40", when the response addresses the complete question and provides accurate information (See Appendix B for more information on the rubric levels).

The writing and mathematics assessment are coded on the same scale, but justifications for each code differ. Raters are given training, as well as examples, in order to help them implement the rubric in a reliable manner. Once all the assessments are scored, they are compiled to determine the student's level of achievement. The Ontario Ministry of Education offers four achievement levels (i.e., 1-4), where level 3 is the provincial standard, level 1 identifies students below the standard, level 2 means that students approach the standard, and level 4 identifies students who are above the standard (EQAO, 2007).

PCAP. The second LSA examined for the current study was the 2007 PCAP on reading. The 2007 PCAP assesses the achievement of 13-year-old students across jurisdictions on three sub-domains of reading: comprehension, interpretation, and response to text. Participants are randomly selected across Canada and are given 90 minutes to complete the assessment booklet,

which contains 49 items. Once this booklet is completed, students are given 30 minutes to complete a student background questionnaire.

There are two forms of the 2007 PCAP assessment, both of which contain 38 multiple choice and 11 open-ended items. Each assessment contains six sections, with the first three sections being allocated 20 minutes each and the last three sections being allocated 10 minutes each. Each assessment has six reading passages which include some visual and graphic elements. As part of the open-ended items, PCAP also includes extended open-ended items that measure students' performance on a problem requiring a complex solution. As described by PCAP, students are assessed on different aspects of reading literacy including *comprehension*, in which readers construct meaning through explicit and implicit information; *interpretation*, in which students use the text to develop understanding; *personal response*, in which students reflect on their own experience; and *critical response*, in which students consider themselves apart from the text and evaluate it objectively. These skills are closely related to thinking skills. Comprehensions is defined as containing elements of lower level thinking skills and locate and cycle strategies, while interpretation, critical response, and personal response include aspects related to critical thinking and integrate and generate strategies. These four literacy skills, which can be considered as thinking skills performed in reading, are each coded based on three performance-level descriptors: (1) Level 1, in which the students demonstrates partial understanding, limited interpretation, limited personal response, or simplistic critical response; (2) Level 2, in which the student demonstrates a clear understanding, reasonable interpretation, appropriate personal response, or appropriate critical response; and (3) Level 3, in which the student demonstrates a nuanced understanding, insightful interpretation, elaborated personal response, or significant critical response (See Appendix C for more information on the rubric

levels). Finally, raters are given training in applying the levels to the open-ended responses, as well as example responses.

Procedure

The following procedure was used in the current study: (1) Obtain the assessment framework, test booklet, and scoring guide documents from PCAP and EQAO; (2) Obtain the item-level data from EQAO (PCAP item-level data could not be obtained); (3) Perform a content analysis of the assessment framework, test booklet, and scoring guide documents of both EQAO and PCAP using the codes developed in the conceptual framework (i.e., assessment perspectives and thinking skills); (3) Analyze the EQAO item-level data and compare the results to the content analysis for the EQAO test booklets; and (4) Report results.

The test booklets, assessment framework, and scoring rubrics were all available on the EQAO website; however, an application for access to item-level data for researchers on the EQAO website had to be submitted. The item-level data were sent in a Microsoft Excel format and included both Grade 3 and Grade 6 responses to both the reading and writing questions for 2007. In addition, both French and English versions were included. PCAP does not provide any detailed information online and thus an application for access was sent to retrieve the assessment framework, scoring guide, and item-level data[2]. Due to confidentiality reasons, only the assessment framework and scoring guide could be obtained. Item-level data are not provided by PCAP, but PCAP enables researchers to examine the test booklet questions in-house at the Toronto headquarters. Because the PCAP test booklets were not accessible to anyone other than the primary researcher, the results are only discussed in a general manner. That is, only overall results are presented. In sum, all the necessary data were obtained from EQAO to conduct both

[2] As of 2013, the PCAP 2007 assessment framework is available online http://www.cmec.ca/docs/pcap/pcap2007/Reading-Framework-EN-Apri2013.pdf

qualitative and quantitative analyses, but only the assessment framework and scoring guide were retrieved from PCAP to conduct the qualitative analyses. The PCAP test booklets were analyzed, but only a general appraisal could be reported.

Once the data were retrieved, analyses were conducted based on Messick's (1995) and Kane's (1990) validity frameworks. Qualitative analyses (i.e., document analysis) of both PCAP and EQAO documents were conducted first, followed by a quantitative analysis of the EQAO item-level data. The quantitative analysis was focused on the examination of thinking skills, for which the qualitative analysis of the thinking skills presented in the test booklets of EQAO, served as a guide.

Qualitative Analysis

The purpose of the qualitative analysis was to identify references to thinking skills in the assessment framework, test booklet, and scoring guide, using codes based on those concepts described in the conceptual framework. Qualitative analysis also served to examine the underlying assessment perspectives in the assessment framework, test booklets, and scoring guide, through codes also based on the two perspectives described in the conceptual framework. The following section reviews the coding design, the content analysis, coding protocol, and the use of an external rater.

As the coding for the analyses was developed based on previous theories, the current study implemented a directed coding design, in which codes are defined a priori using existing theories and research (Hseih & Shannon, 2005). Although this type of content analysis is pertinent to the current study, there are limitations due to potential bias (Hsieh & Shannon, 2005). For example, a researcher may unconsciously find supportive evidence rather than balance both supportive and unsupportive evidence within the transcripts. In order to combat bias

due to supportive versus unsupportive evidence, it has been suggested to use another rater (Hsieh & Shannon, 2005). The inclusion of a second rater ensures that codes developed from the conceptual framework are easily understood and can be applied consistently by another individual outside of the study who has no previous biases. Therefore, for the current study, the primary researcher was the first rater and an external rater was included as the second rater.

The content analysis was conducted using the steps from Zhang and Wildemuth (2009). First, it was determined that the data should be coded in sections in order to accurately pick out themes; second, the coding scheme was developed; and third, the coding was tested on a sample of text in order to ensure coding consistency. The coding scheme was taken from Tables 1, 2, and 4, displayed in conceptual framework. For the assessment frameworks, relevant data from each paragraph was coded for the referenced thinking skills and perspectives. Each item in the test booklets and each level of the scoring guides were coded for the referenced thinking skills. The scoring guide and test booklets were assessed as a whole for the different perspectives because it enabled the researcher to examine the documents all together instead of focusing on small sections of data (i.e., one sentence).

The coding protocol was created using Microsoft Excel. There were three tabs for each LSA: one focusing on the assessment framework, one for the test booklets, and another one for the scoring guide. While each tab was dedicated to each document, the coding for thinking skills and assessment perspectives were included in each tab. Appendix D provides the actual excel tabs used for analysis. The procedure and steps involved in coding the assessment framework and scoring guide are seen in Table 7 and the steps followed for coding the test booklets are found in Table 8. Each paragraph for the assessment framework was analyzed by documenting

Table 7

Coding Protocol: Assessment Framework and Scoring Guide

Coding	Coding Choices	Coding Steps
Assessment perspectives	• Product-based; and • Process-based.	1. Identify the assessment perspective(s), which can be explicitly stated or inferred. More than one perspective can be identified; 2. Provide the evidence from the document; 3. Provide explanations as to why the evidence supports the specific assessment perspective; 4. Provide the page number; and 5. Provide any additional comments.
Thinking Skills	• Lower Level Thinking; • Critical thinking; • Problem solving; and • Creativity.	1. Identify the thinking skill(s); 2. Provide the evidence from the document; 3. Provide an explanation as to why the evidence supports the specific thinking skills; 4. Provide the page number; and 5. Provide any additional comments with regards to other potential skills.

Table 8

Coding Protocol: Test Booklet

Coding	Coding Choices	Coding Steps
Assessment perspectives	• Product-based; and • Process-based	1. Conduct a general examination of the items, taking them all together to identify potential assessment perspective(s).
Thinking skills	• Lower level thinking; • Critical thinking; • Problem solving; and • Creativity.	For each item : 1. Identify the thinking skill(s); 2. Provide the evidence from the document; 3. Provide an explanation as to why the evidence supports the specific thinking skills; 4. Provide the page number; and 5. Provide any additional comments with regards to other potential skills.

the page number, thinking skill(s), perspective(s), and evidence. Each item of the test booklets was analyzed by documenting the test booklet, item, page number, thinking skill(s) and evidence, as well as potential perspectives that take into account the test booklets as a whole. Each level of the scoring guide was analyzed by documenting the thinking skill(s) and evidence, as well as potential perspectives that take into account the scoring guide as a whole. More than one thinking skill or assessment perspective could possibly be referenced within each document. In order to determine the satisfactory underlying incorporation of assessment perspectives and thinking skills, key information from the codes presented in the conceptual framework must be present (Mayring, 2000).

External rater. A training manual was developed in order to provide consistent material for the external raters and also to achieve a high level of inter-rater reliability. A section of the manual was dedicated to the coding protocol, which included definitions of the different assessment perspectives and thinking skills. The inter-rater reliability analysis involved coding the two EQAO test booklets. A total of three external raters were used. The first external rater coded the first booklet (i.e., 12 items) for thinking skills and changes were made to the codes based on their feedback (i.e., clarity); after which, the second external rater used the new codes to examine the two EQAO booklets (i.e., 35 items) for thinking skills. Changes were again made to the codes and then the third external rater coded the two EQAO test booklets for thinking skills, as well as provided an appraisal of underlying assessment perspectives of all three documents. While the goal of using the first two external raters was to help clarify the codes used for the analysis, the third external rater was used to help not only further clarify the codes, but also to help determine which codes best represented the items. The first two external raters had backgrounds in psychology, while the third external rater had a background in economics. The

third external rater's background in economics was important as it is entirely different from the primary researcher, as well as the first and second raters, which in turn gives weight to the applicability of the codes. Overall, if discrepancies were found, they were discussed to ensure the code best represented the item. After the final inter rater reliability analysis, one set of codes was retained.

The inter-rater reliability analysis was conducted to identify the percentage of agreement between the external rater and the primary researcher. Cohen's Kappa was calculated in order to account for agreement based on chance (Cohen, 1960). Cohen's Kappa is calculated by $K = P_A - P_c/1 - P_c$, in which P_A is defined as the proportion of codes on which the raters agree and P_c is the proportion of agreement that is due to chance. The procedure followed the steps provided by Stemler (2001). Cohen (1960) suggests the following Kappa's for different strengths of agreement; (1) <0.00 = poor; (2) 0.00 - 0.20 = slight; (3) 0.21 - 0.40 = fair; (4) 0.41 - 0.60 = moderate; (5) 0.61 - 0.80 = substantial; and (6) 0.81 - 1.00 = almost perfect.

Quantitative Analysis

Once the content analysis was completed, a nonlinear factor analysis was conducted, followed by a reliability analysis (i.e., Cronbach's alpha; Cronbach, 1951) on the EQAO item-level data. The goal of the quantitative analysis was to determine if there were factors aligned with the different thinking skills identified in the qualitative analysis. The following section reviews the steps taken for the quantitative analysis.

Several types of factor analyses are available, including the exploratory linear factor analysis, the confirmatory linear factor analysis, the limited-information nonlinear factor analysis, the full-information nonlinear factor analysis, and essential dimensionality (Dunn, 2002). However, Principle Components Analysis (PCA) is perhaps the most commonly used

CTT approach for data reduction and is considered a good choice when deciding to conduct an exploratory factor analysis (Joliffe & Morgan, 1992). Although, given that both PCAP and EQAO include binary items (i.e., right and wrong), PCA may not be an appropriate choice in this particular context, as it was developed for interval or ratio-type data and may thus inflate values for ordinal variables (Candel, 2001; Linting, Meulman, Groenen, & van der Kooij, 2006; Dunn, 2002). PCA examines the intercorrelations among the variables (Kim & Mueller, 1978) to produce the factors (Dunn, 2002) and most researchers conduct PCA with ordinal data (Linting et al., 2006). On the other hand, nonlinear PCA takes into consideration the possibility of curvilinear factors (de Ayala, 2009; Dunn, 2002) and conducts the factor analysis using dichotomous or binary data (i.e., 0, 1), and the normal ogive model (de Ayala, 2009). Nonlinear PCA analyzes categorical data by taking into account both nominal and ordinal-type variables. For that reason, a nonlinear PCA was conducted (Meulman, van der Kooij, & Heiser, 2004). The factors extracted were based on the number of thinking skills that were found in the literature review and in the conceptual framework. Thus, 4, 3, 2, and 1 factors were examined using the nonlinear PCA.

The quantitative analysis examined the structural element of Messick's (1989, 1995) validity framework, and included a dimensionality analysis. One method that was suggested in the literature was the NOHARM program[3] (Fraser & McDonald, 1988). An alternative to using NOHARM on binary data is TESTFACT (Wood et al., 2004); however, as NOHARM can be used at no cost and has been shown to perform well in dimensionality studies (De Champlain & Gessaroli, 1998), it was chosen for the current study. Although the program is suitable for dichotomous data, there are some cautions. First, the program can only examine dichotomous data and secondly, it does not take into account missing data.

[3] NOHARM 4 can be downloaded at http://noharm.niagararesearch.ca/

The NOHARM program is relatively easy to use, is free to download, and includes a user manual. In order to run the program, the job criteria must be imputed followed by the binary data for all the items. The following data were inputted into a notepad, which creates text files using basic text formatting, and then uploaded in the NOHARM program: (1) The number of items (N); (2) The number of dimensions (M; i.e., factors); (3) The number of subjects or observations (NS); (4) The type of input data (IN; e.g., raw dichotomous scores coded as 0 or 1); (5) Whether it is an exploratory analysis (EX); (6) The generation of starting values (IV); (7) The printing/non printing of the raw product-moment matrix (PS); and (8) The printing/non printing of the residual matrix (i.e., for fit; PR). Thus, the notepad document is formatted so that the first line contains the title, the second line contains the job criteria, and the third line contains the guessing parameters (e.g., zero chance of guessing for 30 questions would include 30 zeroes). Once the file is uploaded and all criteria are met, the program can be run. The output for the nonlinear PCA included information such as the factor loadings and the correlations (see Fraser & McDonald, 1988).

Fit statistics are provided by the NOHARM program. To interpret model fit, a residual matrix is produced, which is the difference between the observed covariances and those after the model is fitted to the data. Thus, a situation where the differences are zero is ideal. NOHARM produces the root mean square (RMS) and the goodness-of-fit (GFI) index to examine model of fit. The RMS can be compared to the standard error of the residuals, where a smaller RMS indicates a good fit (de Ayala, 2009). A GFI close to 1 indicates a good model fit. In order to determine the amount of factors that best represents the data, researchers can compare the goodness-of-fit scores for every factor solution (e.g., compare 2 to 3 factors). When there is a sharp increase in fit, and then the fit scores stabilize (like a scree plot), this would indicate the

best model. Like a factor analysis, the most simple and easiest solution to understand is best (de Ayala, 2009).

Like a linear factor analysis, there is the choice to use a varimax or promax rotation. For the current study, eigenvalues above 0.30 were included (Gorsuch, 1983). To examine reliability of the factors, a reliability analysis using Cronbach's alpha was conducted (Cronbach, 1951). It has been suggested that there are more appropriate forms of reliability analysis for data that are not continuous (e.g., ordinal; Gadermann, Guhn, & Zumbo, 2012). However, it has also been suggested that conducting a reliability analysis, using dichotomous data, is similar when using either the Kuder-Richardson 20 (KR-20; Kuder & Richardson, 1937) or Cronbach's alpha (Cronbach, 1951; SAS, 2012). Acceptable reliability was defined as 0.70 (Nunnally & Bernstein, 1994).

Methodology Summary

Overall, the current study is focused on a validation approach based on the work of both Messick (1989, 1995) and Kane (1990, 1992). The two assessments of interest were the PCAP and EQAO and included the assessment framework, test booklet, and scoring guide documents, as well as item-level data from EQAO. The coding, as presented in the conceptual framework, was used both for the primary researcher and the external rater for the content analysis. Both qualitative and quantitative analyses were conducted to address the research question in conjunction with the validity arguments presented in the conceptual framework. How the research questions are addressed, by the nature of data collected, are presented in Table 9. Elements not addressed due to limitations in conducting a secondary data analysis and issues relating to confidentiality included the external, generalizability, and consequential element of

Table 9

How the Research Questions are addressed

Research Questions	Documents	Goal of the Analysis	Evidence
When examining the design and student data of two LSAs in reading, what common thinking skills are assessed?	Assessment Framework	Examine the document in terms of the thinking skills to be assessed.	Content
	Test Booklets	Examine each item in terms of what thinking skill is elicited by the items.	Substantive
		Conduct a nonlinear factor analysis to determine the dimensionality of the items.	Internal
	Scoring Guide	Examine the document in terms of the thinking skills being assessed.	Content
When examining the design and student data of two LSAs in reading, what are the LSAs' underlying assessment perspectives?	Assessment Framework, test booklets, and scoring guide.	Examine the document in terms of the potential underlying assessment perspectives.	Content

Messick's (1995) validity framework, which were previously described in the conceptual framework.

Chapter Five: Results

This chapter presents the results regarding the PCAP and EQAO and is organized by the research questions. As previously mentioned, the research questions were: When examining the design and student data of two LSAs in reading, (a) what common thinking skills are assessed? And (b) what are the LSAs' underlying assessment perspectives? Therefore, these questions are addressed in the context of the current study's validation approach, in which the evidence included both a content and nonlinear factor analysis.

The Assessment of Thinking Skills in the Context of PCAP and EQAO

The validation arguments for the current study, based on the research questions and focused on thinking skills, were: a) students are being assessed on thinking skills and student results reflect this; b) the assessments are multi-dimensional to reflect the targeted thinking skills; c) the thinking skills are explicitly detailed in the assessment framework, the items themselves, and in the scoring guide. In order to answer the research question related to thinking, both a content analysis of the assessment framework, test booklets, and scoring guide and a dimensionality analysis (non-linear factor analysis) were conducted. The following section reports on the results from both the EQAO and PCAP.

