



uOttawa

L'Université canadienne
Canada's university

**FACULTÉ DES ÉTUDES SUPÉRIEURES
ET POSTDOCTORALES**



**FACULTY OF GRADUATE AND
POSTDOCTORAL STUDIES**

Martha Jane Koch

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Education)

GRADE / DEGREE

Faculty of Education

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

**Implications of the Multiple-use of Large-scale Assessments for the Process of Validation: A Case
Study of the Multiple-use of a Grade 9 Mathematics Assessment**

TITRE DE LA THÈSE / TITLE OF THESIS

Christine Suurtaam

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

Richard Barwell

Barbara Graves

**Susan Brookhart
Duquesne University**

Raymond LeBlanc

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

**Implications of the multiple-use of large-scale assessments for the process of validation:
A case study of the multiple-use of a Grade 9 mathematics assessment**

Martha J. Koch

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies in partial fulfillment of
the requirements for the PhD degree in Education.

Teaching, Learning and Evaluation

Faculty of Education

University of Ottawa

©Martha J. Koch, Ottawa, Canada, 2010



Library and Archives
Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-74235-8
Our file Notre référence
ISBN: 978-0-494-74235-8

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.


Canada

Table of Contents

List of Tables	iv
List of Figures	v
Abstract	vi
Acknowledgements	vii
Prologue	1
Chapter 1 – Introduction	5
Scope of the Inquiry	5
Clarifying Assessment-related Terms	7
A Sociocultural View of Assessment	10
Overview of the Dissertation.....	16
Chapter 2 – A Theoretical Discussion of Multiple-use	21
Today’s Large-scale Assessment Practices	21
Current Practices in the Use of Large-scale Assessments	26
The Problems with Multiple-use	41
The Prevalence of Multiple-use	50
Factors Which May Encourage Multiple-use	52
The Need for Research into Multiple-use	56
Chapter 3 – Theoretical Frameworks for the Process of Validation	58
Clarifying Validity-related Terms	59
Measurement-based Conceptions of the Process of Validation	63
Validation for Large-scale and Classroom Assessments	73
Kane’s Model of Argument-based Validation	74
Re-conceptualizing Validity Theory	85
The Value of Sociocultural Studies in Reframing Validity Theory	87
Chapter 4 – Contextualization of the Case Study	94
Grade 9 Mathematics Instruction and Assessment in Ontario	95
Development, Administration & Validation of the Grade 9 Assessment	99
Research on Using the EQAO Assessment in Students’ Grades	118
Chapter 5 – Methodology, Research Design & Data Construction	127
Methodology of the Inquiry	127
Research Questions and Design	136
The Process of Data Construction	142

Chapter 6 – Observations of the Multiple-use of the EQAO Grade 9 Assessment...	168
Research Question 1 - The Pattern of Uses	168
Research Question 2 - Describing Teachers' Use	181
Research Question 3 - EQAO's Perspective	235
Research Question 4 - Identifying Interactions	243
Discussion of Case Study Observations	250
Chapter 7 – Validating the Multiple-uses of the Grade 9 Assessment	256
Research Question 5 - Applying Kane's Model	256
Research Question 6 - Using Boundary Concepts	268
Chapter 8 – Implications of Multiple-use for the Process of Validation	294
Characteristics of the Process of Validation that Address Multiple-use	295
Contributions to Educational Assessment	301
Limitations of this Inquiry	304
Suggestions for Further Research.....	307
Closing Remarks	309
Epilogue	311
References	315
Appendices	333
Appendix A – EQAO Multiple-choice Items (Grade 9 Applied)	
Appendix B – EQAO Open-response Item (Grade 9 Academic)	
Appendix C – EQAO Generic Rubric for Open-response Items	
Appendix D – EQAO Item Specific Rubric for Population Plans Item	
Appendix E – EQAO Item Population Plans Anchor Response	
Appendix F – Interview Protocol for Department Heads & Teachers	
Appendix G – Interview Protocol for Principals	
Appendix H – Interview Protocol for EQAO Personnel	

List of Tables

Table 1: Achievement Levels for the EQAO Grade 9 Assessment of Mathematics	106
Table 2: Percentage of Ontario Students at or above Level 3 on the EQAO Grade 9 Academic and Applied Assessments of Mathematics	108
Table 3: Studies of the EQAO Grade 9 Assessment of Mathematics Commissioned by EQAO	113
Table 4: Percentage of Teachers who Report Using Province-wide Assessments as Part of Students' Grades by Province or Territory	120
Table 5: Teachers' Responses to <i>What is the contribution of the EQAO assessment to your final report card mark?</i>	123
Table 6: Overview of Interviews in School-level Case Studies	152
Table 7: Teachers' Reported use of the EQAO Grade 9 Assessment of Mathematics as Part of Students' Grades	182
Table 8: Participants at Birch Park Secondary School	185
Table 9: Participants at Scottsdale Collegiate Institute	191
Table 10: Participants at Taylor Brook High School	197