EQAO results. To examine the thinking skills present within the EQAO documents, the assessment framework, test booklets, and scoring guides were content analyzed based on the thinking definitions provided in the conceptual framework. In addition, a non-linear factor analysis was conducted to examine the dimensionality of the EQAO assessment items. Prior to presenting the results of the content analysis, the results of the inter-rater reliability analyses between the primary researcher and the external raters are reported next.

Inter-rater reliability analysis. For the current study, three individuals were asked to serve as external rates. The third external rater was also tasked with coding the assessment framework for thinking skills and assessment perspectives. The results of the third external inter-rater analysis are displayed in Appendix E. Table 10 displays the percentage of agreement and Cohen's Kappa for the three inter-rater reliability analyses. When codes are not applied equally among raters (e.g., the 2nd rater does not use all of the codes), SPSS 19 takes into account the discrepancy and provides a weight so that Cohen's Kappa can be calculated. The inter-rater results for the third external rater revealed a 77.1 % agreement with a Cohen's Kappa of 0.68, which is considered substantial. Disagreement between the primary investigator and the third external rater was found when coding items that may tap into more than one HOT skill. That is, the primary investigator identified two HOT skills within one question, while the external rater identified one of them. While there was one coded list for both the primary researcher and the external raters to produce a Cohen's Kappa; after further discussion, one list was retained that best represented the items.

Assessment framework. Some aspects of critical thinking found in the EQAO's LSA involved the ability to predict, compare, and make judgments (Appendix F). Aspects of critical thinking were mentioned in the assessment framework through the use of words such as 'making inferences', for which students would be required to offer general claims from specific evidence. For example, EQAO noted that"…analyzing textually implicit information and ideas that are directly stated in one part of the text… (Pearson & Johnson, 1978; Raphael, 1986)" (EQAO, 2007; p. 12) or synthesizing "…textually explicit information and ideas that are found in more than one place in the text…" (EQAO, 2007; p. 12). Thus, the student must be able to generate or select the relevant information from multiple areas of the text, which would not have been

Table 10

Results from the inter-rater reliability analyses

Rater	# of items coded	% agreement	Cohen's Kappa	Strength
1	12	41.7	0.24	Slight
2	35	75.0	0.43	Moderate
3	35	77.1	0.68	Substantial

previously exposed to the students. Problem solving was mentioned, with reference to the word 'solve'; however, no further information was provided. Creativity was not apparent in the assessment framework.

Test booklets. The results of the content analyses of both test booklets analyzed as a whole and at the individual item levels for EQAO are in Appendix G. Overall, it was noted that there was a high proportion of items measuring critical thinking (13 items), followed by critical thinking or lower level thinking (12 items), critical thinking and aspects of creativity (7 items), and lower level thinking (3 items). Some questions were coded as tapping into either lower level thinking or critical thinking. These items asked students to provide an alternate term for a concept found within the text. In some cases, if these students had been already exposed to the concept they would easily be able to find an alternate word. However, if a student had not been exposed to the concept, then they would have to synthesize information within the text to determine the meaning.

A nonlinear factor analysis was conducted to examine the structure of the EQAO assessment. There were data missing from the 2007 EQAO reading assessment. The type of missing data were defined as value, meaning that individuals only responded to specific questions and left some out. As part of the EQAO 2007 reading assessment, the following codes were used to denote missing data: missing or not applicable data were coded as -99, an item excluded from the student's results was coded as -97, a student who was dropped from the assessment was coded as -96 (i.e., value type of missing data), and a student who provided no response was coded as -9. In general, all missing data coded as -99, -98, -97 were recoded as 0. As one of the disadvantages to using the NOHARM program is that there is no command for addressing missing data, the type of data were excluded from the analysis.

The raw Excel data set included 256,979 grade 3 and 6 French and English students. As the study was only interested in grade 6 students, grade 3 students were removed from the sample, which yielded 142,897 grade 6 students. French grade 6 students were also removed from the data set, as this was not examined in the current study, yielding a final data set of 137,171 English grade 6 students. The excel data set was inputted into SPSS and the responses for the 35 questions were re-coded in order to have the data presented in a binary format (i.e., 0 and 1). All wrong responses coded as -9, -6, -4, -3, -2, and -1, were coded as 0 and all right responses coded as 1, 2, 3, 4 were coded as 1. Other codes for the multiple choice items, such as missing (-99), excluded item (-97), no response (-9), and system missing responses were coded as missing. For the open-ended questions, codes such as, an excluded item (-97) and no response (-9) were coded as missing, and unrelated responses (-8) were coded as 0. As NOHARM has no function to account for missing data, the cases with any missing data were deleted from the file, which resulted in a final data set of 137,104. The SPSS data were then exported as an excel file and the excel file was saved as a text file.

In order to input the text file into the NOHARM program, there were several variables that needed to be defined in the file. The first three lines of the text file were used for the criteria needed for the NOHARM program (Fraser & McDonald, 1988). There were 35 questions, two factors were being extracted, and there were a total of 137,104 English students from grade 6. It is important to note that 2, 3, and 4 factors were explored, as indicated in the methodology; however, differences in fit were extremely small, except between 1 and 2 factors. Once all the NOHARM job criteria were met (i.e., the conditions were addressed in the text file; see Fraser & McDonald, 1988), the NOHARM program was run.

Several analyses were conducted to examine different factors using both promax and varimax rotations. However, a two-factor solution with a varimax rotation best represented the binary data, as seen in Appendix H. The index of GFI was 0.997, which is considered a near perfect fit (Fraser & McDonald, 1988) and the RMS was 0.001703; indicating little difference between the data and the model. In sum, there was good model fit. The two factor solution best represented the data because, in comparing the GFI among 2, 3, and 4 factors (Table 11), the fit changed significantly between 1 and 2 factors and no significant change was seen from 2 to 3 factors. In essence, the GFI acts like a scree plot. The two factor solution was the most simple to understand and interpret (Kim & Mueller, 1978), as RMS was also smaller between 1 and 2 factors. Only eigenvalues above 0.3 were considered, as suggested by the literature (Gorsuch, 1983). The two factors could be identified as lower level thinking skills and HOT skills. As well, factor 2 contained a higher proportion of open-ended questions (11 items; 10 open-ended; 1 multiple choice) and factor 1 (22 items; 22 multiple choice) contained only multiple choice responses.

For the first factor, which contained 22 items, Cronbach's alpha (Cronbach, 1951) was 0.771, which is considered acceptable (Nunnally & Bernstein, 1994). The Cronbach's alpha (Cronbach, 1951) for the second factor was 0.687. Cronbach's alpha rose up to 0.760 when a multiple choice item was deleted from the scale (i.e., booklet 2, item 8) of the second factor. Thus, factor one best represented multiple choice items and lower level thinking, while the second factor best represented open-ended questions and HOT. However, based on the content analysis, both multiple choice and open ended questions could be tapping into HOT or lower level thinking, depending on the previous experience and knowledge of the student. The first

Table 11

Comparison of GFI and RMS among 1, 2, 3, and 4 factors.

Factors	GFI	RMS
1	0.9939530	0.0025195
2	0.9971088	0.0017422
3	0.9980381	0.0014351
4	0.9985374	0.0012391

factor included items that dealt with describing a person in the text or word meanings. Other items that were coded as a HOT skill included summarizing the text by providing a title.

Scoring guide. The scoring guide did not provide enough information to determine the nature of thinking, as it contained short incomplete sentences. There was an inability to gather enough evidence to suggest which type of thinking skills were present within the EQAO scoring guide because of the lack of description of what the scores meant. This was primarily due to the fact that the scoring guide only provided information in point form.

PCAP results. The content analysis involved the coding of three PCAP documents that included the assessment framework, test booklets, and scoring guide. However, due to confidentiality, only an overall appraisal of the items relating to the test booklet could be reported.

Assessment framework. The PCAP assessment framework provided several references to different thinking skills, as defined in Appendix I. Critical thinking was referenced, as the test developers of PCAP stated "[r]eaders examine the text to develop an understanding of the relationships of discrete elements to the whole" (CMEC, 2007; p. 4) and "…making predictions and asking questions…synthesizing" (CMEC, 2007; p. 2). Problem solving was also referenced (e.g., "…[m]uch as in science and mathematics problem solving, students must structure the problem for themselves in order to solve it…"[CMEC, 2007; p. 6]), as well as some elements of creativity, in which "[r]eaders elaborate personal connections and reactions to texts by providing some extended explanations, examples, and supporting arguments. They state personal points of view and justify them" (CMEC, 2007; p. 4). Lower level thinking skills were referenced with the statement, "[r]eaders identify both explicit…aspects of text…" (CMEC, 2007; p. 4).

Test booklets. The PCAP text booklets contained a large number of items (86 items). The highest proportion of thinking skills were associated with critical thinking (57 items), followed by lower level thinking (13 items), and creativity (2 items). The creative items provided students the chance to include their own ideas, elaborate, with no specific question about the reading content. Some items were also coded as tapping into either critical thinking or lower level thinking (14 items), which was similar to the EQAO test booklets. These questions asked students to define a concept. In such a case, if an individual had been previously exposed to the concept they would have an easier time defining it. However, those students not previously exposed to the item would have to determine the definition by synthesizing the information in the text.

Scoring guide. The stated skills, described in both the assessment framework and scoring guide, overlapped with the different thinking skills (Appendix J). For example, comprehension was similar to lower level thinking skills, such that individuals are asked to identify explicit information gleaned within the text. Interpretation was similar to elements of critical thinking because the goal of some of the items are to enable the students to define, select, and generate responses from content that they have not been previously exposed to. Critical response was also similar to critical thinking as students are required to reflect, compare, and make a judgment. Personal response was closely aligned to critical thinking, and to a certain extent, creativity, only if the definition of creativity includes aspects of personal points of views and ideas.

Underlying Assessment Perspectives

The arguments presented in the conceptual framework on the underlying assessment perspectives were closely aligned with the second research question. The arguments were:

Students' assessment results reflect elements of specific assessment perspectives, which are apparent within the assessment framework, scoring guide, and items themselves.

EQAO results. The authors of EQAO organized the assessment framework such that one chapter was dedicated to the goals of the assessment and one chapter was dedicated to the concept of reading. As such, the definitions provided in the conceptual framework on the different assessment perspectives were used for coding the assessment chapter.

Assessment framework. As depicted in Appendix K, evidence in the assessment framework suggested that the EQAO included information relating to a product-based perspective rather than to a process-based perspective. For instance, the authors of EQAO (2007) stated that "EQAO's large-sale assessments provide reliable, objective and high-quality data that can inform school boards' improvement planning and target setting" (p. 4), which indicated the use of assessing the product of achievement and to use these results to improve educational programs. In addition, it was reported in the document that "[i]n order for students' results on EQAO's large-scale assessment to be comparable across the province, the assessments must be administered, scored and reported on in a consistent manner" (p. 5), which also indicated that the assessment results are compared to other schools (i.e., scoring interpretation). Other key words that tapped into elements of a product-based perspective were the inclusion of "...comparable year-to-year data" (EQAO, 2007; p. 4), "...reliable, objective, and high-quality data..." (EQAO, 2007; p. 4), "...scored at a distance..." (EQAO, 2007; p. 4), "...summative..." (EQAO, 2007; p. 4), "...standardized tasks and under standardized conditions..." (EQAO, 2007; p. 4), "...psychometrically comparable items..." (EQAO, 2007; p. 5), and "...same scoring guides..." (EQAO, 2007; p. 5).

Test booklet and scoring guide. For the test booklets and scoring guides, evidence strongly indicated that the assessment was aligned with a product-based perspective. The test booklets contained a high proportion of multiple choice items that had only one correct response (i.e., selected-response item format). However, some questions contained elements of the process-based perspective. For example, one question in the test booklets required students to "...use information from the selection and your own ideas to support your answer" (EQAO, 2007; Test booklet 1; p. 3), which would indicate students active involvement in the construction of meaning. Although there were questions that asked students to include their own ideas, the scoring guide did not provide any resources for raters to take that information into account with respect to their final score.

The scoring guide, available online, was succinct, structured, and included six levels of descriptors (blank, illegible, code of 10, 20, 30, and 40) for raters. There were two aspects to each code: (1) How well the response addressed the question; and (2) What type of support was used to respond to the question. The level of support is determined by how well the student uses information in the text to support their responses. For example, within the scoring guide it is stated that the "…response does not refer to ideas and information from the reading selection" (EQAO, 2007). This would indicate that test developers view meaning as residing in the text. This theme was common among all levels of the scoring guide which could mean that the scoring guide (Table 12) included elements relating to the product-based perspective.

PCAP results. To examine the underlying assessment perspectives, the definitions provided in the conceptual framework were used to analyze the assessment framework, test booklets, and scoring guide. The PCAP assessment framework clearly described the process of

Table 12

EQAO Assessment: Analyzing the Scoring Guide

Assessment Perspective	Evidence	Reasons
Product-based	"…response does not refer to ideas and information from the reading selection" (EQAO, 2007)	*Assumption*: Meaning resides in the text.
	"…response indicates a significant misunderstanding of the reading selection; ideas and information from the reading selection are inaccurate" (EQAO, 2007)	*Assumption*: Meaning resides in the text.
	"…response is developed with limited support; ideas and information from the reading selection are minimal, vague and/or irrelevant" (EQAO, 2007)	*Assumption*: Meaning resides in the text.
	"…response is developed with some accurate, specific and relevant ideas and information from the reading selection; some ideas and information are inaccurate, vague and/or irrelevant" (EQAO, 2007)	*Assumption*: Meaning resides in the text.

Table 12

Continued

Assessment Perspective	Evidence	Reasons
	"…response is developed with accurate, specific and relevant ideas and information from the reading selection" (EQAO, 2007)	*Assumption*: Meaning resides in the text.
Process-based	None apparent	

reading, including skills measured by the items. In particular, the assessment framework provided a detailed description of the scoring for the open-ended questions.

Assessment framework. Within the assessment framework (Table 13), the results of the content analysis offered evidence that the assessment was geared towards a product-based perspective. Test developers of the PCAP emphasized the use of a formative process in the classroom but also iterated that the overall goals of LSAs are to provide objective results indicating the role of a summative assessment. For example, "[a]ssessments external to schools can have a valuable impact on teaching practices…" (CMEC, 2007; p. 1-2), indicating one of the goals of the assessment is to improve educational resources and programs.

Test booklets. A large proportion of questions in the test booklets[4] were directed towards a product-based perspective, as the questions had only one correct answer (i.e., more than 80 percent were multiple choice). Some of the open-ended questions gave students the opportunity to explain their response based on past experiences. For example, students were asked to comment on the passage, with no specific question regarding the reading selection. This was only required by students on a small proportion of the questions (i.e., two) so PCAP's questions were largely governed by a product-based perspective.

Scoring guide. The scoring guide (Table 14) provided evidence to suggest that the open-response items were geared towards both a product-based and process-based perspective. Elements of the product-based perspective were evident in the comprehension and interpretation descriptors at all levels (i.e., 1, 2, and 3), as the guide focused on how well students addressed questions via elements in the text. For example, the test developers of the PCAP stated "…identifying limited and/or general aspects of the text" (CMEC, 2007; p. 9), "…recognising…ideas in the text" (CMEC, 2007; p. 9), and "…complexity in the text" (CMEC,

[4] Due to the confidentiality of the PCAP assessment, no specific information from the test booklets was reported.

Table 13

PCAP Assessment: Analyzing the Assessment Framework

Assessment Perspective	Evidence	Reasons
Product-based	"In contrast, large-scale assessments are one-time paper-and-pencil measures. The content and the administration procedures are standardized to ensure the results mean the same thing in different contexts. The difference between classroom assessment and large-scale assessment is based on the need for quite different information: immediate and contextualized data for the former as opposed to rigorously comparable results for the latter." (CMEC, 2007; p. 1)	*Scoring Interpretation*: Students' results are compared to other students and schools (e.g., comparable results).
	Assessments external to schools can have a valuable impact on teaching practices and function as a pedagogical resource, provided the education community uses the results in the ways for which they were designed." (CMEC, 2007; p. 1-2)	*Goal of Assessment:* The goal is to use the results to improve educational programs (e.g., impact on teaching practices).

Table 13

Continued

Assessment Perspective	Evidence	Reasons
	"This assessment is designed to report on the four domains indicated above. Reporting this level of specificity will support jurisdictions in developing, adopting, and adapting education policies and programs so as to focus on continuous improvement. "It will also enable provinces and territories to improve their own assessments and to validate their results by comparing them to both national and international results." (CMEC, 2007; p. 4)	*Goal of Assessment:* The goal is to use the results to improve educational programs (e.g., adapting education polices). *Scoring Interpretation*: Students' results are compared to other students and schools (e.g., comparing…national and international results).

Table 13

Continued

Assessment Perspective	Evidence	Reasons
	"The traditional multiple choice format comprises a stem statement and four choices, one of which is correct, while the other three function as distracters. This is the format most familiar to teachers and students. Each item focuses on a single domain." (CMEC, 2007; p. 6)	*Item Format:* Multiple choice (selected-response)
Process-based	None apparent	

Table 14

PCAP Assessment: Analyzing the Scoring Guide

Assessment Perspective	Evidence	Reasons
Product-based	"The reader demonstrates partial understanding: • identifying limited and/or simple aspects of the text" (PCAP, 2007; p.9)	*Assumption*: Meaning resides in the text.
	"The reader demonstrated clear understanding: • identifying both explicit and implicit aspects of the text • recognizing some abstract language, details, and implied ideas in the text" (PCAP, 2007; p.9)	*Assumption*: Meaning resides in the text.
	"The reader demonstrates a nuanced understanding: • recognizing aspects of style, specific selections of detail, organization, and/or complexity in the text " (PCAP, 2007; p.9)	*Assumption*: Meaning resides in the text.

Table 14

Continued

Assessment Perspective	Evidence	Reasons
	"The reader demonstrates limited interpretation of the text: • providing a simplified and/or general perspective of aspects of the text • relying on stated conclusions from the text to inform meaning, and/or connecting aspects of the text to each other" (PCAP, 2007; p.9)	*Assumption*: Meaning resides in the text.
Process-based	"The reader demonstrates thoughtful or insightful interpretation of the text: • synthesizing several elements of the text or expressing thoughtful analysis of a significant element of the text • relying on subtle relationships among elements and ideas to develop a thoughtful or insightful perspective" (PCAP, 2007; p.9)	*Assumptions:* Students are active creators of knowledge (e.g., insightful perspective).