List of Figures

Figure 1. A visual representation of intended and alternate proposed uses.	30
Figure 2. A visual representation of the practice of multiple scoring.	34
Figure 3. A visual representation of the practice of multiple-use showing four uses.	40
Figure 4. A visual representation of the practice of multiple-use showing interactions.	47
Figure 5. A visual representation of intended and alternate proposed uses, multiple scoring and multiple-use practices including interactions among some uses.	49
Figure 6. A visual representation of Kane’s argument-based model of validation. .	84
Figure 7. A visual representation of five multiple-uses of the EQAO Grade 9 Assessment of Mathematics.	177
Figure 8. Percentage of students at or above Level 3 on the EQAO Grade 9 Applied Assessment of Mathematics by school, board and province for Birch Park.	189
Figure 9. Percentage of students at or above Level 3 on the EQAO Grade 9 Academic Assessment of Mathematics by school, board and province for Birch Park.	189
Figure 10. Percentage of students at or above Level 3 on the EQAO Grade 9 Applied Assessment of Mathematics by school, board and province for Scottsdale Collegiate.	195
Figure 11. Percentage of students at or above Level 3 on the EQAO Grade 9 Academic Assessment of Mathematics by school, board and province for Scottsdale Collegiate.	195
Figure 12. Percentage of students at or above Level 3 on the EQAO Grade 9 Applied Assessment of Mathematics by school, board and province for Taylor Brook.	201
Figure 13. Percentage of students at or above Level 3 on the EQAO Grade 9 Academic Assessment of Mathematics by school, board and province for Taylor Brook.	201

Abstract

While most large-scale assessments are designed for one intended use, in many instances, a single administration of an assessment is used for one or more uses in addition to the intended use. I suggest the term *multiple-use* to refer to this practice. The purpose of this dissertation is to consider the implications of the multiple-use of large-scale assessments for the process of validation. I begin by discussing multiple-use from a theoretical perspective, identifying the challenges this practice creates for the process of validation. I maintain that multiple-use increases the stakes associated with an assessment and argue that where multiple-use is known to occur, the separate validation of each use may not be adequate since interactions between multiple-uses may take place. To build on this theoretical discussion, I report on an in-depth case study of one instance of multiple-use. Using questionnaire data, document analysis, school-level case studies and interviews with test development personnel, I describe the multiple-use of the *Education Quality and Accountability Office (EQAO) Grade 9 Assessment of Mathematics* which is administered in Ontario, Canada. I identify the pattern of uses for this assessment and focus on two multiple-uses: the use of the assessment for accountability and the use of the assessment by teachers as part of students' grades. Evidence of interactions between the two uses is provided and the limitations of applying Kane's (2006) argument-based model of validation given these interactions are discussed. I demonstrate how the concepts of boundary objects and boundary encounters, situated within sociocultural theory, contribute to the process of validation for this assessment and may be beneficial for the validation of other instances of multiple-use. A number of ways of rethinking the process of validation to better address the multiple-use of large-scale assessments are suggested and areas for further research are identified.

Acknowledgements

I want to express my appreciation to the many people who have helped me complete this dissertation. I am grateful to the teachers, principals, and EQAO personnel who took the time to participate in interviews, complete questionnaires and respond to my inquiries. Their contributions have made this dissertation possible.

I want to thank Dr. Christine Suurtamm, my thesis supervisor, for listening to my ideas, asking questions and offering suggestions that, time and again, clarified and extended my thinking and helped me express my thoughts more effectively. More than that, Chris has been a constant source of encouragement and a wonderful mentor from whom I have learned a great deal about research and a myriad of other things.

I also want to offer my thanks to the members of my thesis committee. Dr. Barbara Graves' enthusiasm for exploring ideas is a tremendous inspiration and I am grateful for her advice and friendship over many years. Dr. Richard Barwell offered many thoughtful comments and his keen awareness of research as an ongoing process has been a genuine insight for me. My understanding of sociocultural theory and the value of narrative for research has been greatly enriched by conversations and course work with Dr. Raymond LeBlanc. I have learned a great deal from discussions with my thesis committee and I know their contributions will add to the quality of my future research endeavours.

Many thanks are also due to my family and friends whose faith in my ability to complete this project gave me the confidence to overcome the obstacles I encountered. The energy and enthusiasm of my children, Keaghan, Nathan and Hannah have contributed to this project in countless ways beginning years before the research started. My thanks also to Paul for his willingness to talk through ideas at any time of the day or night, year after year and for all the other ways he enriches my life.

Finally, I want to acknowledge the generous financial support I have received from the Social Sciences and Humanities Research Council of Canada, the Faculty of Education and the Faculty of Graduate and Postdoctoral Studies at the University of Ottawa.

Prologue

The inspiration for this inquiry did not emerge from my earlier research endeavours or from course work in my doctoral program; it came to me in the fall of 2004 when I was attending an information evening for parents at my son's high school. Sometimes called "curriculum night", the information evening is a common occurrence at high schools in Ontario. At this event, which typically takes place a few weeks into the first semester of the school year, parents move from classroom to classroom while teachers outline the curriculum that will be covered and describe how students will be assessed. I had completed a Masters degree in educational measurement a few years earlier, had been involved in a number of research projects in educational assessment and had been teaching curriculum and evaluation courses in a Bachelor of Education program for several years. My son was in his first semester of Grade 9 and I was particularly interested in learning about the assessment practices his teachers were planning to use.

As I listened to my son's mathematics teacher explain his assessment plan, I was surprised to learn that each student's performance on the *Education Quality and Accountability Office (EQAO) Grade 9 Assessment of Mathematics* (EQAO, 2009a) would contribute 10% toward their overall course grade. I was familiar with this province-wide assessment from my work in educational measurement. I knew that it had been introduced a few years earlier as part of an increased focus on accountability in Ontario and that it was designed to provide data on student achievement for schools and school districts but I had never heard of it being used as part of students' grades. I couldn't imagine how this was accomplished. I knew that the results from one year's testing were not returned to schools until the first semester of the following school year and I couldn't understand how these

results could be included in students' end-of-semester report card grades. As I sat in my son's mathematics classroom, more and more questions came to mind. Glancing around the room, it seemed that other parents were not particularly interested in this aspect of the teacher's assessment plan, so despite my growing curiosity, I asked only how the EQAO scores were received in time to be incorporated into students' report card grades. The teacher, who was also the mathematics department head, explained that a selection of items were chosen and marked by each classroom teacher before the tests were sent to EQAO for official marking and that these teacher-derived scores were included in each student's final course grade. He indicated that this practice had been adopted by the school's mathematics department to encourage students to take the EQAO assessment more seriously and to ensure that more value would come from the time and effort spent preparing for and completing the EQAO assessment.

Rather than setting my mind at ease, the teacher's response triggered many additional concerns leaving me with the view that this use of the EQAO assessment was inappropriate and misguided. My concerns were related to a number of issues based on established principles of measurement. These principles, outlined in documents such as the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999), are intended to guide the development, use and validation of large-scale assessments. I wondered if the use of EQAO Grade 9 assessment items as part of students' grades was a common practice across the province, recognizing that if the practice was not universally adopted it would create different testing conditions in each school or district. Students in some schools would complete the EQAO assessment with additional motivation

and this might jeopardize the comparability of the assessment results across schools and districts. A number of other issues became apparent to me. Was this use of the assessment by classroom teachers approved by EQAO and, if so, why had EQAO decided to allow it? What implications might arise from the idiosyncratic approaches teachers could use to select and mark items from the assessment? How appropriate is the use of an externally developed assessment as a measure of students' learning in a specific classroom context? What are the perspectives of teachers regarding this use of the EQAO Grade 9 assessment? How much information are students and parents given about this practice and what are their views of having the assessment included in students' grades?

After that evening, I continued to think about this use of the EQAO Grade 9 assessment and I began to read about how this practice might be viewed within the field of educational measurement. As I had recently been involved in validity research for another large-scale assessment and many of my concerns about this use of the EQAO Grade 9 assessment related to validity, I began to focus specifically on how the process of validation might proceed given this practice. I questioned the theoretical and practical implications of using an assessment designed for accountability purposes for that purpose as well as including the items as part of students' grades. How would measurement practitioners approach the validation of such a practice and what relevance, if any, would their approach to validation have for mathematics teachers using the assessment as part of their students' grades? Through informal discussions with mathematics teachers in various school districts in Ontario, I learned that the use of the EQAO Grade 9 assessment for students' grades was not limited to my son's school but was taking place in many high schools and yet I could not find any research about this practice. I also began to wonder if using one administration of a

large-scale assessment for more than one purpose, a practice I started to refer to as *multiple-use*, might be taking place in other educational assessment contexts.

I had been interested in pursuing doctoral studies for several years but had not found a suitable topic. The prospect of investigating this practice seemed both intriguing and challenging. Since attending that parent information meeting my perspective regarding this practice has shifted and expanded considerably. What follows in this manuscript is an account of the changes in my perspective and of the inquiry process through which those changes took place.

CHAPTER 1

INTRODUCTION

The purpose of this inquiry is to gain an understanding of the implications of the multiple-use of large-scale assessments for the process of validation. In this introductory chapter I explain the scope of the inquiry, clarify my use of some assessment-related terms and describe the sociocultural view of assessment which I adopt. I also provide an overview of the remaining chapters in the dissertation.

Scope of the Inquiry

This inquiry begins with a theoretical discussion of the practice of multiple-use. I define and describe multiple-use and identify the challenges this practice creates for the process of validation.¹ I suggest that the practice of multiple-use effectively increases the stakes associated with an assessment. Moreover, I argue that where multiple-use is known to occur, the separate validation of each use may not be adequate because multiple-uses may interact with one another. Descriptions of the use of large-scale assessments in the literature provide evidence that multiple-use is neither a recent phenomenon nor a rare occurrence and yet very little research into this practice has been conducted. The few studies which have been done identify some issues multiple-use creates for validation but neither the problem of increased stakes nor the potential for interactions between multiple-uses to occur have been previously considered.

In the second phase of this inquiry I build on my theoretical discussion of multiple-use by conducting an in-depth case study of one instance of this practice. Using questionnaire data, document analysis, school-level case studies and interviews with test

¹ An extensive discussion of how the process of validation has been conceptualized and of the way I use this term is included in chapter 3.

development personnel, I investigate the multiple-use of the Education Quality and Accountability Office (EQAO) Grade 9 Assessment of Mathematics. This assessment, which is referred to in the prologue, is administered annually to about 150,000 students in all publicly funded high schools in Ontario, Canada (EQAO, 2009a). After identifying the pattern of uses associated with the EQAO Grade 9 assessment, I focus on two of the multiple-uses: the use of the assessment for accountability and the use of the assessment by teachers as part of students' grades. I describe the procedures teachers use to include the EQAO Grade 9 assessment as part of their students' grades as well as their views of this practice. I also consider teachers' use of the assessment for students' grades from the perspective of EQAO. Including both perspectives in this study enables me to identify some interactions which take place between these two uses. In light of these interactions, I consider the adequacy of an argument-based model of validation (Kane 1992; 2006), situated within a measurement-based theoretical perspective, for this instance of multiple-use. I also demonstrate how using the concepts of boundary objects (Star & Griesemer, 1989; Bowker & Star, 1999) and boundary encounters (Cobb, McClain, Lamberg & Dean, 2003) which are rooted in a sociocultural theoretical perspective can contribute to the process of validation for the EQAO Grade 9 assessment.