Table 14

Continued

Assessment Perspective	Evidence	Reasons
	"The reader demonstrates a limited personal response to the text: • expressing personal connections to minor details that are valid but simplistic, vague, or common • providing a response that is supported with little or no specific, textual, or personal references" (PCAP, 2007; p.9)	*Assumptions:* Students are active creators of knowledge (e.g., personal connections).
	"The reader demonstrates appropriate personal response to the text: • expressing personal connections linked to key ideas • supporting a response with examples, general explanations, or statements that are direct or clearly inferred from the text or personal knowledge" (PCAP, 2007; p.9)	*Assumptions:* Students are active creators of knowledge (e.g., personal connections).

Table 14

Continued

Assessment Perspective	Evidence	Reasons
	"The reader demonstrates elaborated personal response to the text: • expressing personal implications and insights • supporting a response with specific details, examples, precise explanations, or thoughtful justifications" (PCAP, 2007; p.9)	*Assumptions:* Students are active creators of knowledge (e.g., personal implications).
	"The reader demonstrates appropriate critical response: • expressing a perspective on the text based upon some social, cultural, or literary awareness"(PCAP, 2007; p.9)	*Assumptions:* Students are active creators of knowledge (e.g., expressing a perspective).
	"The reader demonstrates significant critical response: • supporting response with reference to voice, style, stance of the writer, organization and structure, and analogies or illustrations from the world at large." (PCAP, 2007; p.9)	*Assumptions:* Students are active creators of knowledge (e.g., from the world at large).

2007; p. 9). Conversely, the codes at the higher levels for personal and critical response descriptors reveal the significance of the student being active in the construction of meaning. For example, the information from the scoring guide states that students rely "…on subtle relationships among elements and ideas to develop a thoughtful or insightful perspective" (CMEC, 2007; p. 9).

Results Summary

The research questions asked what common thinking skills were evident in the design and student data of two existing LSAs in reading, and what underlying assessment perspectives prevailed within these LSAs. As a result, both a content and dimensionality analysis were conducted to examine these two questions. The content analysis, which had the purpose to evaluate content and substantive validity, included the examination of the assessment frameworks, test booklets, and scoring guides. The nonlinear factor analysis, which had the purpose to evaluate internal validity, included the EQAO student item-level data. Tables 15 and 16 include a summary of the results. The overall appearance of thinking skills and assessment perspectives are determined by the number of codes in each document.

Both the EQAO and PCAP referenced several thinking skills. The EQAO documents did not explicitly define any particular thinking skills, but implicitly referred to the type of skills needed to perform the assessment. There were not many references to problem solving skills, as the only key word 'solve' was used; yet, there were references to critical thinking and lower level thinking, with some elements of creativity (e.g., support with your own ideas). For instance, the EQAO assessment involved students "…inferring or predicting textually implicit ideas using background knowledge and experience together with information in the text…(Pearson & Johnson, 1978; Raphael, 1986)" (EQAO, 2007; p.12). No specific thinking skills could be

Table 15

*Overall results of the thinking skills present within the EQAO and PCAP**

Themes		Number of References for EQAO				Number of References for PCAP			
		Framework	Booklets	Scoring	Overall	Framework	Booklets	Scoring	Overall
Lower level thinking		2	3	LI	5	1	13	0	16
HOT	Critical Thinking	8	13	LI	21	15	57	6	78
	Problem solving	1	0	LI	1	1	0	0	1
	Creativity and Critical Thinking	0	7	LI	7	1	2	2	5
Critical thinking items or lower level thinking		0	12	LI	12	0	14	0	14

*The numbers represent the number of references for each concept; LI: Limited Information

Table 16

Overall results of the underlying assessment perspectives in the EQAO and PCAP

Themes	EQAO				PCAP			
	Framework	Booklets*	Scoring*	Overall	Framework	Booklets	Scoring	Overall
Assessment perspective								
Product-based	21	X	X	21	4	X	X	4
Process-based	1			1	0	x	x	0

*An overall appraisal of the test booklets and scoring guide was conducted. X: dominant appearance; x: lesser appearance (< 5)

gleaned from the scoring guide due to the limited information provided. In sum, all the different thinking skills were implicitly referenced across all of EQAO's documents.

PCAP provided plenty of implicit references of HOT skills within the assessment framework and referred to critical thinking, problem solving, and aspects of creativity. The PCAP also provided an in-depth scoring guide, which indicated the implicit scoring of HOT-related skills. PCAP, thus, also implicitly referenced all of the thinking skills across all the documents.

The nonlinear factor analysis revealed that a two factor solution best represented the data. Factor 1 was characterized by multiple choice items and lower level thinking skills, indicating the role that multiple choice items has in assessing these skills. In contrast, factor 2 was characterized by open-ended questions and HOT skills, demonstrating the role that open-ended questions have in assessing HOT skills. This also identifies the existence of two dimensions in the assessment.

Results revealed that the most prominent assessment perspective evident in the EQAO and PCAP assessment framework and scoring guide was the product-based perspective. The test developers of EQAO and PCAP noted the importance of a process-based perspective within the classroom, but stressed that LSAs are developed to be derived from a product-based perspective. The product-based perspective was also emphasized within the PCAP assessment framework, test booklets, and scoring guide. However, elements of the process-based perspective were also apparent in the test booklets and in the scoring guide of the PCAP assessment.

Overall, both the EQAO and the PCAP assessments implicitly referred to thinking skills and assessment perspectives within their assessment framework, test booklets, and scoring

guides. However, multiple perspectives were present, in addition to different thinking skills. The following section will interpret the results in light of the literature.

Chapter Six: Discussion

LSAs are an important part of the education system as they provide numerous measures of accountability (Taylor & Tubianosa, 2001). They evaluate several disciplines, including reading, which is a critical aspect to the lives of both adults and students because of its necessity in a variety of contexts (McKown & Barnett, 2007). In that respect, the focus of the current study was to examine, using a validation approach (i.e., Messick, 1995; Kane 1990, 1992, 2001), common thinking skills measured by two existing LSAs of reading and their underlying assessment perspectives. The following chapter presents a discussion of the themes that resulted from the content and nonlinear factor analyses. Furthermore, practical and theoretical recommendations, contributions to the field of education, strengths, and limitations are also addressed.

The goals of the current study were twofold: to examine how thinking skills are assessed in LSAs and to identify the underlying assessment perspectives used by LSAs to support the construct of thinking skills. Given that the study was framed around the validation approaches offered by Kane (1992) and Messick (1995), two types of analyses were undertaken: a content analysis and a nonlinear factor analysis. Concepts derived from the literature review, including thinking skills and assessment perspectives, were used to conduct the content analysis. Several themes emerged upon analysis of both EQAO's (i.e., content analysis and nonlinear factor analysis) and PCAP's (i.e., content analysis) documents, including the assessment framework, test booklets, and scoring guides. These themes are discussed in more detail in the following sections.

The Case of Thinking Skills

The teaching of thinking skills has been important to the educational system for decades (Lewis & Smith, 1993). Yet, the assessment of thinking skills has become even more prevalent in recent years (e.g., Schraw & Robinson, 2011), which has resulted in the targeting of these skills by LSAs. It is important to note that the assessment of thinking skills is not unique to education, as other areas, such as medicine, assess similar skills (e.g., MCAT ®, 2012). While many assessments, such as PCAP and the EQAO, are not necessarily high stakes for students and teachers, the results still have an impact on aspects of the educational system, such as on curriculum planning (Abedi, 1997). This can be particularly important when results are used to measure how students are doing with regards to these skills.

In both LSAs reviewed for this study, thinking skills were not explicitly defined in the respective frameworks, but instead the frameworks referred to specific reading skills elicited by the test questions. Moreover, when examining the assessment framework and test booklets, there was a discrepancy between the two documents as to which skills were being targeted. Both LSA frameworks emphasized the assessment of HOT skills; however, when examining the test booklets, an equal proportion of lower level thinking and HOT skills were assessed. After thorough examination of the assessment frameworks, test booklets, and scoring guides, several themes emerged with regards to the testing of thinking skills in both LSAs: (1) prevalence of critical thinking skill questions; (2) items measuring critical thinking and aspects of creativity; (3) the actual nature of critical thinking skills questions; (4) the link between thinking skills and document processing strategies; (5) an alternate explanation for the appearance of two factors; and (6) the importance of assessing HOT.

Prevalence of critical thinking skills questions. Critical thinking was the most commonly assessed HOT skill in the assessment framework, test booklets, and scoring guide for both LSAs. Although this finding may be expected, given critical thinking's pervasiveness in philosophy (Lewis & Smith, 1993) and associated domains (e.g., reading), there are other potential reasons. First, the dominance of the assessment of critical thinking skills could be due to the common association between HOT and critical thinking within the literature. Researchers tend to use critical thinking and HOT interchangeably. For example, Paul and Nosich (1992) proposed a national assessment of HOT and labelled the skill being assessed as critical thinking. Meaning, critical thinking is the HOT skill. They also referred to elements of problem solving skills and creativity within the definition of critical thinking. While this is just one example, it illustrates the issues in understanding which specific skills are considered HOT and the impact of having many terms and definitions provided within the literature (Lewis & Smith, 1993).

The prevalence of critical thinking skills in both LSAs may imply that the test developers considered critical thinking as HOT that encompassed many skills (e.g., problem solving). However, it also may stem from the importance of assessing critical thinking in the curriculum. The Ontario curriculum for language (Ministry of Education, 2009), which is used to develop the EQAO assessment, has sections describing the importance of assessing critical thinking in the language curriculum (e.g., reading). For example, "[s]uccessful language learners: think critically" (Ministry of Education, 2009; p. 4) and

> The program in all grades is designed to develop a range of essential skills in these four interrelated areas, built on a solid foundation of knowledge of the conventions of Standard English and incorporating the use of analytical, critical, and metacognitive thinking skills. (Ministry of Education, 2009; p. 9)

Aspects of the different reading skills may also be closely linked to critical thinking rather than other HOT skills. The reading process involves several types of skills, including both basic and HOT skills (e.g., critical thinking) (Alderson, 2000). Based on the literature, critical thinking skills are often assessed within the domains of philosophy (Lewis & Smith, 1993), which is reflected in the curriculum (Ministry of Education, 2009). The reading skills described by the test developers, thus, would tap into aspects of critical thinking more so than other skills, such as problem solving. Again, this overlap of HOT and critical thinking may be the reason why there is a high prevalence of critical thinking skills in two LSAs that assess reading.

Critical thinking and aspects of creativity. Another noteworthy observation was the nesting or coupling of aspects of creativity within critical thinking skills questions. The part of the assessment items that were considered to tap into creative skills, included those that asked students to provide their own views. Although this does not represent all aspects of creativity, as defined in the conceptual framework, it still offered a potential mechanism for students to provide a unique or novel response. Also, as in the EQAO assessment, the part of the question that addressed individual's personal views was not addressed in the scoring guide, and therefore was not systematically considered as part of the students' results. The nesting of creative skills within critical thinking may imply that test developers have difficulty in assessing this skill accurately in a large-scale format or, again, that test developers consider creativity to be an aspect of critical thinking. Either way, it is not an efficient use of time or useful for students to provide responses that are not included in their final results. Although PCAP does provide a detailed scoring guide, there is no information as to how a students' unique response is included in the score. Overall, the definition of creativity provides the basis as to how the construct will be assessed.

Creativity is considered to be difficult to assess due to the many definitions and terms used within the literature, as well as the implementation of a product-based assessment approach of the last hundred years (Plucker & Makel, 2010). However, as Plucker and Makel (2010) suggest, there are promising approaches to assessing creativity in large-scale formats. Although this approach measured creativity on its own (i.e., outside of a content domain), test developers wishing to assess creativity in reading may want to take this research into account.

The actual nature of critical thinking skills questions. In addition to the observation that some critical thinking skills questions were also addressing a potential aspect of creativity, there is another element that was brought up in light the results. That is, items identified as measuring critical thinking skills could actually be measuring lower level thinking skills if the students had been exposed to the information prior to the assessment. Therefore, based on a student's previous knowledge of the topic, some of the questions could have been tapping into either lower level or into critical thinking skills. Questions that ask about the meaning of a word or about syntax rules in the context of the text were more likely to fall into this category. This finding once again illustrates the importance of clear definitions and terms.

Thinking skills and document processing strategies. It was important to identify the link between reading and thinking and the work by Mosenthal and Kirsch (Mosenthal, 1996; Mosenthal & Kirsch, 1991a) provided a basis for this comparison. It was difficult to determine which thinking skills and which document strategies were associated in the scoring guide and assessment frameworks. However, within the assessment framework and scoring guide, it was apparent that lower level thinking was similar to locate strategies, while critical thinking, creativity, and problem solving had similar characteristics of cycle, integrate, and generate strategies. The examination of the test booklets provided the best evidence for these links, as

document processing strategies are used when there is a goal, which in the case of the test booklet, are the assessment questions. Based on this examination, lower level thinking, again, bore a resemblance to locate strategies. Questions that could tap into either lower level or critical thinking skills were similar to cycle strategies, as students would have to conduct several locate strategies to determine the correct response. This helps explain why such items can tap into both lower and HOT simultaneously depending on the student's previous experience. Items tapping into critical thinking were closely aligned with integrate strategies, as individuals would be required to compare relevant information to determine the appropriate response. No problem solving items were apparent in the test booklets, so no linking to a specific document processing strategy could be made. While some items required students to provide their own ideas, which is an element of creativity, these items could also be tapping into generate strategies (i.e., use of previous knowledge and experience). These results reveal that thinking skills are closely linked to reading strategies and therefore can be explicitly examined in the context of a LSA.

The match between goals and skills. Both LSA frameworks highlighted the assessment of HOT skills and attributed little to lower level thinking skills, while the current study found that both levels were equally present in the test booklets. As such, students, parents, and other stakeholders may be under the impression that the assessments reflect students' HOT skills, when in reality there may be a higher proportion of lower level thinking skills being evaluated. A clear understanding of what is actually being assessed is vital since administrators may use assessment results to adopt educational policies (Abu-Alhija, 2007). For instance, if administrators believe that students are performing well with HOT tasks related to reading, educational reading programs may be readapted to reflect students perceived high performance related to HOT tasks. However, this change may prove to be to the detriment of the school

program because the results upon which the changes were based only mirrored their achievement with lower level thinking tasks. As a result, future funding may be impacted if the school then performs poorly. Nevertheless, it is good to strive for and develop HOT rather than lower level skills, as the hierarchical nature of HOT implicitly assumes that lower level thinking skills are also included. However, care should be taken to ensure that the assessment items still mirror the goals of the assessment.

Appearance of Two Factors. Although the nonlinear factor analysis revealed that one factor best represented HOT and the other factor represented lower level thinking, another interpretation is that the two factors represented the item formats (i.e., multiple choice and open ended response). As a result, this may lead to a different conclusion that there is still one construct being measured. Unfortunately, the item level data of PCAP was not examined, which could have further clarified this issue. Future research could conduct nonlinear factor analyses on the item formats separately to determine the thinking skills being assessed and whether there is still the presence of two factors.

Importance of Assessing HOT. It is important to note that the test developers from both PCAP and EQAO explicitly stated that the goal of both assessments was to evaluate reading, with specific reference to their stated reading skills. However, as was noted in the current study and by the work of Mosenthal and Kirsch (Mosenthal, 1996; Mosenthal & Kirsch, 1991a, 1991b), there are links between reading and thinking. Therefore, both LSAs are also implicitly, if not explicitly because of the close link between reading and thinking, also assessing thinking skills. The current study found that within the assessment frameworks the test developers from both LSAs indicated the importance of evaluating reading skills that were closely linked to HOT, such as critical thinking. However, the results from the test booklets revealed a higher proportion

of items assessing lower level thinking. Although lower level thinking skills build on to HOT skills, the goal to measure HOT did not necessarily mirror what was being assessed by the test items. Therefore, assessing lower level thinking is important; yet, the goals of the assessment must reflect these aims.

Underlying Perspectives and Interacting Elements

Critical thinking was predominantly referenced and measured within both LSAs. Therefore, if test developers are interested in assessing HOT skills, the underlying perspective may be indirectly impacting the assessment of these skills. That is, assessing critical thinking may be difficult if test developers require reliable results, but also wish to evaluate HOT skills. However, adequate representation of HOT skills is achievable in LSAs when the right balance of the process-based and the product-based assessment perspectives is adopted. If test developers explicitly state that there are multiple goals to the study, which they already do (e.g., EQAO), it would enable them to include elements of each perspective. They may choose to not only examine the product but also the process of learning; giving them the ability to use elements of each assessment perspective. In any case, the best approach is to explicitly state the underlying elements of the development, administration, and scoring of the assessment (Principles for Fair Student Assessment Practices for Education in Canada, 1993), so as to not incur misunderstandings regarding the assessment results.

Another option for evaluating HOT is to just focus on improving learning and incorporate a process-based assessment perspective. Yet, questions remain as to if it is possible for test developers to align the LSA with a process-based assessment perspective. This perspective is largely associated with performance assessments and the ability to measure HOT and is more expensive to achieve the reliability levels typically associated with product-based assessments

(Stecher, 2010). However, recent research has revealed that performance assessments are possible with increased reliability, if there is thoughtful planning (Edmunds, 2006; Stecher, 2010) including proper rater training (Stecher, 2010). The assessment of HOT skills depends on whether test developers wish to change the format of LSAs altogether or to adapt those already in existence to compliment the underlying perspectives.

The Developmental Nature of Thinking and Thinking Skills

In light of these results, there may be other aspects to consider in the assessment of thinking skills, such as the relationships between the different HOT skills and the potential developmental nature of thinking skills. The current study separated thinking into lower and higher level, with lower level thinking skills building to produce HOT skills. However, there may be hierarchical relationships among the different HOT skills. For example, problem solving skills may be considered part of critical thinking skills (e.g., Paul & Nosich, 1992). The examination of this dichotomy is commonly found within the literature (e.g., Alexander et al., 2011).

An additional aspect relates to the developmental nature of thinking. Cognitive development, in the area of developmental psychology, refers to the changes in mental processes during growth and aging (Shaffer, Wood, & Willoughby, 2005). Particularly, the works of Piaget and Vygotsky have been pivotal in helping to explain the development of cognitive processes, especially in children (Barrouillet & Gauffroy, 2013; Shaffer et al., 2005). As the current study examined how thinking skills are assessed in two LSAs for grade 6 students (e.g., Piaget's third stage of cognitive development; Shaffer et al., 2005), it is possible that only certain thinking skills could be evaluated for that age group. Although not explicitly identified by both LSAs,

these assessments may have been developed to take into account the cognitive processes for a specific age group and thus, the type of skills evaluated would reflect this.

Validation using Messick and Kane

The current study implemented the argument approach to validity suggested by Kane (1992), with evidence provided by Messick (1995). Two interpretive arguments were put forth: (1) students are assessed on thinking skills and student results reflect this, and (2) There is an underlying perspective framing the development and scoring of the assessment. While the first argument included two inferences: (1) the assessment is multidimensional, and (2) thinking skills are explicitly detailed, the second argument included one: (1) the student documents take into account an assessment perspective. Based on these inferences, the content, substantive, and internal elements of Messick's (1995) construct validity framework were incorporated. The use of Messick's (1995) and Kane's (1992) approach to validity provided a means in which to examine the two research questions. However, there is also the possibility to further examine the thinking skills and underlying assessment perspectives through other types of evidence. For example, it would be interesting to compare the consequences of achievement scores in different settings (e.g. high stakes versus low stakes) to examine the underlying skills and perspectives (i.e., generalizability element). Additionally, conducting focus groups and interviews with those directly impacted by achievement results would have been advantageous. This would enable the researcher to understand individuals' perceptions of the skills being assessed and the underlying influences (i.e., consequential element). Although the examination of these areas would be useful, due to resources and cost, these aspects were not undertaken in the current study.

The Use of Definitions

The research from the literature review helped develop the definitions proposed in the current study. Most importantly, the definitions proposed in the current study reflected the background and perspective of the primary researcher. For example, creativity defined in the current study could be considered a narrower view of the construct, as the more you define creativity the less it is considered creative. However, this was the definition that best represented the need of the researcher (i.e., propose codes to identify skills within two LSAs) and what was found within the literature. As a result, the perspective of test developers, including the goal of the assessment, should be considered when developing definitions of constructs that will be evaluated. It may be useful to develop definitions that reflect the goals of the assessment perspective (e.g., process- and product- based assessment perspective). Additionally, if test developers wish to incorporate both perspectives, perhaps a definition that reflects characteristics of both a product-based and process-based perspective is needed.

Theoretical and Practical Recommendations

As indicated from the assessment frameworks of both LSAs, test developers may lack the necessary resources for adopting different thinking skills and underlying assessment perspectives. Additional research is required to help clearly articulate underlying assessment perspectives, including models and theories that could frame LSAs in terms of their goals. This would aid test developers of LSAs in identifying theories and models and in adopting the most appropriate approach for their respective context. Wolfe, Childs, and Elgie (2004), suggest the inclusion of a bridging document that showcases clear links between the assessment blueprint and how it is mapped to the curriculum.

While there exists a prevalence of critical thinking skills in LSAs that measure reading, questions assessing problem solving could also be designed (Haladyna, 1997). For example, it may be interesting to examine LSAs in mathematics and science to examine how problem solving skills are being assessed, as well as to identify which perspectives underlie these assessments. Nevertheless, clear definitions are needed if test developers are interested in assessing thinking skills (Treffinger et al., 2002).

Critical thinking is a process; in effect, to assess such a skill, the goal of the assessment should be to understand the learning process and test developers should utilize elements of the process-based assessment perspective to do so. In practice, there should be a balance of different elements of underlying perspectives that meet the goals of the assessment, as demonstrated in the conceptual framework. Ultimately, it is up to the test developer to determine which perspective is adopted and which skills are assessed to best represent the goals of the study.

With regards to validity, the consequential element is the most debated element of Messick's (1995) construct validity framework (Popham, 1997). Some studies have suggested that consequences should be assessed by the test developer and not test validators, as it is up to the test developers to understand the impact of score responses (Nicholas & Williams, 2009). Due to the limitations of a secondary data analysis, the examination of the consequences of test use was not carried out. However, the discrepancies found across the LSA documents in the current study points to the potential consequences of how teachers and students may interpret the results. For example, if teachers are under the impression that a LSA will help assess the complex thinking skills of their students, and this is used in conjunction with their classroom assessments, students' achievement results may be misrepresented. Future studies could address such consequences by conducting focus groups and interviews with students, teachers, and

administrations. There is also a broader question of how the achievement results of LSAs can potentially impact students over their careers. Additionally, research could also examine the impact of LSA results on strategic planning at the provincial or district level, multicultural student groups, student retention rates, curricular reforms, professional development initiatives, and other related reforms (e.g., Literacy and Numeracy Secretariat).

It may also be beneficial to use the conceptual framework from the current study to compare the EQAO with another LSA that evaluates reading or one that evaluates another discipline (e.g., mathematics) to determine similarities and/or differences in how thinking is evaluated. It would also be valuable to examine the linguistic comparison of French and English versions and the assessment of thinking skills. Additionally, researchers could examine the difference in score interpretations if the EQAO was considered a high-stakes test versus a low-stakes test. Finally, researchers could conduct focus groups or interviews in which they ask students if specific thinking skills, based on the conceptual framework from the current study, were apparent when they took the EQAO or PCAP assessment.

Examining multidimensionality, such as the existence of thinking skills, is appropriate when implementing an IRT method to determine construct irrelevant and construct underrepresentation variance. The current study was not able to implement such analyses due to the limitations associated with a secondary data analysis. However, the appropriateness and refinement of dimensionality analyses for dichotomous data needs to be explored (Dunn 2002).

In addition to the descriptive account of the substantive element in the current study, test developers or researchers could incorporate think-aloud procedures and see how they link to what is proposed by the assessment; such as those that tap into lower level thinking versus HOT skills. This approach would help ensure that the skills represent the intended construct domain.

More specifically, older students could be provided with a coding sheet with explanations as to the different thinking skills (i.e., lower level thinking, critical thinking, creative thinking, and problem solving) and asked what skills are elicited by the items. On the other hand, younger students could be asked what they think about when they respond to the items and researchers could code their results based on the different thinking skills. Although time consuming, this would greatly augment the ability of test developers to evaluate skills more closely aligned with the construct domain.

Contribution of the Research to the Field of Education

There are many ways to assess and define thinking skills. By examining in depth how the two LSAs define and measure these skills, and the underlying assessment perspectives adopted, this study shed light into the extent to which the theoretical, conceptual, operational, and statistical models correspond to one another. Moreover, research studies that adopt a validation perspective are usually based on a traditional understanding of validity that does not incorporate multiple lines of evidence (Gustafson & Rosén, 2005; Kapinus, 2003). Adding to the literature, the current study contributes to research in the field of education by developing frameworks for the examination of thinking skills and assessment perspectives, using a validation approach. There is the potential for test developers to use the benchmark definition of thinking skills and examine how their assessment may vary depending on the underlying assessment perspective (i.e., product or process based). Enabling test developers to adapt the frameworks for the specific purpose of acknowledging the presence of thinking skills and assessment perspectives while conceptualizing, developing, and reporting results.

Strengths and Limitations

There are strengths associated with the current study, including the provision of a comprehensive overview of assessment perspectives and thinking skills. All of which are areas that have had many different definitions. In addition, the literature review facilitated a coding framework developed to examine LSAs in reading, which could be adapted and used with other LSAs. Moreover, application of the NOHARM program provided awareness as to the limitations associated with the use of a regular linear factor analysis in analyzing binary data. Overall, the current study linked thinking to reading, explored different ways of classifying thinking, examined the factor structure of an EQAO data set, and examined the distribution of thinking skills in two LSAs.

There are also some limitations to consider when examining these results: (1) the results from the PCAP and EQAO cannot be generalized to other LSAs, (2) there was no examination of curriculums and associated test blueprints for each LSA, and (3) the documents were examined from an external perspective. Though the results specific to either the PCAP or EQAO could not be generalized to other regional, national or international assessments, as identified by the first limitation, the thematic codes based on different assessment perspectives and thinking skills have the potential to be adapted and used with other LSAs. Secondly, there is one curriculum associated with the EQAO assessment, while the PCAP is based on a general curriculum (i.e., elements of curriculums across Canada); therefore, making it difficult to examine the PCAP's associated curriculum. The test blueprint for both LSAs was not initially examined because it provided limited information as to the underlying perspectives. With regards to the third limitation, though examining documents from a distance could be considered a strength (no bias) in the current study, it may also be considered a limitation. That is because the primary

researcher is not trained as a rater for both LSAs and was not privy to specific documents and knowledge regarding the scoring of the test items. Additionally, the external raters who helped develop the codes were not in education and not privy to useful assessment information, such as the curriculum. Having external raters, such as teachers, that were aware of the curriculum and the assessment of thinking skills in the classroom could have impacted how the codes were developed and which skills were deemed important. Overall, despite these limitations, the current study provides a guide for future studies in this area.

Discussion Summary

The current study clarified the mechanisms that make up a LSA, including thinking skills and assessment perspectives from a validation approach based on the work by Messick (1995) and Kane (1990). A content analysis was carried out on two LSAs, with a follow up on one of the LSAs with a nonlinear factor analysis. Results revealed that there were interacting elements between the assessment and the items as to an assessment perspective and concepts of thinking. The tools provided in the current study have the potential to help inform both researchers and practitioners about the interaction between the assessment approach and the perspective related to a specific discipline (e.g., reading) and their impact on given assessment purposes.

Chapter Seven: Conclusion

The purpose of the current study was to examine the operationalization and conceptualization of thinking skills within LSAs that focus on reading. Moreover, it was framed within a validation perspective implementing Messick's (1995) and Kane's (1990, 1992) approach to validity. As a result, elements regarding the underlying assessment perspective were investigated. LSAs on reading were chosen as reading is a large component of students' and adults lives (McKown & Barnett, 2007). The following chapter will present an overview of the study

LSAs serve a variety of purposes (Taylor & Tubianosa, 2002), evaluate both content- and thinking-related skills (Rijmen, 2011), and are, even implicitly, developed based on theoretical models (e.g., assessment model) (Principles for Fair Student Assessment Practices for Education in Canada, 1993). As a result, it was important that the current study take these factors into account when examining the construct of thinking skills, through a validation framework. Therefore, the goal was to conduct a study, using a validation approach, which examines how thinking is assessed within two LSAs that focus on reading.

There are three areas that have predominantly shaped LSAs. The scientific paradigm, in the early 20th century, was driven by scientific inquiry (Cochrane, 2007; Kuhn 1961). The emergence of standardized tests, based on scientific inquiry (Giordano, 2005) led to the mental measurement movement (Simon et al., 2012). And finally, educators and educational leaders became interested in the use of standardized testing in the area of education, which led to the educational assessment movements, including the examination of programs and the improvement of learning (Giordano, 2005). Accordingly, there are continued efforts to diagnose, improve, and certify learning (Abu-Alhija, 2007; Giordano, 2005; Simon et al., 2012; Ungerleider, 2003).

The goal of the assessment is to guide the development of the actual assessment (Anderson & Morgan, 2008). Based on the different assessment movements, two assessment perspectives emerged; the product-based and process-based perspective. Elements of the product-based perspective include the notion that the goal of the assessment is to regulate or certify learning, the assessment includes selected-response items, achievement results are compared to a standard or to other students and schools, and meaning resides in the text. On the other hand, elements of the process-based perspective include the idea that the goal of the assessment is to improve learning, knowledge is constructed, achievement results are included in the learning process, and items include partial scoring. These two perspectives are not mutually exclusive and LSAs can potentially include elements that tap into both.

Previous research has resulted in a myriad of definitions and terms associated with thinking skills (e.g., Lewis & Smith, 1993). The most commonly cited thinking skills included lower level thinking, critical thinking, problem solving, and creativity. Test developers of EQAO and PCAP defined, in their own way, the skills that were going to be assessed (e.g., reading skills). These skills may not directly be identified as thinking skills, but were closely aligned with such skills. In some cases, the reading skills were described in a general manner. Work by Mosenthal and Kirsch (Mosenthal, 1996; Mosenthal & Kirsch, 1991a) provided the framework to link thinking and reading and specifically to locate, cycle, integrate, and generate strategies.

From the literature review, it was revealed that validity is associated with the uses and interpretation of the results (Messick, 1995). Multiple sources of evidence should be included (Zumbo, 2009) and that the researcher determines the type of evidence needed to support the validity argument (Kane, 1990, 1992). The work of Messick (1995) and Kane (1992) are especially useful in conducting such a validation.

Also, it was considered the validator's responsibility to determine the stated arguments and to choose what type of evidence to include (Kane, 1990, 1992). Based on the validity argument for the current study, it was deemed necessary to analyze the content, substantive, and internal elements of Messick's (1995) construct validity framework, due to the current study's emphasis on the conceptualization and operationalization of thinking skills. As previous research suggested, content can be approached through the examination of the domain definition, domain relevance, and domain representation (Sireci, 1998a). Furthermore, if the use of think-aloud procedures cannot be collected; the substantive element can be addressed by describing the processes that are elicited by students when completing the items (Lane, 1999). Finally, within the literature, it was determined that factor analyses were commonly used to address the internal structure (e.g., Kuppermintz et al., 1995). However, previous research relied on linear factor analyses with dichotomous data, which has been considered not necessarily appropriate (Dunn, 2002). Thus, based on these findings, the current study implemented a nonlinear factor analysis using the NOHARM program (Fraser & McDonald, 1988).

Conceptually, and based on previous research, two perspectives of assessment and four thinking skills were defined and used as the basis of the content analysis. Bearing in mind that previous research had yet to address the theoretical underpinnings of a LSA while conducting the study via a validation approach, these definitions provided a foundation as to how both LSAs were to be examined. In sum, a thorough conceptual framework was developed in order to examine both LSAs. Although the current study implemented a qualitative analysis, the analysis contained a priori coding, which lent itself to a more quantitative approach.

Results from the current study revealed that critical thinking skills and integrate and cycle strategies were prominent within all three documents for both LSAs. After further examination

of the EQAO documents through a nonlinear factor analysis, it was revealed that some HOT questions may be tapping into lower level thinking, this would, as a result, decrease the number of questions assessing HOTs. However, there was a lack of information regarding the assessment of problem solving skills in the documents of both PCAP and EQAO, with no questions dedicated to these skills.

The assessment framework for both LSAs identified elements of process-based assessment perspective and also focused on the assessment of HOT skills. However, the product-based assessment was predominant in the test booklets and scoring guides. Lower level thinking skills were also predominant in these two documents. This indicated the potential interaction between the stated thinking skills and the underlying perspectives. These results demonstrated the need for LSAs to not only document the perspective of assessment, but also the theoretical foundations of what is to be evaluated (e.g., reading).

Results from the nonlinear factor analysis revealed that the open-ended questions were more apt to assess HOT-related skills, while the multiple choice items were more apt to assess lower level thinking skills. Although research has suggested that this is not always the case (Taylor & Tubianosa, 2001), these results pointed to the importance that item format has on the goals of the test items. Yet, it would have been ideal to also examine the item-level data of the PCAP assessment. The current study demonstrates that it would be in the best interest of both the test developers and researchers to clarify not only the content- and thinking-related goals of the assessment and the theoretical underpinnings, but also how the format of the assessment will facilitate the implementation of these goals.

Based on the result, several themes on the thinking skills and underlying assessment perspectives emerged from the analyses of both LSAs: (1) prevalence of critical thinking skills

questions; (2) items measuring critical thinking and aspects of creativity; (3) the actual nature of critical thinking skills questions; (4) the link between thinking skills and document processing strategies; (5) an alternate explanation for the appearance of two factors; and (6) the importance of assessing HOT. Additionally, it was determined that there were different underlying assessment perspectives present within all three documents. It was suggested that, with the different types of thinking skills being assessed and different underlying assessment perspectives across all three documents, there are potential interactions between the different skills and perspectives. Also, an LSA based on the product-based assessment perspective would enable test developers to assess skills, such as lower level thinking in which answers are explicitly stated. Conversely, an LSA based on the process-based assessment perspective would enable test developers to measure complex skills that require elaborated responses. From this, the current study revealed a need to explicitly detail perspectives of assessment and not just the overall goals of the assessment, and document this link between the assessment framework and the items themselves. However, the question remains as to how LSAs should be defined and where they should fit within the educational context. These questions are pertinent to both practitioners and researchers and warrants further research.

Conclusion Summary

This study examined the conceptualization of thinking skills using Messick's (1995) and Kane's (1990, 1992) approach to validity, as well as examining the underlying assessment perspectives of LSAs that evaluate reading. Although the results from both the PCAP and EQAO are not generalizable to other LSAs, the conceptual definitions derived from the current study have the potential to aid in the future identification of assessment perspectives and thinking skills apparent within other LSAs.

References

Abedi, J. (1997). *Dimensionality of NAEP subscale scores in mathematics*. CSE Technical Report 428. Retrieved from http://www.cse.ucla.edu/products/Reports/TECH428.pdf.

Abu-Alhija, F. N. (2007). Large-scale testing: Benefits and pitfalls. *Studies in Educational Evaluation, 33*, 50-68. doi: 10.1016/j.stueduc.2007.01.005

Alderson, J. C. (2000). *Assessing reading*. Cambridge, New York: Cambridge University Press.

Alexander, P. A., Dinsmore, D. L., Fox, E., Grossnickle, E. M., Loughlin, S. M., Maggioni, L., Parkinson, M. M., & Winters, F. I. (2011). Higher order thinking and knowledge: Domain-general and domain-specific trends and future directions. In G. Schraw & D. R. Robinson (Eds.), *Assessment of higher order thinking skills* (p. 19-46). Charlotte, North Carolina: Information Age Publishing, Inc.

Alexander, P. A., & Fox, E. (2004). A historical perspective on reading research and practice. In R. B. Ruddell & N. J. Unrau (Eds.), *Theoretical models and processes of reading* (pp.33-68). Newark, DE: Reading Association, Inc.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2002). *Standards for educational psychological testing.* Washington, DC: American Educational Research Association.