As will become apparent, my study of the multiple-use of the EQAO Grade 9 assessment contributes to the process of validation for this large-scale assessment. However, my principal reason for conducting this study is to better understand the practice of multiple-use broadly and to identify some of the implications of this practice for the process of validation. Shepard states that research that is "embedded in the dilemmas of practice" (2000, p.13) may be the most useful for learning about and understanding complex areas of

practice. In keeping with Shepard's view, in the final part of this dissertation I draw on my empirical findings concerning the multiple-use of the EQAO Grade 9 assessment to build on my theoretical discussion of multiple-use. I illustrate some implications of multiple-use for the process of validation and suggest some ways of rethinking the process of validation to better address the practice of multiple-use. Moss observes that 'richly contextualized cases of practice' can help to expand our conceptual frameworks (2003, p.14). Accordingly, this in-depth case study of the multiple-use of the EQAO Grade 9 assessment is intended to contribute to the dialogue currently taking place in educational assessment and measurement surrounding the re-conceptualization of the process of validation.

Clarifying Assessment-related Terms

In this section, I clarify my use of some assessment-related terms. In particular, I discuss the distinction I make between *assessment* and *measurement* as well as my use of the terms *large-scale assessment* and *classroom assessment*.

The terms assessment and measurement have been used by researchers in a variety of ways. In my use of these terms I draw on the work of Delandshere and Petrosky (1998) who define measurement as "the assignment of numbers to properties of objects or events according to rules whose validity can be tested empirically" (p.16). Delandshere and Petrosky maintain that assessment differs from measurement in that assessment encompasses all the procedures that are used to obtain information about students' learning and to make judgments concerning the significance, importance, or value of that information, whether or not numbers are assigned. Thus, according to Delandshere and Petrosky, measurement can be an aspect of assessment but assessment is a broader concept that is not limited to the process of assigning numbers to objects or events. In this dissertation I use the terms

assessment and measurement in a similar manner and in my inquiry into the multiple-use of large-scale assessments I pursue the wider range of issues that are encompassed by assessment including, but not limited to, those issues that relate to measurement.

As this dissertation concerns both large-scale and classroom assessment practices, these terms also need to be clarified. The term large-scale assessment generally refers to assessments that are administered to large numbers of students under standardized conditions where students respond to the same, or nearly the same, items (Wilson, 2007, p.1099). These assessments are developed by external agencies and are administered to students relatively infrequently. Standardized approaches to administration and scoring are used to enhance the comparability of results from year to year or across jurisdictions and, in many instances, scoring is conducted by the agency that developed the assessment. As noted by Wilson, the term large-scale assessment does not exclusively refer to machine-scored assessments with multiple-choice items; large-scale assessments often include a variety of item formats and scoring techniques. In contrast, classroom assessment is typically used to refer to those assessment activities that provide information about individual students' learning and that can be used day-by-day to guide the teaching and learning activities that take place in classrooms (Shepard, 2006). Classroom assessments are conducted much more frequently than large-scale assessments and can employ a wide range of assessment strategies and tools including informal techniques such as observation and questioning as well as more structured activities such as performance tasks, quizzes, and tests. While most classroom assessments are created by individual teachers, commercially produced classroom assessments are also available for teachers' use.

Many important theoretical and practical differences between large-scale and classroom assessment have been identified and explored in recent publications and a strong case has been made for further developing theories specifically for classroom assessment (Brookhart, 2003, 2004; Moss, 2003; Shepard, 2006; Wilson & Kenney, 2003). However, there are also some important similarities between large-scale and classroom assessment as well as situations where making a distinction between these assessment types is less clear or may be of limited value (Delandshere, 2001, 2002). Edwards (2005) observes that many binary distinctions which are made in educational discourse (e.g. formal vs informal learning) oversimplify the complex practices which take place. I agree with his observation and believe that it is applicable to large-scale and classroom assessment practices. For instance, this dissertation explores teachers' use of a large-scale assessment as part of their classroom assessment strategy. Thus, one assessment can be characterized as both a large-scale and a classroom assessment. Stated another way, in this instance large-scale and classroom assessment practices essentially overlap and the importance and value of distinguishing between the two assessment types must be carefully considered. Accordingly, while I use the terms large-scale and classroom assessment with their commonly understood meanings (as described in the previous paragraph), I see the distinction between these terms as problematic in that there are situations where making this distinction leads to a simplified or reduced view of the complex assessment activities that are taking place. I return to this idea and elaborate on it more fully later in the dissertation.

Having clarified my use of some assessment-related terms, I turn to a description of the theoretical view of assessment which most strongly influences my thinking and which I consider to be the overarching frame for this inquiry.

A Sociocultural View of Assessment

Throughout much of the 20th century, both measurement and assessment were often viewed as relatively objective and context-free technical procedures (Delandshere, 2001; Moss, Girard & Haniford, 2006; Shepard, 2006). According to this view the central concern was to investigate the psychometric properties of assessments and enhance those properties by improving processes such as item development, scoring, equating, and scaling. While this psychometric or technical view originated in measurement-based approaches to large-scale assessment, it has also been applied to classroom assessment contexts (Brookhart, 2003, 2004; Shepard, 2006; Moss, 2003) and continues to be the principal focus for many measurement researchers. However, in recent years, a number of researchers have suggested that considering assessment practices from a sociocultural perspective may be beneficial for both large-scale and classroom assessments (Delandshere, 2001, 2002; Gipps, 1999, 2002; Moss, 2007; Moss et al., 2006; Stobart, 2008). These authors maintain that focusing on the technical aspects of measurement is a limited approach because it sets aside the complex human interactions that take place in assessment and fails to consider the social, political and ethical aspects of assessment practice. Adopting a sociocultural approach to assessment takes these aspects of assessment practice into account.

As will become apparent, I refer to a number of theoretical perspectives in this inquiry into the multiple-use of large-scale assessments but the perspective which most strongly influences my research is rooted in sociocultural studies. In this section, I describe some key aspects of the sociocultural view of assessment that I adopt. In particular, I discuss the importance of assessment to support students' learning, the view of assessment as a

value-laden and culturally situated process and the importance of context and contextualization in assessment practice.

Assessment to Support Students' Learning

From a sociocultural perspective, teaching, learning and assessment are regarded as intertwined processes (Cobb, 1994; Gipps, 2002; Moss, 2003; Shepard, 2001). Assessment practices both influence and are influenced by the approaches to teaching and learning that are carried out in a given setting. While assessment may be conducted for many reasons such as reporting on students' achievement or monitoring the effectiveness of an instructional program, the main purpose of assessment, from a sociocultural perspective, is to support and enhance learning. Delandshere and Petrosky (1998) note that a shift toward the view of assessment as principally to support students' learning took place in the 1990s. I have found this emphasis on assessment to be fairly common in theoretical discussions of assessment (e.g. Wilson, 2007), in standards intended to guide assessment practice (e.g. Joint Committee on Standards for Educational Evaluation [JCSEE], 2003; National Council of Teachers of Mathematics [NCTM], 1995) and in curriculum documents for various educational jurisdictions (e.g. Ontario Ministry of Education [OME], 2005).

Many phrases are used to refer to assessment that supports student learning such as formative assessment (e.g. Black & Wiliam, 1998; Brookhart, 2001), assessment for learning (e.g. Gipps, 1999; Stiggins, 2005), and assessment as learning (e.g. Earl, 2003). While each phrase reflects a slightly different meaning, they all emphasize that assessment activities should be consistent with the approach to learning that is being adopted in a classroom. Assessment supports learning when it is embedded in instructional activities rather than being a separate process that takes place after learning has occurred (Shepard, 2000). Gipps

(1999, 2002), drawing on the work of Vygotsky, points out that in order to support learning, assessment activities should include the use of the same tools that were used during the learning process and assistance and support from peers and adults during some assessment activities should take place. In order to support and enhance learning, assessment must also provide opportunities for informative feedback that can be used by students and their teachers. Indeed, many researchers suggest that providing feedback to students is one of the principal means by which assessment supports the learning process (e.g. Black & Wiliam, 1998; Brookhart, 2001).

In a sociocultural view of assessment, what is assessed is not only the product or outcome of learning but the learning process that is taking place. Assessing the learning process provides valuable insights into how students are learning and suggests ways that they can be supported in that learning. From a sociocultural perspective, assessment itself is understood as an ongoing and dynamic process. Assessment is a dynamic process because learning is continually taking place as assessment occurs such that what is being assessed is constantly shifting and changing. This view of assessment recognizes that learning cannot be adequately measured by one instrument (Brookhart, 2003; Gipps, 1999) and that assessing complex processes such as problem solving and critical thinking requires innovative and varied assessment practices (Delandshere & Petrosky, 1998). Thus, in a sociocultural approach to assessment multiple sources of evidence are considered and the interpretations made on the basis of that evidence are constantly evolving (Moss, 2003; Shepard, 2001).

Assessment as a Value Laden, Social Process

Sociocultural views of assessment move away from thinking of assessment as a series of isolated events that objectively measure the acquisition of knowledge and toward a

view where assessment is seen as an inherently value-laden process which takes place in a specific cultural and social setting. Erickson (1986) notes that sociocultural approaches to inquiry stress the importance of looking at various levels of a system rather than considering a single level in isolation. For this reason, sociocultural studies of assessment practice typically include a consideration of the social, cultural and political aspects of the individual, the classroom, the school, the district and beyond. The sociocultural view recognizes that decisions are made about what will be assessed on the basis of the values of the individuals and groups who are involved in the process. Making those values explicit helps others understand the inferences which can be made from the assessment process. Assessment involves many individuals and groups and those involved may interpret an assessment in a variety of ways. Thus, multiple interpretations of assessment practices are possible and considering these different interpretations is an important aspect of this theoretical perspective (Gipps, 1999; Shepard, 2000, 2001).

Another important consequence of viewing assessment as a social practice is that the relationships among those involved in the process are also considered. There is an emphasis on reducing the hierarchical relationships among the individuals and groups who are involved in assessment practice. For instance, sociocultural views of assessment move away from a view of the teacher as the assessor and the student as the passive object of the assessment (Gipps, 2002). The sociocultural approach seeks to appreciate the learner's perspective and include it in assessment activities. Assessment is characterized by greater student involvement including approaches such as having students work collaboratively with teachers to develop assessment tasks and establish the evaluation criteria for those tasks. A

high level of student involvement is also instantiated by an increased emphasis on the use of peer and self assessment techniques (Brookhart, 2001; Shepard, 2006).

Context and Contextualization in Assessment Practice

Several authors have observed that, from a sociocultural perspective, assessment is best understood by considering the contexts where it takes place (Delandshere, 2002; Gipps, 1999; Moss, 2003). However, as noted by Chaiklin and Lave (1996) and Edwards (2005), the concept of context can be conceptualized in a number of ways. One view of context, arguably the most enduring and powerful, is as a bounded setting, container or backdrop for human activity (Chaiklin & Lave, 1996; Edwards, 2005). According to this view, the classroom is a context or setting in which learning activities take place. What goes on inside the classroom is seen as largely distinct or separate from what goes on outside the classroom. Context functions as a relatively static and stable container or backdrop for the activities that take place; it precedes the activities which are taking place and is seen as separate from those activities. In this view, context is an important consideration in the sense that it can be used to help explain the activities which have taken place.