Anderson, L. W. (2003). *Classroom assessment: Enhancing the quality of teacher decision making*. New York, New York: Taylor & Francis

Anderson, P., & Morgan, G. (2008). *Developing tests and questionnaires for a national assessment of educational achievement*. Washington, DC: The International Bank for Reconstruction and Development/The World Bank.

Arnove, R. F., Altbach, P. G., & Kelly, G. P. (1992). Emergent issues in education: Comparative perspectives. Retrieved from http://books.google.ca

Bacci, E. D., Koger, M. E., Hoffman, R. G., & Thacker, A. A. (2003). *Relationships among Kentucky's core content test, ACT scores, and students self-reported high school grades for the classes of 2000 through 2002*. Retrieved from http://www.kde.state.ky.us/KDE/Administrative+Resources/Testing+and+Reporting+/Kentucky+School+Testing+System/Accountability+System/HumRRO+KCCT+ACT+Validity+Study.htm.

Barrouillet, P., & Gauffroy, C. (2013). Introduction–from Piaget to dual-process theories: the complexities of thinking and reasoning development. In P. Barrouillet & C. Gauffroy, *The development of thinking and reasoning* (p. 1-10). New York, N.Y.: Psychology Press.

Bloom, B. S. (Ed.) (1956) Taxonomy of educational objectives. Handbook 1: Cognitive domain. White Plains, New York: Longmans.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1986). *Taxonomy of educational objectives*. New York: Longmans Green.

Booker, M. J. (2007). A roof without walls: Benjamin Bloom's taxonomy and the misdirection of American education. *Academic Questions,* 20, 347-355. doi: 10.1007/s12129-007-9031-9

Borseboom, D., Cramer, A. O. J., Kievet, R. A., Scholten, A. Z., & Franić, S. (2009). The end of construct validity. In R.W. Lissitz, *The concept of validity: Revisions, new directions, and applications* (pp. 19-37). Charlotte, NY: Information Age Publishing Corp.

Bracey, G. W. (2001). *The 11th Bracey report on the condition of public education*. Retrieved from http://www.america-tomorrow.com/bracey/EDDRA/bracey11.htm

Brown, F. G. (1983). Principles of educational and psychological testing (3rd Ed.). New York, NY: CBS College Publishing.

Candel, M. J. J. M. (2001). Recovering the metric structure in ordinal data: linear versus nonlinear principal components analysis. *Quality & Quantity, 35*, 91-105. doi: 10.1023/A:1004873031561

Carnine, D., & Kameenui, E. J. (1992). Introduction. In D. Carnine & E. J. Kameenui, *Higher order thinking: Designing curriculum for mainstream students*. Austin, Texas: Donald D. Hammil Foundation.

Chalhoub-Deville, M. (2009). Content validity considerations in language testing contexts. In R. W. Lissitz, *The concept of validity: Revisions, new directions, and applications* (pp.241-263). Charlotte, NY: Information Age Publishing Corp.

Chiu, C. Y., & Seo, M. (2009). *Cluster analysis for cognitive diagnosis: An application to the 2001 PIRLS reading assessment*. IERI Monograph article. Retrieved from http://www.ierinstitute.org/html/dissemination_area.html.

Chudowsky, N., Pellegrino, J. W., Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Retrieved from http://books.google.ca/books?hl=en&lr=&id=Omh-Gi8_K9kC&oi=fnd&pg=PA1&dq=educational+assessment&ots=B9BuUtGvSe&sig=ZVMtT6TkqwnGW-pDAPgAJW_BkBc#v=onepage&q&f=false.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334. doi: 10.1007/BF02310555

Cochrane, D. (2007). *The scientific and epistemological paradigm shift of the early 1900's*. Retrieved from http://www.astrosoftware.com/ParadigmShift.htm

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37- 46. doi: 10.1177/001316446002000104

Councils of Ministers of Education, Canada (CMEC). (2007). *PCAP-13 Reading Assessment Framework*. Unpublished report.

Councils of Ministers of Education, Canada (CMEC). (2008). *Education in Canada*. Retrieved from http://cicic.ca/docs/cmec/EducationCanada2008.en.pdf

Councils of Ministers of Education, Canada (CMEC). (2010). *A framework for statistics on learning and education in Canada*. Retrieved from http://www.cmec.ca/Publications/Lists/Publications/Attachments/257/cesc-data-framework-sept2010.pdf

Councils of Ministers of Education, Canada (CMEC). (2012). *PCAP-13 2007: Report on reading strategies and reading achievement*. Retrieved from http://www.cmec.ca/411/Programs-and-Initiatives/Assessment/Pan-Canadian-Assessment-Program-(PCAP)/PCAP-13-2007-Report-on-Reading-Strategies-and-Reading-Achievement/index.html

Craft, S., Jeffrey, B., & Leibling, M. (Eds.). (2001). *Creativity in education*. Retrieved from http://books.google.ca/books?hl=en&lr=&id=eJJY8d26ktkC&oi=fnd&pg=PR5&dq=creativity+in+education&ots=6Xj59cuvol&sig=ONh5VyZWL7L0lRIWrUUuPf_11Ag#v=onepage&q=creativity%20in%20education&f=false.

Creswell, J. W. (2003). A framework for design. In *Research Design: Qualitative, quantitative and mixed method approaches* (2nd ed., pp. 3-26). Thousand Oaks, CA: Sage Publications.

Creswell, J. W. (2007). Philosophical, paradigm and interpretive frameworks. In *Qualitative inquiry & research design: Choosing among five approaches* (2nd ed., pp. 15-22). Thousand Oaks, CA: Sage Publications.

Creswell, J. W., & Plano-Clark, V. L. (2007). Choosing a mixed method design. In *Designing and conducting mixed methods research* (pp. 59-88). Thousand Oaks, CA: Sage Publications.

Creswell, J. W., Plano-Clark, V. L., & Garrett, A. L. (2008). Methodological issues in conducting mixed methods research designs. In M.M. Bergman (Ed.), *Advances in mixed method research* (pp. 66-83). Thousand Oaks, CA: Sage Publications.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, New York: The Guildford Press.

De Champlain, A. F., & Gessaroli, M. E. (1998). Assessing dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education, 11*, 231-235. doi: 10.1207/s15324818ame1103_2

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, 13, 55-77. doi: 10.1207/s15324818ame1301_3

Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education, 37*, 830-837. doi: 10.1046/j.1365-2923.2003.01684.x

Dunn, J. L. (2002). *Exploring the dimensionality of the SAIP mathematics assessment*. Unpublished Master's Thesis, University of Toronto, Toronto, Canada.

Edmunds, J. (2006). *How to assess student performance in history: Going beyond multiple-choice tests*. Produced by the SERVE center at the University of North Carolina at

Greensboro. Retrieved from http://www.serve.org/uploads/publications/AssessHistory.pdf

Education Quality and Accountability Office (EQAO). (2007). *Framework: Assessment of reading, writing and mathematics, junior division (grades 4-6).* Retrieved from http://www.eqao.com/pdf_e/08/6e_Framework_07_web.pdf.

Education and Quality Accountability Office (EQAO). (2010). *What parents need to know about province-wide testing*. Retrieved from http://www.eqao.com/pdf_E/10/WhatParentsNeedtoKnow.pdf.

Education and Quality Accountability Office (EQAO). (2012a). *Grade 6: Assessment of reading, writing and mathematics, junior division*. *Student booklet: Language 1*. Retrieved from http://www.eqao.com/pdf_e/12/6e_Lang1_0612_web.pdf

Education Quality and Accountability Office (EQAO). (2012b). *Teacher bulletin*. Retrieved from http://www.eqao.com/pdf_e/12/36e_Tchr_Bulletin_0212_web.pdf.

Edwards, R. (1989). *Margaret Thatcher, Thatcherism and education. McGill Journal of Education, 24*(2), 203-214.

Farr, R., & Carey, R. F. (1986). *Reading: What can be measured?* 2nd edition. Newark, DE: International Reading Association. 218 pages. Retrieved from http://www.eric.ed.gov/PDFS/ED266438.pdf.

Fogarty, R., & McTighe, J. (1993). Educating teachers for higher order thinking: The three story intellect. *Theory into Practice, 32*(3), 161-169.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267-269.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation, 17*(3), 1-13.

Giordano, G. (2005). *How testing came to dominate American schools: The history of educational assessment*. New York, NY: Peter Lang Publishing Inc.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist, 39*(2), 93-104. doi: 10.1037/0003-066X.39.2.93

Gong, B. (1999). *Relationships between student performance on the MCAS (Massachusetts Comprehensive Assessment System) and other tests - Collaborating District A, grades 4 and 10*. Retrieved from http://nciea.org/publications/StudPerf_Mass_Gong99.pdf.

Gorsuch, R. L. (1983). *Factor analysis (2nd ed.)*. Hillsdale, NJ: Erlbaum.

Gough, P. B., & Turner, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*, 6-10. doi: 10.1177/074193258600700104

Greeno, J. G. (1989). A perspective on thinking. *American Psychologist, 44*(2), 134-141.

Gustafsson, J. E., & Rosén, M. (2005). The Dimensional structure of reading assessment tasks in the IEA Reading Literacy Study 1991 and the Progress in International Reading Literacy Study 2001. *Educational Research and Evaluation, 12*(5), 445-468. doi: 10.1080/13803610600697179

Hager, P. & Butler, J. (1996) Two models of educational assessment. *Assessment & Evaluation in Higher Education, 21*(4), 367-378. doi: 10.1080/0260293960210407

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston: Allyn and Bacon.

Hamilton, L. S., Nussbaum, E. M., Kupermintz, H., Kerkhoven, J. I. M., & Snow, R. E. (1995). Enhancing the validity and usefulness of large-scale assessments: II NELS: 88 Science achievement. *American Educational Research Journal, 32*(3), 555-581. doi: 10.3102/00028312032003555

Hamilton, L. S., Nussbaum, M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education*, *10*(2), 181-200. doi: 10.1207/s15324818ame1002_5

Heald-Taylor, B.G. (1998). Three paradigms of spelling instruction in grades 3 to 6. *The Reading teacher, 51*, 404-412.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277-1288. doi: 10.1177/1049732305276687

International Association for the Evaluation of Educational Achievement (IEA). (2007). *Mission statement*. Retrieved from http://www.iea.nl/mission_statement.html.

International Association for the Evaluation of Educational Achievement (IEA). (2012). *PIRLS 2006 assessment framework and specifications - 2nd edition*. Retrieved from

Ivie, S. D. (1998). Ausubel's learning theory: An approach to teaching higher order thinking skills. *The High School Journal, 82*(1), 35-42.

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

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26. doi: 10.3102/0013189X033007014

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112-133. doi: 10.1177/1558689806298224

Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58*(5), 736-753.

Joliffe, I., & Morgan, B. (1992). Principal component analysis and exploratory factor analysis. *Statistical Methods in Medical Research, 1*, 69-95. doi: 10.1177/096228029200100105

Kane, M. T. (1990). *An argument-based approach to validation*. ACT research report series. Retrieved from https://www.act.org/research/researchers/reports/pdf/ACT_RR90-13.pdf.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535. doi: 10.1037/0033-2909.112.3.527

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342. doi: 10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*, 31-41. doi: 10.1111/j.1745-3992.2002.tb00083.x

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*(3), 135-170. doi: 10.1207/s15366359mea0203_1

Kane, M. T. (2006). *Validation*. In R. Brennan (Ed.), Educational Measurement (4th ed.) (pp. 17-64). Westport, CT: Greenwood Publishing.

Kane, M. T. (2008). Terminology, emphasis, and utility in validation: Comments on Lissitz and Samuelsen. *Educational Researcher, 37*(2), 76-82. doi: 10.3102/0013189X08315390

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R.W. Lissitz, *The concept of validity: Revisions, new directions, and applications* (pp. 39-64). Charlotte, NY: Information Age Publishing Corp.

Kapinus, B. (2003). *PIRLS-IEA reading literacy framework: Comparative analysis of the 1991 IEA reading student and the progress in international literacy study*. Paper presented at the meeting of the National center for Education Statistics, Washington, DC.

Kelly, P. A., O'Malley, K. J., Kallen, M. A., & Ford, M. E. (2005). Integrating validity theory with use of measurement instruments in clinical settings. *Health Services Research, 40*(5), 1605-1619. doi: 10.1111/j.1475-6773.2005.00445.x

Kim, J. -O., & Mueller, C. W. (1978). *Introduction to factor analysis: What it is and how to do it.* Beverley Hill, California: Sage Publications, Inc.

King, F. J., Goodson, L., & Rohani, F. (1998). *Higher order thinking skills*. Retrieved from http://www.cala.fsu.edu/files/higher_order_thinking_skills.pdf.

Koch, M. J., & DeLuca, C. (2012). Rethinking validation in complex high-stakes assessment contexts. *Assessment in Education; Principles, Policies & Practice, 19*(1), 99-116.

Koretz, D. McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992). *The reliability of scores from the 1992 Vermont portfolio assessment*. Retrieved from http://www.rand.org/content/dam/rand/pubs/drafts/2005/DRU159.pdf

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160.

Kuhn, T. S. (1961). The function of measurement in modern physical science. *Isis, 52*(2), 161-193.

Kupermintz, H., Ennis, M. M., Hamilton, L.S., Talbert, J. E., & Snow, R. E. (1995). Enhancing the validity and usefulness of large-scale assessments: I. NELS: 88 Mathematics Achievement. *American Educational Research Journal, 32*(3), 525-554. doi: 10.3102/00028312032003525

Kupermintz, H., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: III. NELS: 88 Mathematics achievement to 12th Grade. *American Educational Research Journal, 34(1)*, p. 124-150. doi: 10.3102/00028312034001124

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*(2), 293-323. doi: 10.1016/0010-0285(74)90015-2

Lane, S. (1999). *Validity evidence for assessments*. Reidy Interactive Lecture Series. Retrieved from http://www.nciea.org/publications/ValidityEvidence_Lane99.pdf.

Leighton, J. P. (2011). A cognitive model for the assessment of higher order thinking in students. In G. Schraw & D.R. Robinson (Eds.), *Assessment of higher order thinking skills* (p. 19-46). Charlotte, North Carolina: Information Age Publishing, Inc.

Leighton, J., & Gierl, M. (2007). Why cognitive diagnostic assessment? In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 3–18). Cambridge: Cambridge University Press.

Lewis, A., & Smith, D. (1993). Defining higher order thinking. *Theory into Practice, 32*(3), 131-137. doi: 10.1080/00405849309543588

Linn, R. L. (1986). Educational testing and assessment: Research needs and policy issues. *American Psychologist, 41*(10), 1153-1160.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-14. doi: 10.3102/0013189X029002004

Linting, M., Meulman, J. J., Groenen, P. J. F., & van der Kooij, A. J. (2007). Nonlinear principal components analysis: Introduction and application. *Psychological Methods, 12*(3), 336-358. doi: 10.1037/1082-989X.12.3.336

Lissitz, R. W. (2009). *The concept of validity: Revisions, new directions, and applications.* Charlotte, NY: Information Age Publishing Corp.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-348. doi: 10.3102/0013189X07311286

Luce-Kapler, R., & Klinger, D. (2005). Uneasy writing: The defining moments of high-stakes literacy testing. *Assessing Writing, 10*, 157-173. doi: 10.1016/j.asw.2005.08.002

Magone, M. E., Cai, J. Silver, E. A., & Wang, N. (1994). Validating the cognitive complexity and content quality of a mathematics performance assessment. *International Journal of Educational Research*, *21(3)*, 317-340. http://dx.doi.org/10.1016/S0883-0355(06)80022-4

Marzano, R. J., Bandt, R. S., Hughes, C. S., Jones, B. F., Presseisen, B. Z., Ranken, S.C . & Suhor, C. (1988). *Dimensions of thinking: A framework for curriculum and instruction*. Alexandria, Virginia: The Association for Supervision and Curriculum Development.

Maxcy, J. (2003). Pragmatic threads in mixed methods research in the social sciences. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 51-89). Thousand Oaks, CA: Sage Publications.

Mertens, D. M. (2005). An introduction to research. In *Research and evaluation in education and psychology: integrating diversity with quantitative, qualitative and mixed methods* (pp. 8- 39). Thousand Oaks, CA: Sage Publications.

Mayring, P. (2000). *Qualitative content analysis*. Retrieved from http://217.160.35.246/fqs-texte/2-00/2-00mayring-e.pdf.

Mazzeo, C. (2001). Frameworks of state: Assessment policy in historical perspective. *Teachers College Record, 103*(3), 367-397.

McKown, B. A., & Barnett, C. L. (2007). *Improving reading comprehension through higher-order thinking skills.* Unpublished Manuscript. Retrieved from http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/28/09/00.pdf.

Medical College Admission Test (MCAT). (2012). *Medical College Admission Test (MCAT®)*. Retrieved from https://www.aamc.org/students/applying/mcat/.

Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives, 6*(13), 1-30.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955-965. doi: 10.1037/0003-066X.30.10.955

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11) 1012-1027. doi: 10.1037/0003-066X.35.11.1012

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11. doi: 10.3102/0013189X018002005

Messick, S. (1990). *Validity of test interpretation and use*. Retrieved from http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/14/82/4e.pdf.

Messick, S. (1994a). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23. doi: 10.3102/0013189X023002013

Messick, S. (1994b). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning*. Research report RR-94-45

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi: 10.1037/0003-066X.50.9.741

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*, 35-44.

Meulman, J. J., Van Der Kooij, A. J., & Heiser, W. J. (2004). *Principal component analysis with nonlinear optimal scaling transformations for ordinal and nominal data. Chapter 3*. Retrieved from http://atgstg01.sagepub.com/upm-data/5040_Kaplan_Final_Pages_Chapter_3.pdf.

Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 2,* 367-378. doi: 10.1177/01466210022031813

Ministry of Education (2009). *The Ontario curriculum: Grades 1-8*. Retrieved from www.edu.gov.on.ca

Morgan, D. L. (2007). Paradigms lost and pragmatism regained: Methodological implications of combining qualitative and quantitative methods. *Journal of Mixed Methods Research, 1*(1), 48-62. doi: 10.1177/2345678906292462

Moseley, D., Elliot, J., Gregson, M., & Higgins, S. (2005). Thinking skills frameworks for use in education and training. *British Educational Research Journal,* (3), 367-390. doi: 10.1080/01411920500082219

Mosenthal, P. B. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology, 88*(2), 314-332. doi: 10.1037/0022-0663.88.2.314

Mosenthal, P. B., & Kirsch, I. S. (1991a). Extending prose comprehension through knowledge modeling. *Journal of Reading, 35*, 58-61.

Mosenthal, P. B., & Kirsch, I. S. (1991b). Toward an explanatory model of document literacy. *Discourse Processes, 14*, 147-180. doi: 10.1080/01638539109544780

Mosenthal, P. B., Kirsch, I. S. (1992). Types of document knowledge: From structures to strategies. *Journal of Reading, 36*, 64-67.

Moss, P. A., & Koziol, S. M. (1991). Investigating the validity of a locally developed critical thinking test. *Educational Measurement: Issues and Practice*, *10*(3), 17-22. doi: 10.1111/j.1745-3992.1991.tb00199.x

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011: Assessment framework*. Retrieved from http://timssandpirls.bc.edu/pirls2011/downloads/PIRLS2011_Framework.pdf

Mullis, I. V. S., Martin, M. O., & Sainsbury, M. (2013). PIRLS 2016 reading framework. In I. V. S. Mullis & M. O. Martin (Eds.), *PIRLS 2016 assessment framework*. TIMSS and PIRLS International Study Center, Chestnut Hill, MA.

Newman, F. M. (1990). Higher order thinking in teaching social studies: A rationale for the assessment of classroom thoughtfulness. *Journal of Curriculum Studies, 22*, 41-56. doi: 10.1080/0022027900220103

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice, 28*(1), 3-9. doi: 10.1111/j.1745-3992.2009.01132.x

Nussbaum, E. M., Hamilton, L. S., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: IV. NELS: 88 Science Achievement to 12th Grade. *American Educational Research Journal, 34(1)*, p. 151-173. doi: 10.3102/00028312034001151

Nunnally, J. C., & Bernstein, I. H. (1994), Psychometric Theory, Third Edition, New York: McGraw-Hill.

OECD. (n.d.). *OECD Programme for International Student Assessment*. Retrieved from http://www.pisa.oecd.org/pages/0,3417,en_32252351_32235907_1_1_1_1_1,00.html.

OECD. (2003). *The PISA 2003 assessment framework*. Retrieved from http://browse.oecdbookshop.org/oecd/pdfs/browseit/9603051E.PDF.

OECD. (2008). *Assessment for learning: Formative assessment*. Retrieved from http://www.oecd.org/site/educeri21st/40600533.pdf

OECD (2009). *PISA 2009 assessment framework: Key competencies in reading, mathematics and science*. Retrieved from http://www.oecd.org/pisa/pisaproducts/44455820.pdf.

OECD (2013). *Draft literacy reading framework*. Retrieved from http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Reading%20Framework%20.pdf

Ontario Ministry of Education (2006). *The Ontario curriculum, grades 1-8: Language, 2006 (revised).* Retrieved from http://www.edu.gov.on.ca/eng/curriculum/elementary/language18currb.pdf.

Paul, L., & Elder, L. (2005). *A guide for educator to critical thinking competency standards: standards principles, performance indicators, and outcomes with a critical thinking master rubric*. Retrieved online at www.criticalthinking.org.

Paul, R., & Nosich, G. M. (1992). *A model for the national assessment of higher order thinking*. Retrieved online from www.criticalthinking.org.

Plucker, J. A., & Makel, M. C. (2010). Assessment of creativity. In J. C. Kaufman & R. J. Sternberg, *the Cambridge handbook of creativity*. New York, N. Y.; Cambridge University Press.

Popham, W. J. (1997). Consequential validity: Right concern - wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13. doi: 10.1111/j.1745-3992.1997.tb00586.x

Power, B. (2007). *What are the seven reading comprehension strategies*? Retrieved from http://www.choiceliteracy.com/articles-detail-view.php?id=85

Principles for Fair Student Assessment Practices for Education in Canada. (1993). Edmonton, Alberta: Joint Advisory Committee. (Mailing Address: Joint Advisory Committee, Centre for Research in Applied Measurement and Evaluation, 3-104 Education Building North, University of Alberta, Edmonton, Alberta, T6G 2G5).

Quellmalz, E. S. (1985). Needed: Better methods for testing higher-order thinking skills. *Educational Leadership*, 29-35.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13-16. doi: 10.1111/j.1745-3992.1998.tb00827.x

Rijmen. F. (2011). *Hierarchical factor item response theory models for PIRLS: Capturing clustering effects at multiple levels*. IERI Monograph article. Retrieved from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_04_Chapter_3.pdf.

Rosenblatt, L. M. (1982). The literary transaction: Evocation and response. *Theory into Practice, 21*(4), 268-277.

Rosenblatt, L. M. (1985).Viewpoints: Transaction versus interaction: A terminological rescue operation. *Research in the Teaching of English, 19*(1), 96-107.

Rosenblatt, L. M. (1986). The aesthetic transaction. *Journal of Aesthetic Education, 20*(4), 122-128.

Rosenblatt, L. M. (1993). The transactional theory: Against dualisms. *College English, 55*(4), 377-386.

Rosenblatt, L.M. (1995). Continuing the conversation: A clarification. *Research in the Teaching of English, 29*(3), 349-354.

Rousseau, M., Simon, M., Bertrand, R., & Hachey, K. (2010). Reporting missing data: A study of selected articles published from 2003-2007. *Quality & Quantity, 46*(5), 1393-1406.

Ruddell, R. B., & Unrau, N. J. (2004). *Theoretical models and processes of reading* (5th ed.). Newark, DE: Reading Association, Inc.

SAS (2010). *Base SAS 9.2 Procedures Guide: Statistical Procedures* (3rd Edition). Retrieved from http://support.sas.com/documentation/cdl/en/procstat/63104/PDF/default/procstat.pdf.

Schraw, G., & Robinson, D. R. (2011). *Assessment of higher order thinking skills*. University of Texas, at Austen: Texas, IAP–Information Age Publishing, Inc.

National Council for Excellence in Critical Thinking (1987). *Definition of critical thinking*. Presentation by Michael Scriven and Richard Paul at the 8th Annual International Conference on Critical Thinking and Education Reform. Retrieved from http://www.criticalthinking.org/pages/defining-critical-thinking/766.

Serafini, F. (2001). Three paradigms of assessment: Measurement, procedure, and inquiry. *The Reading Teacher, 54*(4), 384-393.

Shaffer, D. R., Wood, E., & Willoughby, T. (2005). Developmental psychology: Childhood and adolescence (2nd Ed.). Toronto, Ontario: Thomson, Nelson.

Shepard, L. A. (2000). *The role of classroom assessment in teaching and learning*. CSE Technical report 517. Retrieved from http://datause.cse.ucla.edu/DOCS/las_rol_2000.pdf

Short, K., & Burke, C. (1994). *Curriculum as inquiry*. Paper presented at the fifth whole language umbrella conference, San Diego, CA.

Simner, M. L. (2000). *A joint position statement by the Canadian Psychological Association and the Canadian Association of School Psychologists on the Canadian press coverage of the province-wide achievement test results*. Retrieved from http://www.cpa.ca/documents/joint_position.html

Simon, M., Ercikan, K. & Rousseau, M. (2012). Improving large-scale assessment in education. In M. Simon, K. Ercikan, & M. Rousseau, *Improving large-scale assessment in education: Theories, issues and practice*. New York, NY: Taylor & Francis.

Sireci, S. G. (1998a). Gathering and analyzing content validity data. *Educational Assessment, 5*(4), 299-321. doi: 10.1207/s15326977ea0504_2

Sireci, S. G. (1998b). The construct of content validity. *Social Indicators Research, 45*, 83-117. doi: 10.1023/A:1006985528729

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R.W. Lissitz, *The concept of validity: Revisions, new directions, and applications* (pp. 19-37). Charlotte, NY: Information Age Publishing Corp.

Skinner, B. F. (1977). *Operant Behaviour*. In W. E. Honig and J. E. R. Staddon (Eds.), Handbook of operant behaviour. New York, N. Y.: Prentice Hall.

Stanovich, K. E. (1980). Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly, 16* (1), 32-71.

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation*, 7(17). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=17

Taylor, A. R., & Tubianosa, T. S. (2001). *Student Assessment in Canada: Improving the learning environment through effective education* (SAEE Research Series No.9). Kelowna, BC: Society for the Advancement of Excellence in Education.

Tindal, G. & Nolet, V. (1995). Curriculum-based measurement in middle and high schools: critical thinking skills in content areas. *Focus on Exceptional Children , 27*(7), 1-22.

Topping, K. J., & Bryce, A. (2004). Cross-age peer tutoring of reading and thinking: Influence on thinking skills. *Educational Psychology, 24*(5), 595-621. doi: 10.1080/0144341042000262935

Tracey, D. H., & Morrow, L. M. (2012). *Lenses on reading: An introduction to theories and models* (2nd Ed.). New York, NY: The Guilford Press.

Treffinger, D. J., Young, G.C., Selby, E. C., & Shepardson, C. (2002). *Assessing creativity: A guide for educators*. Retrieved from http://www.gifted.uconn.edu/NRCGT/reports/rm02170/rm02170.pdf

Underbakke, M., Borg, J. M., & Peterson, D. (1993). Researcher and developing the knowledge base for teaching higher order thinking. *Theory into Practice, 32*(3), 138-146.

Ungerleider, C. (2003). Large-scale student assessment: Guidelines for policymakers. *International Journal of Testing, 3*(2), 119-128.

Volante, L. (2007). Educational quality and accountability in Ontario: Past, present, and future. *Canadian Journal of Educational Administration and Policy, 58*, 1-21.

Volante, L. (2010). Standards-based reform: can we do better? *Education Canada, 47* (1), 54-56.

Volante, L. & Jaafar, S. B. (2008). Profiles of education assessment systems Worldwide: Educational assessment in Canada. *Assessment in Education: Principles, Policy, & Practice, 15*(2), 201-210.

Wolfe, R., Childs, R., & Elgie, S. (2004). *Final report of the external evaluation of EQAO's assessment proces*s. Ontario Institute for Studies in Education of the University of

Toronto. Retrieved online from http://www.ontla.on.ca/library/repository/mon/8000/244807.pdf.

Wolming, S., & Wikström, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy & Practice, 17*(2), 117-132. doi: 10.1080/09695941003693856

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, D. (2004). TESTFACT. [Computer program] Lincolnwood, IL: Scientific Software International.

Woods, C. M. (2002). Factor analysis of scales composed of binary items: Illustration with the Maudsley Obsessional Compulsive Inventory. *Journal of Psychopathology and Behavioural Assessment, 24*(4), 215-223. doi: 10.1023/A:102077931134

Zhang, Y., & Wildemuth, B. M. (2009). *Qualitative analysis of content*. In B. Wildmuth (Ed.), Applications of social science research methods to questions in information and library science (pp.308-319). Westport, CT: Libraries Unlimited. PDF.

Zumbo, B. (2009). Validity as contextualized and pragmatic explanation and its implications for validation practice. In R.W. Lissitz, *The concept of validity: Revisions, new directions, and applications* (pp.65-82). Charlotte, NY: Information Age Publishing Corp.

Appendix A

University of Ottawa Ethics

File Number: 09-10-33 **Date (mm/dd/yyyy):** 10/15/2010

Université d'Ottawa
Bureau d'éthique et d'intégrité de la recherche

University of Ottawa
Office of Research Ethics and Integrity

Ethics Approval Notice

Social Science and Humanities REB

Principal Investigator / Supervisor / Co-investigator(s) / Student(s)

First Name	Last Name	Affiliation	Role
Marielle	Simon	Faculty of Education	Supervisor
Krystal	Hachey	Faculty of Education	Student Researcher

File Number: 09-10-33

Type of Project: Secondary use of data for the completion of a Doctorate Thesis

Title: Large-Scale Assessments and Higher Order Thinking

Approval Date (mm/dd/yyyy)	Expiry Date (mm/dd/yyyy)	Approval Type
10/15/2010	10/14/2011	I
		(I: Conditional Approval)

Special Conditions / Comments:

The abovementioned research project has been reviewed by the Social Sciences and Humanities Research Ethics Board.

Before research using the requested data may begin, letters of permission from the Pan-Canadian Assessment Program (PCAP) and the Education Quality and Accountability Office (EQAO) will be required on file.

These letters of permission can be sent to our offices as soon as they become available.

File Number: 09-10-33 Date (mm/dd/yyyy): 09/21/2011

Université d'Ottawa University of Ottawa
Bureau d'éthique et d'intégrité de la recherche Office of Research Ethics and Integrity

Ethics Approval Notice

Social Science and Humanities REB

Principal Investigator / Supervisor / Co-investigator(s) / Student(s)

First Name	Last Name	Affiliation	Role
Marielle	Simon	Education / Education	Supervisor
Krystal	Hachey	Education / Education	Student Researcher

File Number: 09-10-33

Type of Project: Secondary use of data

Title: Large-Scale Assessments and Higher Order Thinking

Renewal Date (mm/dd/yyyy)	Expiry Date (mm/dd/yyyy)	Approval Type
10/15/2011	10/14/2012	Ia

(Ia: Approval, Ib: Approval for initial stage only)

Special Conditions / Comments:
N/A

Appendix B

EQAO Scoring Rubric

Code	Descriptor
	■ blank: nothing written or drawn in the space provided
	■ illegible: cannot be read; completely crossed out/erased; not written in English ■ irrelevant content: does not attempt assigned question (e.g., comment on the task, drawings, "?", "!", "I don't know") ■ off topic: no relationship of written work to the question
10	■ response does not refer to ideas and information from the reading selection ■ response indicates a significant misunderstanding of the reading selection; ideas and information from the reading selection are inaccurate
20	■ response addresses only part of the question ■ response is developed with limited support; ideas and information from the reading selection are minimal, vague and/or irrelevant
30	■ response addresses the complete question ■ response is developed with some accurate, specific and relevant ideas and information from the reading selection; some ideas and information are inaccurate, vague and/or irrelevant
40	■ response addresses the complete question ■ response is developed with accurate, specific and relevant ideas and information from the reading selection

Appendix C

PCAP Performance-Level Descriptors (CMEC, 2007)

Performance-Level Descriptors

	Level 1	Level 2	Level 3
Comprehension	The reader demonstrates partial understanding: • identifying limited and/or simple aspects of the text • recognizing direct vocabulary, concrete details, and directed stated ideas	The reader demonstrates clear understanding: • identifying both explicit and implicit aspects of the text • recognizing some abstract language, details, and implied ideas in the text	The reader demonstrates a nuanced understanding: • identifying principal ideas and important details • recognizing aspects of style, specific selections of detail, organization, and/or complexity in the text
Interpretation	The reader demonstrates limited interpretation of the text: • providing a simplified and/or general perspective of aspects of the text • relying on stated conclusions from the text to inform meaning, and/or connecting aspects of the text to each other	The reader demonstrates reasonable interpretation of the text: • noting relationships and integrating these elements to develop a reasonable perspective • relying on explicit and inferred support	The reader demonstrates thoughtful or insightful interpretation of the text: • synthesizing several elements of the text or expressing thoughtful analysis of a significant element of the text • relying on subtle relationships among elements and ideas to develop a thoughtful or insightful perspective
Personal Response	The reader demonstrates a limited personal response to the text: • expressing personal connections to minor details that are valid but simplistic, vague, or common • providing a response that is supported with little or no specific, textual, or personal references	The reader demonstrates appropriate personal response to the text: • expressing personal connections linked to key ideas • supporting a response with examples, general explanations, or statements that are direct or clearly inferred from the text or personal knowledge	The reader demonstrates elaborated personal response to the text: • expressing personal implications and insights • supporting a response with specific details, examples, precise explanations, or thoughtful justifications
Critical Response	The reader demonstrates tangential or simplistic critical response: • commenting on minor aspects or providing a simple evaluation • providing little support beyond repetition or personal authority	The reader demonstrates appropriate critical response: • expressing a perspective on the text based upon some social, cultural, or literary awareness • expressing general and plausible commentary on the quality, currency, or relevance of the text • providing supportive appropriate evidence from the text or other sources	The reader demonstrates significant critical response: • evaluating the text and the relationships of elements based on social, cultural, and/or literary knowledge • providing evaluation of the quality, sources, currency, or relevance of text issues, themes, and/or elements of style • supporting response with reference to voice, style, stance of the writer, organization and structure, and analogies or illustrations from the world at large.

Appendix D

Excel Tabs

The following table provides the excel tab format for coding the assessment framework and scoring guide.

Instructions
1. Read the assessment framework and scoring guide 2. Determine if there is evidence that references any assessment perspective. 3. There is space to address aspects of both assessment perspectives if they are present.

ASSESSMENT PERSPECTIVE

Assessment		**EQAO**
Assessment Perspective		
Evidence		
Explanation		
	Page number	
Assessment Perspective		
Evidence		
Explanation		
	Page number	

THINKING SKILLS

Assessment		**EQAO**
Thinking Skill		
Evidence		
Explanation		
	Page Number	
Thinking Skill		
Evidence		
Explanation		
	Page Number	
Thinking Skill		
Evidence		
Explanation		
	Page Number	
Thinking Skill		
Evidence		
Explanation		
	Page Number	

The following table provides the excel tab format for coding the test booklets (this one was developed for EQAO)

Instructions

1. Read the reading passage.
2. Examine each question related to the reading passage.
3. Determine, for each question, what thinking skills the question is assessing.

Legend

Thinking skills: What thinking skills would students need to use to answer the question?