Many sociocultural theorists have problematized the notion of context as a bounded setting and suggest that context may be better viewed as a relational and emergent process that encompasses the relationships among individuals, the tools that mediate their activities and the processes in which they are engaged (Chaiklin & Lave, 1996; Edwards, 2005). In the view of context as a relational process, one context cannot be separated from another and rather than preceding the activities which are taking place, contexts co-emerge with the social practices that are taking place. Edwards (2005) suggests that the term *contextualization* might be used to capture this more dynamic, relational and emergent sense

of context. He also proposes that because all contexts include traces of other contexts, the notion of *polycontextualization* might more accurately describe what takes place.

While Edwards (2005) describes context, contextualization and polycontextualization with reference to the discourse of lifelong learning, I think his approach also reflects views of context in educational assessment where context is sometimes treated as a noun or objective thing and sometimes treated as a verb or an action. I see the notion of context as a bounded setting within which assessment activities take place often being used in the measurement field, particularly with reference to large-scale standardized assessment activities where context is typically viewed as the conditions which precede and surround an assessment activity and where attempts are made to control or manipulate those conditions. In measurement, context is seen as separate from the assessment activity and is often used to help explain what took place. For example, when administering a large-scale assessment, invigilators are required to follow standardized procedures and must report any unexpected events or changes in the testing conditions which took place.

In sociocultural views of assessment the notion of context as a relational process seems to be more evident. Assessment is seen as contributing to the classroom culture but is also seen as responding to that culture. For example, the classroom culture may be highly collaborative or it may be strongly competitive and this culture both influences and is influenced by the approach to assessment which is taking place. At the same time, activities which are taking place at the school, district and other levels of the education system are also influencing and being influenced by the assessment practices in the classroom. In this way, assessment practices are dynamic processes. Thus, similar to the notions of contextualization and polycontextualization proposed by Edwards, context is something that precedes an

assessment event, while contextualization is an action that co-emerges as various assessment activities take place.

Both context as bounded setting and contextualization as a relational process are recurring concepts in this inquiry into the implications of the multiple-use of large-scale assessments for the process of validation. As will become evident, these concepts have influenced the methodology of this inquiry and are tied to the concepts of boundary objects and encounters which are used in the analysis of the case study observations. Accordingly, following on Edwards (2005), in my references to context as it is used in measurement I use the term context to signify a bounded setting and when I wish to signify the notion of context as an emergent, dynamic and relational process, I use the term contextualization. However, I also want to point out that the distinction between context and contextualization, like the distinction between large-scale and classroom assessment, can sometimes be an oversimplification and context may, at times, be used to signify both the setting and the relational processes which take place as part of assessment practice. I return to a discussion of the connections between context, contextualization and boundary concepts in chapter 7.

In this section I have summarized some key aspects of the sociocultural view of assessment that I adopt. A more extensive exploration of this theoretical framework emerges throughout the dissertation as I continue to identify how sociocultural perspectives have influenced my inquiry process.

Overview of the Dissertation

In the final section of this chapter, I provide an overview of the sequence and content of the remaining chapters in this dissertation. In Chapter 2, I define the practice of multiple-use and describe the problems that the multiple-use of large-scale assessments creates for

current conceptions of validation. I also consider the prevalence of multiple-use, suggest some factors which may contribute to this practice taking place, and describe the need for further research in this area.

Because this inquiry focuses on the process of validation in the context of multiple-use, in chapter 3, I clarify my use of validity-related terms, discuss some measurement-based principles for validation and introduce Kane's argument-based approach (2006). I also summarize some of the limitations of current conceptions of validity which are identified in the literature and describe recent calls which have been made for the re-conceptualization of validity theory. I explore the suggestion that drawing on other disciplines in addition to educational measurement can be worthwhile for reframing notions of validity (Messick, 1989; Moss 2003, 2007; Mislevy, 2007; Mislevy, Moss & Gee, 2009). In particular, I focus on Moss's contention that concepts which originate in sociocultural theory such as the notion of a boundary object can contribute to conceptions of validity and the process of validation for educational assessments (Moss et al., 2006).

Chapter 4 moves from a theoretical discussion of multiple-use to my case study of the multiple-use of the EQAO Grade 9 Assessment of Mathematics. I begin to contextualize this study by including an overview of Grade 9 mathematics courses in Ontario, a summary of policies and practices which relate to mathematics assessment in Ontario high schools, and a description of the development, administration and existing validity evidence for the EQAO Grade 9 Assessment of Mathematics. I also summarize existing literature related to teachers' use of the EQAO Grade 9 assessment as part of students' grades as this is one of the multiple-uses which I focus on in the case study.

Chapter 5 provides an account of the methodology of the case study. I begin by explaining my epistemological stance and then describe how I combined this stance with ideas about the re-conceptualization of validity theory and with my sociocultural view of assessment to establish the methodology for this inquiry. I also discuss the research questions which structure the case study. As noted, the overall purpose of this inquiry is to consider the implications of the multiple-use of large-scale assessments for the process of validation. To address this purpose, I developed six specific research questions which I include in this introductory chapter as a way of outlining the study, even though many of the terms used in the questions have not yet been defined. The research questions are:

1. What is the pattern of uses associated with the EQAO Grade 9 Assessment of Mathematics?
2. How does teachers' use of the EQAO Grade 9 Assessment of Mathematics as part of students' grades take place?
3. How is teachers' use of the EQAO Grade 9 Assessment of Mathematics as part of students' grades viewed by EQAO?
4. Is there evidence to suggest that there are interactions between teachers' use of the assessment as part of students' grades and EQAO's use of the assessment for accountability purposes?
5. How can Kane's argument-based model of validation be applied to the multiple-use of the EQAO Grade 9 Assessment of Mathematics?
6. In what ways can the concepts of boundary objects and boundary encounters contribute to the process of validation given the multiple-use of the EQAO Grade 9 Assessment of Mathematics?

In addition to providing a more extensive discussion of these questions, in chapter 5 I explain the research design and describe the methods I used in my case study of the multiple-use of the EQAO Grade 9 assessment.

Chapter 6 presents the observations that address the first four research questions. With regard to the first question, I draw on my observations to describe the pattern of uses associated with the EQAO Grade 9 assessment. I then focus on two of the uses identified in

this pattern: the use by teachers as part of students' grades and the use by EQAO for accountability. To address the second question, I provide a description of my findings regarding how teachers use the EQAO Grade 9 assessment as part of their students' grades. These findings include a discussion of the procedures and views of classroom teachers. To address the third research question, I draw on my observations to summarize EQAO's policies and perspectives on the use of the assessment as part of students' grades. For the fourth question, I present the evidence I found of interactions between teachers' use of the Grade 9 assessment as part of students' grades and EQAO's use of the assessment for accountability.

Chapter 7 addresses the fifth and sixth research questions. In light of the interactions presented in chapter 6, I consider how the process of validation might take place for the EQAO Grade 9 assessment. I provide an account of my attempt to apply Kane's argument-based model of validation to the multiple-use of the EQAO Grade 9 assessment (Kane, 2006). After showing the difficulties and limitations of applying Kane's approach to this instance of multiple-use, I describe my application of the concepts of boundary objects (Star & Griesemer, 1989; Bowker & Star, 1999) and boundary encounters (Cobb et al., 2003) and discuss how these concepts contribute to the process of validation for the EQAO Grade 9 assessment.

In the final chapter I return to a discussion of the overall purpose of the inquiry. I identify and discuss the insights for the process of validation which emerge from my investigation of the multiple-use of the EQAO Grade 9 assessment. I describe how the insights gained from this particular instance of multiple-use help to extend and refine the theoretical discussion of the problem of multiple-use and suggest some changes which might

be made to the process of validation where multiple-use is known to occur. Finally, I identify some of the limitations of this research and suggest possibilities for further research into the multiple-use of large-scale assessments.

CHAPTER 2

A THEORETICAL DISCUSSION OF MULTIPLE-USE

In this chapter I define the practice of multiple-use and describe the problems it creates for the validation of large-scale assessments. I begin by outlining some recent trends in large-scale educational assessment and then describe three practices, one of which is multiple-use, which are currently occurring within this large-scale assessment milieu. Each practice relates to the scoring, interpretation and use of large-scale assessments and, when considered together, these practices demonstrate the complex ways that large-scale assessments are being used in many educational jurisdictions. I describe all three practices, rather than presenting multiple-use alone, because the similarities and differences among the practices highlight some important characteristics of multiple-use and help to illustrate the particular challenges that multiple-use creates for validation. In addition, aspects of each of the three practices are evident in the use of the EQAO Grade 9 Assessment of Mathematics and explaining these practices provides relevant background information for understanding the multiple-uses of this assessment. After describing these practices, I focus on the problems that the practice of multiple-use creates for the process of validation. I explain these problems, consider some factors which may result in an increase in the occurrence of multiple-use and conclude the chapter with a discussion of the need for research into the practice of multiple-use.

Today's Large-scale Assessment Practices

Over the past few decades the use of large-scale educational assessments has grown considerably in many countries with results from such assessments increasingly being relied upon by educators and policymakers for a range of decisions. For example, during the period

from 1974 to 1999, Phelps (2000) found that the number of large-scale educational assessment programs increased in 27 countries and provinces and decreased in only 3 locations included in his study. The increased focus on large-scale assessment includes a greater emphasis on the use of assessments for accountability purposes and has been accompanied by rapid changes in large-scale assessment methods as well as increased recognition of the impact of large-scale assessments on student learning, classroom practice, and educational policy. Each of these aspects of large-scale assessment practice is briefly discussed as part of my contextualization of the practice of multiple-use.

Large-scale Assessments in Accountability Systems

In recent years, many new large-scale educational assessments have been introduced as part of educational reform efforts (Baker & Linn, 2002; Clarke, Madaus, Horn & Ramos, 2000; Fullan, 2001; Levin, 1998). In many jurisdictions, these large-scale assessments are the basis of accountability systems that have been designed to evaluate and report on the progress of schools. The increased use of large-scale assessments for accountability has been observed in England (Gipps, 2003; Stobart, 2009), the United States (Clarke et al., 2000; Johnson & Johnson, 2006; Kohn, 1999; Koretz & Hamilton, 2006; Wilson, 2007), Canada (Crundwell, 2005; Klinger, DeLuca & Miller, 2008) and a number of other countries (Phelps, 2000).² An important characteristic of assessments within accountability systems is the emphasis placed on communicating results to the public. The results from these assessments are typically widely disseminated and can be readily accessed through print resources and internet websites (Crundwell, 2005; Koretz & Hamilton, 2006; Volante, 2007). Results are not only provided in the form of summary reports indicating school and

² Phelps located 59 large-scale assessment programs which were introduced in various countries and provinces between 1974 and 1999; 22 of these were for accountability purposes; to provide information for “system diagnosis and monitoring” (Phelps, 2000, p.16).

district performance, but the data sets from which these reports were developed can often be obtained through the websites of educational jurisdictions (Sicolý, 2002). Thus, the increased use of large-scale assessments as part of accountability systems has resulted in easier and more widespread access to large-scale assessment data than has previously occurred.