Test Booklet	Item	Page number	Thinking Skills	Evidence	Other
1	1	2			
1	2	2			
1	3	2			
1	4	2			
1	5	3			
1	6	3			
1	7	5			
1	8	5			
1	9	5			
1	10	5			
1	11	6			
1	12	6			
2	1	3			
2	2	3			
2	3	3			
2	4	3			
2	5	3			
2	6	4			
2	7	4			
2	8	4			
2	9	4			
2	10	4			
2	11	5			
2	12	5			

Test Booklet	Item	Page number	Thinking Skills	Evidence	Other
2	17	8			
2	18	8			
2	19	8			
2	20	9			
2	21	9			
2	23	15			
2	24	15			
2	25	15			
2	26	15			
2	27	16			
2	28	16			

Appendix E

Cohen's Kappa Output from SPSS

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Rater1 * Rater2	35	100.0%	0	0.0%	35	100.0%

Rater1 * Rater2 Crosstabulation

Count

		Rater2			
		Lower Level Thinking	Critical Thinking	Creativity	Lower Level thinking or Critical Thinking
Rater1	Lower Level Thinking	3	0	0	0
	Critical Thinking	0	12	0	0
	Lower Level thinking or Critical Thinking	0	0	0	12
	Critical Thinking and Creativity	0	3	4	0
Total		3	15	4	12

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.678	.087	7.052	.000
N of Valid Cases		35			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Inter-Rater Results by Item

Test Booklet	Item	Page number	Principal Investigator Thinking Skills	Rater 1 Thinking Skills
1	1	2	**Lower Level Thinking**	**Lower Level Thinking**
1	2	2	**Critical thinking**	**Critical Thinking**
1	3	2	**Critical Thinking or Lower Level Thinking**	**Critical Thinking or Lower Level Thinking**
1	4	2	**Critical Thinking or Lower Level Thinking**	**Critical Thinking or Lower Level Thinking**
1	5	3	**Critical Thinking**	**Critical Thinking**
1	6	3	Creativity and **Critical Thinking**	**Critical Thinking**

Test Booklet	Item	Page number	Principal Investigator Thinking Skills	Rater 1 Thinking Skills
1	7	5	**Critical Thinking**	**Critical Thinking**
1	8	5	**Critical Thinking or Lower Level Thinking**	**Critical Thinking or Lower Level Thinking**
1	9	5	**Critical Thinking or Lower Level Thinking**	**Critical Thinking or Lower Level Thinking**
1	10	5	**Critical Thinking or Lower Level Thinking**	**Critical Thinking or Lower Level Thinking**
1	11	6	Creativity and **Critical Thinking**	**Creativity**

Test Booklet	Item	Page number	Principal Investigator Thinking Skills	Rater 1 Thinking Skills
1	12	6	**Creativity** and Critical Thinking	**Creativity**
2	1	3	**Critical Thinking**	**Critical Thinking**
2	2	3	**Lower Level Thinking**	**Lower Level Thinking**
2	3	3	**Critical Thinking**	**Critical Thinking**
2	4	3	**Critical Thinking or Lower Level Thinking**	**Critical Thinking or Lower Level Thinking**
2	5	3	**Lower Level Thinking**	**Lower Level Thinking**

Test Booklet	Item	Page number	Principal Investigator Thinking Skills	Rater 1 Thinking Skills
2	6	4	**Critical Thinking or Lower Level Thinking**	**Critical Thinking or Lower Level Thinking**
2	7	4	**Critical Thinking**	**Critical Thinking**
2	8	4	**Critical Thinking**	**Critical Thinking**
2	9	4	**Critical Thinking**	**Critical Thinking**
2	10	4	**Critical Thinking or Lower Level Thinking**	**Critical Thinking or Lower Level Thinking**
2	11	5	Creativity and **Critical Thinking**	**Critical Thinking**
2	12	5	**Critical Thinking**	Creativity or **Critical Thinking**

Test Booklet	Item	Page number	Principal Investigator Thinking Skills	Rater 1 Thinking Skills
2	17	8	**Critical Thinking or Lower Level Thinking**	**Critical Thinking or Lower Level Thinking**
2	18	8	**Critical Thinking**	**Critical Thinking**
2	19	8	**Critical Thinking**	**Critical Thinking**
2	20	9	**Creativity** and Critical Thinking	**Creativity**
2	21	9	**Creativity** and Critical Thinking	**Creativity**
2	23	15	**Critical Thinking or Lower Level Thinking**	**Critical Thinking or Lower Level Thinking**
2	24	15	**Critical Thinking or Lower Level Thinking**	**Critical Thinking or Lower Level Thinking**

Test Booklet	Item	Page number	Principal Investigator Thinking Skills	Rater 1 Thinking Skills
2	25	15	**Critical Thinking**	**Critical Thinking**
2	26	15	**Critical Thinking or Lower Level Thinking**	**Critical Thinking or Lower Level Thinking**
2	27	16	**Critical Thinking**	**Critical Thinking**
2	28	16	Creativity and **Critical Thinking**	**Critical Thinking**

Appendix F

EQAO Thinking Skills: Analyzing the Assessment Framework

Thinking Skills and Document Processing Strategies	Evidence	Reason
Critical thinking, integrate, and generate strategies	"…understanding implicitly stated information and ideas (making inferences)…" (EQAO, 2007; p. 11).	The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response.
	"…synthesizing textually explicit information and ideas that are found in more than one place in the text… (Pearson & Johnson, 1978; Raphael, 1986)" (EQAO, 2007; p. 12).	The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response. Thus, the student must "...generate a set of characteristics which are not memorized" (Haladyna, 1997; p.21)
	"In the reading component, students use reading strategies to interact with a variety of	The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the

Thinking Skills and Document Processing Strategies	Evidence	Reason
	fiction and non-fiction reading selections to construct an understanding of the meaning of the texts." (EQAO, 2007; p. 10)	appropriate response.
	"They must also connect their understanding of the text to their personal knowledge and experience" (EQAO, 2007; p. 10)	The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response.
	"...responding to reading by making connections between information and ideas in a reading selection and the reader's personal knowledge and experience (interpreting a reading selection by integrating its information and ideas with personal	The student must generate or select the relevant characteristics to connect their understanding to the text. The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response

Thinking Skills and Document Processing Strategies	Evidence	Reason
	knowledge and experience).” (EQAO, 2007; p. 11)	
	“predicting textually implicit ideas using background knowledge and experience together with information in the text…(Pearson & Johnson, 1978; Raphael, 1986)” (EQAO, 2007; p. 12).	The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response.
	“…providing scriptally implicit ideas and information by making personal connections with background knowledge and experience (Pearson & Johnson, 1978; Raphael, 1986)” (EQAO, 2007; p. 12).	The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response.

Thinking Skills and Document Processing Strategies	Evidence	Reason
Problem solving, integrate, and generate strategies	"During reading, students may use 'curing systems' - that is, clues from context or from their understanding of language structures and/or letter-sound relationships - to help them solve unfamiliar words, and comprehension strategies to help them make meaning of the text." (EQAO, 2007; p. 10).	Providing a problem solution.
Creativity	None apparent	
Lower level thinking and locate strategies	"…Students are asked to demonstrate their understanding of explicit (directly stated)…" (EQAO, 2007; p. 10).	The ability to understand directly stated information demands only routine applications of skills.

Thinking Skills and Document Processing Strategies	Evidence	Reason
	"…analyzing textually explicit information and ideas that are directly stated in one part of the text…(Pearson & Johnson, 1978; Raphael, 1986)" (EQAO, 2007; p. 12)	The ability to understand directly stated information demands only routine applications of skills.

Appendix G

EQAO Thinking Skills: Analyzing the Test Booklets

Test Booklet	Item	Page number	Thinking Skills and Document Processing Strategies	Evidence
1	1	2	Lower level thinking and locate strategies	Individuals are asked to restate information in the text.
1	2	2	Critical thinking and integrate strategies	The information is not explicitly stated in the text. Instead, the student needs to correctly interpret the information presented (i.e., why are puppies introduced to busy areas?) and compare what is already established in the text.
1	3	2	Critical thinking or lower level thinking and cycle strategies	The information is not explicitly stated in the text. The student needs to correctly interpret the information given (i.e., what does 'residential areas' mean') based on the context of the text. This item could also be tapping into lower level thinking if the student had been previously exposed to the word 'residential'. The word is not unique to the text. That is, the student would not need to use the text to understand the word. Students must conduct several locate strategies to determine the definition of the word.

Test Booklet	Item	Page number	Thinking Skills and Document Processing Strategies	Evidence
1	4	2	Critical thinking or lower level thinking and cycle strategies.	The information is not explicitly stated in the text. The student must identify relevant characteristics of the text to determine how to best summarize the information presented (i.e., what could be an appropriate title for a specific paragraph?). Additionally, if the student understood the meaning of "simple commands" they could also identify the appropriate title of "basic commands". Students must conduct several locate strategies to determine an appropriate title.
1	5	3	Critical thinking and integrate strategies.	The information is not explicitly stated in the text. The student needs to correctly interpret the information given (i.e., why does a guide dog need to concentrate?). Students need to use information in the text, compare information in the text, and make a judgement.
1	6	3	Critical thinking, creativity, and generate strategies.	The information is not explicitly stated in the text. This item taps into critical thinking as the student must use information in the text to make a judgement, as well as provide their own ideas. One could consider the inclusion of personal views as an element of

Test Booklet	Item number	Page number	Thinking Skills and Document Processing Strategies	Evidence
				creativity. Although this is a limited view of creativity, it still does provide students an opportunity to include their novel ideas. Due to the use of previous knowledge, this item also taps into generate strategies.
1	7	5	Critical thinking and integrate strategies	The information is not explicitly stated in the text. The student must identify and compare relevant characteristics of the text to determine how to best summarize the information presented (i.e., what is the text mainly about?).
1	8	5	Critical thinking or lower level thinking and cycle strategies.	The information is not explicitly stated in the text. The student needs to correctly interpret the information given (i.e., Based on the information provided what does squall mean?) based on the context of the text. This item could also be tapping into lower level thinking if the student had been previously exposed to the word 'squall'. The word is not unique to the text. That is, the student would not need to use the text to understand the word. Students would have to conduct several locate strategies to determine the definition of the term.

Test Booklet	Item number	Page number	Thinking Skills and Document Processing Strategies	Evidence
1	9	5	Critical thinking or lower level thinking and cycle strategies	The information is not explicitly stated in the text. The student needs to correctly interpret the information given (i.e., Based on the information provided what does the information in the parentheses mean?) based on the context of the text. This item could also be tapping into lower level thinking if the student had been previously exposed to the use of parentheses. The rules of parentheses are not unique to the text. That is, the student would not need to use the text to understand. Students would have to conduct several locate strategies to determine the reason for the parentheses.
1	10	5	Critical thinking or lower level thinking and cycle strategies	The information is not explicitly stated in the text. The student needs to correctly interpret the information given (i.e., Based on the information provided what does looters mean?) based on the context of the text. This item could also be tapping into lower level thinking if the student had been previously exposed to the word ‘looters’. The word is not unique to the text. That is, the student would not need to use the text to understand the word. Students would have to conduct several locate strategies to determine the

Test Booklet	Item number	Page number	Thinking Skills and Document Processing Strategies	Evidence
				definition of the term.
1	11	6	Critical thinking, creativity, and generate strategies	The information is not explicitly stated in the text. This item taps into critical thinking as the student must use information in the text to make a judgement, as well as provide their own ideas. Although this is a limited view of creativity, it still does provide students an opportunity to include their potential novel ideas. Due to the use of previous knowledge, this item also taps into generate strategies.
1	12	6	Critical thinking, creativity, and generate strategies	The information is not explicitly stated in the text. This item taps into critical thinking as the student must use information in the text to make a judgement, as well as provide their own ideas (i.e., how to protect the Hamilton from environmental threats). One could consider the inclusion of personal views as an element of creativity. Although this is a limited view of creativity, it still does provide students an opportunity to include their potential novel ideas. Due to the use of previous knowledge, this item also taps into

Test Booklet	Item	Page number	Thinking Skills and Document Processing Strategies	Evidence
				generate strategies.
2	1	3	Critical thinking and integrate strategies	The information is not explicitly stated in the text. The student needs to correctly interpret and compare the information given (i.e., Based on the information the first conversation occurred…).
2	2	3	Lower level thinking and locate strategies	Individuals are asked to restate information in the text.
2	3	3	Critical thinking and integrate strategies	The information is not explicitly stated in the text. The student must identify and compare relevant characteristics of the text to determine how to best summarize the information presented (i.e., what is the text mainly about?).
2	4	3	Critical thinking or lower level thinking and cycle strategies	The information is not explicitly stated in the text. The student needs to correctly interpret the information given (i.e., Based on the information provided what does founded mean?) based on the context of the text. This item could also be tapping into

Test Booklet	Item number	Page number	Thinking Skills and Document Processing Strategies	Evidence
				lower level thinking if the student had been previously exposed to the word 'founded'. The word is not unique to the text. That is, the student would not need to use the text to understand the word. Students must perform several locate strategies to determine the definition of the term.
2	5	3	Lower level thinking and locate strategies	Individuals are asked to restate information in the text.
2	6	4	Critical thinking or lower level thinking and cycle strategies	The information is not explicitly stated in the text. The student needs to correctly interpret the information given (i.e., Based on the information provided what does moved mean?) based on the context of the text. This item could also be tapping into lower level thinking if the student had been previously exposed to the word 'moved'. The word is not unique to the text. That is, the student would not need to use the text to understand the word. Students must perform several locate strategies to determine the definition of the term.

Test Booklet	Item	Page number	Thinking Skills and Document Processing Strategies	Evidence
2	7	4	Critical thinking and integrate strategies	The student needs to correctly interpret and compare relevant information (i.e., Based on the information provided, why do the members take the jars?).
2	8	4	Critical thinking and integrate strategies	The student needs to correctly interpret and compare relevant information (i.e., Based on the information provided, Hannah can be described as…)
2	9	4	Critical thinking and integrate strategies	The student needs to correctly interpret and compare the relevant information (i.e., Based on the information provided, Hannah is described as an amazing kid because?).
2	10	4	Critical thinking or lower level thinking and cycle strategies	Individuals are asked to identify what type of text (e.g., journal) the reading passage is. The student would have to try and determine, based on the information provided, how the text is written. On the other hand, if the student had been previously exposed to understanding the different types of texts, then the question would elicit lower level thinking. Students must perform several locate strategies to determine type of text.
2	11	5	Critical thinking,	The information is not explicitly stated in the text. This item taps into critical thinking as

Test Booklet	Item number	Page number	Thinking Skills and Document Processing Strategies	Evidence
			creativity, and generate strategies	the student must use information in the text to make a judgement, as well as provide their own ideas. One could consider the inclusion of personal views as an element of creativity. Although this is a limited view of creativity, it still does provide students an opportunity to include their potential novel ideas. Due to the use of previous knowledge, this item also taps into generate strategies.
2	12	5	Critical thinking and integrate strategies	The student is asked to reflect on the text, evaluate, compare, and make a judgement.
2	17	8	Critical thinking or lower level thinking and cycle strategies	The student is asked to reflect on what the apostrophe means in the context of the sentence. If the student is aware of the meaning behind the use of an apostrophe, then the question would elicit lower level thinking. Students must perform several locate strategies to determine use of the apostrophe.
2	18	8	Critical thinking and integrate strategies	The student needs to correctly interpret and compare the relevant information (i.e., Based on the information provided, the pocket watch was lost because?).

Test Booklet	Item	Page number	Thinking Skills and Document Processing Strategies	Evidence
2	19	8	Critical thinking and integrate strategies	The student is asked to reflect on the text, evaluate, compare, and make a judgment.
2	20	9	Critical thinking, creativity, and generate strategies	The information is not explicitly stated in the text. This item taps into critical thinking as the student must use information in the text to make a judgement, as well as provide their own ideas. One could consider the inclusion of personal views as an element of creativity. Although this is a limited view of creativity, it still does provide students an opportunity to include their potential novel ideas. Due to the use of previous knowledge, this item also taps into generate strategies.
2	21	9	Critical thinking, creativity, and generate strategies	The information is not explicitly stated in the text. This item taps into critical thinking as the student must use information in the text to make a judgement, as well as provide their own ideas. One could consider the inclusion of personal views as an element of creativity. Although this is a limited view of creativity, it still does provide students an opportunity to include their potential novel ideas. Due to the use of previous knowledge,

Test Booklet	Item number	Page number	Thinking Skills and Document Processing Strategies	Evidence
				this item also taps into generate strategies.
2	23	15	Critical thinking or lower level thinking and cycle strategies	Students are asked to determine how the reader can tell if someone is speaking. If aware of the use of quotation marks, then the question would elicit lower level thinking. However, if the student did not understand the use of quotation marks, they would have to reflect on the text. Students would have to perform several locate strategies to determine the purpose of the quotation marks.
2	24	15	Critical thinking or lower level thinking and cycle strategies	Students are asked to determine the use of an exclamation mark. If aware of the use of an exclamation mark, then the question would elicit lower level thinking. However, if the student did not understand the use of exclamation marks, they would have to reflect on the text. Students would have to perform several locate strategies to determine the purpose of the exclamation mark.
2	25	15	Critical thinking and integrate strategies	The student needs to correctly interpret and compare the relevant information (i.e., Based on the information provided, Su-Ling can be described as…)

Test Booklet	Item	Page number	Thinking Skills and Document Processing Strategies	Evidence
2	26	15	Critical thinking or lower level thinking and cycle strategies	The student needs to correctly interpret the information given (i.e., Based on the information provided, what does 'practical and down-to-earth mean?). If the student had already been previously exposed to such a term, then the question would elicit lower level thinking. Students would have to perform several locate strategies to determine the definition of the term.
2	27	16	Critical thinking and integrate strategies	The student needs to correctly interpret and compare the relevant information (i.e., Based on the information provided, Devon speaks very little…)
2	28	16	Critical thinking, creativity, and generate strategies	The information is not explicitly stated in the text. This item taps into critical thinking as the student must use information in the text to make a judgement, as well as provide their own ideas. One could consider the inclusion of personal views as an element of creativity. Although this is a limited view of creativity, it still does provide students an opportunity to include their potential novel ideas. Due to the use of previous knowledge, this item also taps into generate strategies.

Appendix H

EQAO: The Nonlinear Factor Analysis of the 35 items (2 factors)

Item	Test Booklet	Type of Question	Thinking Skills	Factor 1: Lower Level Thinking	Factor 2: HOT Skills
1	1	Multiple choice	Lower Level Thinking	**0.383**	0.276
2	1	Multiple choice	Critical thinking	0.290	0.149
3	1	Multiple choice	Critical Thinking or Lower Level Thinking	**0.542**	0.005
4	1	Multiple choice	Critical Thinking or Lower Level Thinking	**0.355**	0.100
5	1	Open-ended	Critical Thinking	0.392	**0.577**
6	1	Open-ended	Creativity and Critical Thinking	0.375	**0.622**
7	1	Multiple choice	Critical Thinking	**0.387**	0.147
8	1	Multiple choice	Critical Thinking or Lower Level Thinking	**0.447**	0.165
9	1	Multiple choice	Critical Thinking or Lower Level Thinking	**0.381**	0.098
10	1	Multiple choice	Critical Thinking or Lower Level Thinking	**0.532**	-0.006
11	1	Open-ended	Creativity and Critical Thinking	0.071	**0.609**
12	1	Open-ended	Creativity and Critical Thinking	0.222	**0.667**

Item	Test Booklet	Type of Question	Thinking Skills	Factor 1: Lower Level Thinking	Factor 2: HOT Skills
1	2	Multiple choice	Critical Thinking	**0.550**	0.122
2	2	Multiple choice	Lower Level Thinking	**0.648**	0.180
3	2	Multiple choice	Critical Thinking	**0.581**	0.164
4	2	Multiple choice	Critical Thinking or Lower Level Thinking	**0.603**	0.079
5	2	Multiple choice	Lower Level Thinking	**0.410**	0.192
6	2	Multiple choice	Critical Thinking or Lower Level Thinking	**0.391**	0.295
7	2	Multiple choice	Critical Thinking	0.186	0.280
8	2	Multiple choice	Critical Thinking	0.345	**0.355**
9	2	Multiple choice	Critical Thinking	**0.526**	0.338
6	2	Multiple choice	Critical Thinking or Lower Level Thinking	**0.391**	0.295
7	2	Multiple choice	Critical Thinking	0.186	0.280
8	2	Multiple choice	Critical Thinking	0.345	**0.355**
9	2	Multiple choice	Critical Thinking	**0.526**	0.338
10	2	Multiple choice	Critical Thinking or Lower Level Thinking	**0.382**	0.178

Item	Test Booklet	Type of Question	Thinking Skills	Factor 1: Lower Level Thinking	Factor 2: HOT Skills
11	2	Open-ended	Creativity and Critical Thinking	0.209	**0.875**
12	2	Open-ended	Critical Thinking	0.306	**0.726**
17	2	Multiple choice	Critical Thinking or Lower Level Thinking	**0.486**	0.341
18	2	Multiple choice	Critical Thinking	**0.343**	0.248
19	2	Multiple choice	Critical Thinking	**0.376**	0.267
20	2	Multiple choice	Creativity and Critical Thinking	0.093	**0.917**
21	2	Multiple choice	Creativity and Critical Thinking	0.124	**0.899**
23	2	Multiple choice	Critical Thinking or Lower Level Thinking	**0.416**	0.345
24	2	Multiple choice	Critical Thinking or Lower Level Thinking	**0.375**	0.099
25	2	Multiple choice	Critical Thinking	**0.517**	0.407
26	2	Multiple choice	Critical Thinking or Lower Level Thinking	**0.566**	0.202
27	2	Open-ended	Critical Thinking	0.191	**0.877**
28	2	Open-ended	Creativity and Critical Thinking	0.212	**0.895**

Appendix I

PCAP Thinking Skills: Analyzing the Assessment Framework

Thinking Skills and Document Processing Strategies	Evidence	Reason
Critical thinking, integrate, and generate strategies	"…making inferences…" (CMEC, 2007; p. 2).	The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response.
	"They synthesize this information to draw conclusions about the broader meaning and intent of the text, that is, they consider relationships among elements and ideas in the text to construct deeper meaning and discern more significant implications." (CMEC, 2007; p. 4).	The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response.
	"Readers examine the text to develop an	The response is not explicitly stated in the

Thinking Skills and Document Processing Strategies	Evidence	Reason
	understanding of the relationships of discrete elements to the whole" (CMEC, 2007; p. 4).	text; instead the student must determine, based on the information presented, the appropriate response. Students synthesise information to understand elements of the text.
	"Readers stand apart from the text, considering it objectively and/or evaluating its quality and/or appropriateness to the world at large." (CMEC, 2007; p. 5).	The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response. Students synthesise information to understand elements of the text.
	"Interpretation: Readers examine the text to develop an understanding of the relationships of discrete	The response is not explicitly stated in the text; instead the student must determine,

Thinking Skills and Document Processing Strategies	Evidence	Reason
	elements to the whole. Readers use symbols, patterns, text features, and other elements to analyze the story in narrative texts, the general idea in information texts, and the arguments in persuasive text…they consider relationships among elements and ideas in the text to construct deeper meaning and discern more significant implications" (CMEC, 2007; p. 5).	based on the information presented, the appropriate response.
	"…making predictions and asking questions…synthesizing" (CMEC, 2007; p. 2).	The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response. Critical thinking is the process of evaluating information as well as predicting outcomes.

Thinking Skills and Document Processing Strategies	Evidence	Reason
	"Readers elaborate personal connections and reactions to text by providing some extended explanations, exampled, and supporting arguments. They state personal points of view and justify them." (CMEC, 2007; p. 4)	Students are required to reflect, compare, and make a judgement.
	"Critical Response: Readers stand apart from the text, considering it objectively and/or evaluating its quality and/or appropriateness to the world at large. Readers evaluate content, elements of style, or the author's stance. They evaluate choice of content, sources, quality, currency, or relevance of information. Readers support their responses by providing specific, appropriate details and evidence from the text and	PCAP's definition of critical response is similar to critical thinking, such that students are required to reflect, compare, and make a judgement.

Thinking Skills and Document Processing Strategies	Evidence	Reason
	other sources about issues, themes, and elements of style" (CMEC, 2007; p. 5).	
	"…evaluating its quality…" (CMEC, 2007; p. 5).	Critical thinking is the process of evaluating information as well as predicting outcomes.
	"Non-fiction texts also include those written to argue a particular perspective or point of view and those written to persuade the reader to take some particular stand or action (persuasion/argument)." (CMEC, 2007; p. 5).	Students are required to reflect, compare, and make a judgement.
	"Readers evaluate content, elements of style, or the author's stance. They evaluate choice of content, sources, quality, currency, or relevance of information." (CMEC, 2007; p. 5).	Students are required to reflect, compare, and make a judgement.

Thinking Skills and Document Processing Strategies	Evidence	Reason
	The task examines the degree to which students move beyond denotation to connotation, beyond explicit to inferred meaning, beyond concrete references and illustrations to abstraction and application, and beyond personal response to critical response." (CMEC, 2007; p. 6).	Students are required to reflect, compare, and make a judgement.
	"This knowledge helps student to understand ideas and information and make ...predictions when they are reading" (CMEC, 2007; p. 12).	The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response.
	"Readers use symbols, patterns, text features, and other elements to analyze the story…" (CMEC, 2007; p. 4).	Critical thinking is the process of evaluating information.

Thinking Skills and Document Processing Strategies	Evidence	Reason
Problem solving, integrate, and generate strategies	"This extended constructed response requires the student to demonstrate the full reading process involving integrated use of comprehension, interpretation, and personal and critical response. This measures student performance at problem definition, as well as at problem resolution. Much as in science and mathematics problem solving, students must structure the problem for themselves in order to solve it. Here students must, through comprehension, offer an interpretation, select a stance (personal, critical, or some combination) and define for themselves the depth to which they choose to go, as well as to which they can go." (CMEC, 2007; p. 6).	Students are required to analyze a problem.

Thinking Skills and Document Processing Strategies	Evidence	Reason
Creativity and generate strategies.	"Personal response: Readers step out of the text to reflect on their own experience, identify themselves with aspects of the text, establish connections with their own experience, and/or react emotionally to the text. Readers elaborate personal connections and reactions to texts by providing some extended explanations, examples, and supporting arguments. They state personal points of view and justify them. They find evidence in text to support personal claims and viewpoints about issues, themes, and situations" (CMEC, 2007; p. 4).	PCAP's definition of personal response is closely aligned with creativity, in which students are able to elaborate, state personal points of view and provide their own experiences.
Lower level thinking and locate strategies	"Comprehension: The distinction between explicit and implicit comprehension is not a dichotomous one;	PCAP's definition of comprehension is closely aligned with lower level thinking

Thinking Skills and Document Processing Strategies	Evidence	Reason
	rather, it moves incrementally from comprehension of directly stated information to comprehension gleaned both directly and indirectly through an increasing appreciation of nuance and inference. Readers identify both explicit and implicit aspects of text, using both concrete and abstract vocabulary, stated conclusions, principal ideas, important details, and/or some aspects of style" (CMEC, 2007; p. 4).	skills, in which individuals are asked to identify information gleaned within the text.

Appendix J

PCAP Thinking Skills: Analyzing the Scoring Guide

Thinking Skills	Evidence	Reasons
Critical thinking	"…relying on subtle relationships among elements…" (CMEC, 2007; p. 9).	Students not only need to understand the information correctly but also to define, generate and select that information.
	"…synthesizing several elements of the text and ideas" (CMEC, 2007; p. 9).	The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response.
	"The reader demonstrates significant critical response; evaluating the text and the relationships of elements based on social, cultural and /or literary knowledge; providing evaluation of the quality, sources, currency or relevance of text issues, themes and/or elements of style…" (CMEC, 2007; p. 9).	The response is not explicitly stated in the text; instead the student must determine, based on the information presented, the appropriate response.

Thinking Skills	Evidence	Reasons
	"…supporting response with specific details, examples, precise explanations, or thoughtful justification" (CMEC, 2007; p. 9).	Students are required to reflect, compare, and make a judgement.
	"…supporting response with reference to voice, style, stance of the writer, organization, and structure and analogies, or illustrations from the world at large" (CMEC, 2007; p. 9).	Students are required to reflect compare and make a judgement.
Problem solving	None apparent	
Creativity	"The reader demonstrates thoughtful or insightful interpretation of the text…" (CMEC, 2007; p. 9).	Students are able to include insightful interpretations, which involves the inclusion of their creativity.
	"The reader demonstrates elaborated personal response to the text: expressing personal implications and insights…" (CMEC, 2007; p. 9).	Students are able to elaborate on personal insights, which involve the inclusion of

Thinking Skills	Evidence	Reasons
		their creativity.
Lower level thinking	None apparent	

Appendix K

EQAO Assessment: Analyzing the Assessment Framework

Assessment Perspective	Evidence	Reason
Product-based	"The Education Quality and Accountability Office (EQAO) is an arm's-length agency of the provincial government that measures the achievement of students across Ontario in reading, writing and mathematics, and reports the results to parents, educators and government. EQAO assessments are based on the expectations in *The Ontario Curriculum*." (EQAO, 2007; p. 3)	*Object of Assessment*: Test developers are interested in measuring the product of achievement (e.g., the achievement of students across Ontario). *Scoring Interpretation*: Test developers are interested in comparing students based on the expectations in the Ontario curriculum. In addition, student results are compiled and reported in aggregate.

Assessment Perspective	Evidence	Reason
	"EQAO results are reported at the provincial, school board and school levels." (EQAO, 2007; p. 3)	*Scoring Interpretation*: Students' results are compiled and reported in aggregate (e.g., reported at…levels).
	"Large-scale assessments, like those conducted by EQAO, measure student achievement across the province at critical times in students' school careers." (EQAO, 2007; p. 4)	*Scoring Interpretation* Students' results are compiled and reported in aggregate (e.g., across the province).
	"The purpose of EQAO's large-scale assessments is to provide comparable year-to-year data to give the public information on student achievement." (EQAO, 2007; p. 4)	*Object of Assessment*: Test developers are interested in measuring the product of achievement (i.e., data).
	"EQAO's large-scale assessments provide reliable, objective and high-quality data that can inform school boards' improvement planning and target setting." (EQAO, 2007; p. 4)	*Goal of Assessment:* The goal is to use the results to improve educational programs (e.g.,

Assessment Perspective	Evidence	Reason
		improvement planning and target setting). *Object of Assessment*: Test developers are interested in measuring the product of achievement (i.e., data).
	"EQAO's large-scale assessment materials are created and scored 'at a distance.' The assessment scorers do not know the students personally." (EQAO, 2007; p. 4)	*Assumptions:* Knowledge is viewed as separate from the learner, such that assessments can be scored without awareness of the student (e.g., scorers do not know students).
	"EQAO's large-scale assessments are summative; they present a snapshot of student achievement or learning at the time the	*Object of Assessment:* Test developers are interested in the

Assessment Perspective	Evidence	Reason
	assessment is administered." (EQAO, 2007; p. 4)	product of achievement (i.e., results at a given point in time; snapshot).
	"EQAO's large-scale assessments require students to demonstrate their knowledge and skills independently on standardized tasks and under standardized conditions, although some accommodations are allowed for students with special education needs." (EQAO, 2007; p. 4)	*Assumptions:* Meaning resides in the text (e.g., standardized tasks…and conditions).
	"EQAO's large-scale assessments measure achievement against expectations from the prescribed curriculum and contain tasks and items that sample from and represent the curriculum for the domain assessed." (EQAO, 2007; p. 5)	*Scoring Interpretation:* Achievement results are compared to a standard (e.g., prescribed curriculum).
	"EQAO's large-scale assessments provide the same (in a given year) or psychometrically comparable items (from year to year) for all students." (EQAO, 2007; p. 5)	*Object of Assessment*: Test developers are interested in measuring the product of

Assessment Perspective	Evidence	Reason
		achievement (i.e., items are comparable from year to year).
	"In order for students' results on EQAO's large-scale assessments to be comparable across the province, the assessments must be administered, scored and reported on in a consistent and standardized manner." (EQAO, 2007; p. 5)	*Object of Assessment*: Test developers are interested in measuring the product of achievement (i.e., student's results). *Assumptions:* Meaning resides in the text (e.g., consistent and standardized manner). Meaning, results are scored consistently and without the involvement of the student. *Scoring Interpretation*: Student

Assessment Perspective	Evidence	Reason
		results are compared (e.g., comparable across the province).
	"For EQAO's large-scale assessments, all scorers use the same scoring guides and are trained and monitored to ensure objectivity and consistency. "(EQAO, 2007; p. 5)	*Assumptions:* Meaning resides in the text (e.g., ensure objectivity). Meaning, results are scored consistently and without the involvement of the student. Scoring guides are the same for all students. There are no individual differences.
	"The assessment is used to determine how well students are achieving these expectations, and their level of achievement." (EQAO, 2007; p. 5)	*Scoring Interpretation:* Students are compared to a standard (i.e., Ontario Curriculum).
	"The purpose of the Assessment of Reading, Writing and Mathematics, Junior Division (Grades 4-6) is to assess the level at	*Scoring Interpretation:* Students' results are compared to a standard

Assessment Perspective	Evidence	Reason
	which students are meeting curriculum expectations in reading, writing and mathematics at the end of the junior division (up to the end of Grade 6).” (EQAO, 2007; p. 6)	(i.e., Ontario Curriculum).
	“The Individual Student Report includes: • the student’s overall results for reading, writing and mathematics and • comparative data showing the individual student’s results in relation to school, board and provincial results.” (EQAO, 2007; p. 6)	*Scoring Interpretation:* Students’ results are compared to other students and schools.
	“School reports include • overall school-level results for reading, writing and mathematics, with comparisons to board and provincial results; • results by subgrouping, such as by gender and English-language learner and special needs status;	*Scoring Interpretation:* Students’ results are compared to other students and schools.

Assessment Perspective	Evidence	Reason
	• contextual data on demographics and participation in the assessment; • results over time; • results of the student questionnaire accompanying the assessment; • a Student Roster report that shows individual student results for each item on the assessment, with overall board and provincial results for comparison and • profiles of strengths and areas for improvement in reading, writing and mathematics." (EQAO, 2007; p. 6)	
	"Board reports include • overall board results for reading, writing and mathematics, with comparisons to provincial results; • contextual data, results over time, reports by subgroup and	*Goal of Assessment:* The goal is to use the results to improve educational programs (e.g., areas for improvement).

Assessment Perspective	Evidence	Reason
	questionnaire data and • profiles at the board level of strengths and areas for improvement in reading, writing and mathematics." (EQAO, 2007; p. 6)	*Scoring Interpretation*: Students' results are compared to other students and schools (e.g., comparison to provincial results).
	"Provincial reports include • overall provincial results for reading, writing and mathematics, including results by board; • contextual data, results over time, reports by subgroup and questionnaire data; • instructional strategies for success and • case studies (school success stories)." (EQAO, 2007; p. 6)	*Scoring Interpretation:* Students' results are compared to other students and schools (e.g., overall provincial results).
	"EQAO provides the Ontario school system with valid, reliable and comparable year-to year data on student achievement." (EQAO, 2007; p. 6)	*Object of Assessment*: Test developers are interested in measuring the product of

Assessment Perspective	Evidence	Reason
		achievement (i.e., data)
	"Schools and boards can confidently use this data along with other contextual and assessment information (e.g., on demographics, attendance and pass rates) to determine how well their improvement strategies, such as staff development or new learning resources, are working." (EQAO, 2007; p. 6-7)	*Goal of Assessment:* The goal is to use the results to improve educational programs (e.g., to determine… how learning resources are working).
	"EQAO's junior division assessment • provides data to assist schools and boards in improvement planning and target setting; • supports the successful implementation of the curriculum; • improves understanding of assessment practices among educators across the province through capacity building (for example, by assisting teachers to understand the curriculum and to	*Goal of Assessment:* The goal is to use the results to improve educational programs (e.g., improvement planning and target setting).

Assessment Perspective	Evidence	Reason
	develop related classroom assessments) and • improves understanding of assessment practices among the public." (EQAO, 2007; p. 7)	
Process-based	"They are used by the Ministry of Education, district school boards and schools to improve learning, teaching and student achievement. An Individual Student Report is also provided by EQAO for each student who writes an EQAO assessment." (EQAO, 2007; p. 3)	*Assessment Goal:* The goal of the assessment is to use the results to improve learning.