In Canada, where education is a provincial rather than a federal responsibility, every province currently has a large-scale assessment program which is explicitly identified as being used for accountability purposes (Klinger et al., 2008). In Ontario, the focus on large-scale assessment for accountability began in 1996 when the provincial government established the Education Quality and Accountability Office (EQAO). EQAO is an independent, arms-length agency which is governed by a board of directors that is accountable to the public. In other words, EQAO is not under the jurisdiction of the Ontario Ministry of Education. The office was established with a mandate to develop tests to be administered to students across the province, evaluate the public accountability of school districts, and evaluate the quality and effectiveness of education throughout the province (Government of Ontario, 1996). Mandatory province-wide testing began across Ontario soon after EQAO was established. Currently, EQAO administers four assessment programs (EQAO, 2009a):

- the *Assessment of Reading, Writing and Mathematics, Primary Division (Grades 1-3)* administered in Grade 3;
- the *Assessment of Reading, Writing and Mathematics, Junior Division (Grades 4-6)* administered in Grade 6;
- the *Grade 9 Assessment of Mathematics* administered in Grade 9; and
- the *Ontario Secondary Schools Literacy Test (OSSLT)* administered in Grade 10

With about one-third of all students participating in these assessments every year, this is the most extensive province-wide assessment program ever administered in Ontario (Wolfe, Childs & Elgie, 2004). Results from EQAO assessments must be made available to the

public annually (Government of Ontario, 1996) and take the form of school, district and province-wide summary reports.³ In addition, a data portal is available on the EQAO website through which individuals may obtain data for use in conducting their own analyses (EQAO, 2009a).

Rapid Changes in Large-scale Assessment Practices

The increased reliance on large-scale assessments for accountability purposes in recent decades has been accompanied by rapid and continuous changes in assessment methods. These changes include innovative approaches to the design and development of large-scale assessments such as the introduction of new item types and assessment formats as well as new approaches to administering, scoring, and reporting assessment results. For example, many large-scale assessments are moving away from paper-and-pencil formats toward computer-based, continuous testing approaches where test-takers schedule their exam at a time which is convenient for them (Clarke et al., 2000; Drasgow, Luecht & Bennett, 2006). Large numbers of assessment items are needed to support these continuous testing programs and a number of automated systems for item generation and the creation of parallel forms of an assessment are being developed to meet this need (Drasgow et al., 2006).

A number of researchers have suggested that developments in measurement theory lag behind the rapid pace of these changes in assessment methods. For example, Earl and Torrance (2000) caution that “districts, states, provinces, and nations are experimenting with various approaches to assessment, and the knowledge base on large-scale assessment is being developed at the same time as the assessment procedures are being implemented in what amounts to high-stakes field tests” (p.115). Similarly, Koretz and Hamilton (2006)

³ Individual results are provided for students and their parents and are retained in each student’s school file but are not made available to the public.

suggest that recent changes in large-scale assessment “push the field beyond the range of currently available analytical methods” (p.571). The gap between measurement theory and recently developed assessment methods suggests that the implications of these methods for various aspects of assessment including the process of validation may not be well understood.

Recognition of the Impact of Large-scale Assessments

A growing body of research provides evidence of the significant impact of large-scale assessments on student learning, classroom practice and policymaking. For example, Barnes, Clarke & Stephens (2000) examine the impact of two different large-scale assessment programs on secondary mathematics teaching in two states in Australia. Their findings indicate that in the state where the mandatory large-scale assessment included an extended problem-solving task, increased use of problem-solving and investigation in classroom instruction and assessment practices was found as compared with the classroom practices observed in the other state which did not include this type of task in their large-scale assessment program. In another recent example, the consequences of a statewide assessment administered to Grade 4 students in Louisiana are explored (Johnson & Johnson, 2006). The authors describe their experiences teaching in a public school where Grade 4 students who fail the state mandated tests of English and mathematics must repeat the grade. Johnson and Johnson provide a detailed account of the negative effects of this large-scale assessment program on students, teachers and classroom practice. Further, studies investigating the impact of large-scale assessments of secondary students’ literacy in Ontario, Canada (Luce-Kapler & Klinger, 2005; Ricci, 2004) and of students’ writing skills in Alberta, Canada (Slomp, 2008) provide evidence that these assessments lead to policy

changes at the school district level, contribute to a narrowing of teachers' instructional and assessment activities, and result in considerable use of classroom time and school district resources for test preparation activities. As observed by Clarke et al. (2000), "the effects of testing are now so diverse, widespread and serious in the US that we believe it is necessary to establish . . . mechanisms for catalyzing inquiry about and systematic independent scrutiny of them" (p. 177). These studies are just a few examples which demonstrate the impact large-scale assessments can have across a range of grade levels and for a variety of subject areas. For this dissertation, such findings underscore the need to ensure that the interpretations and decisions made on the basis of large-scale assessments can be warranted (Moss, 2007; Kane, 2006; Stobart, 2009). Since the process of validation is the principal means by which interpretations and decisions are warranted and the impact of large-scale assessments can be substantial, the process of validation is a central concern in educational assessment.

Current Practices in the Use of Large-scale Assessments

Within this context of increased reliance on large-scale assessments for accountability and rapid changes in assessment methods, a range of practices in the use of large-scale assessments are occurring. In this section I describe three of these practices: intended and alternate proposed uses, the multiple-scoring of responses to items, and the practice of multiple-use. For each practice, I provide a definition and a hypothetical example and discuss a few instances where the practice is described in the assessment literature. I also offer a visual representation of each practice to illustrate the basic elements of the practice and suggest some ways that these elements may relate to one another. Following each visual representation, I discuss the issues which emerge from each practice with regard to the

process of validation. I introduce these validation issues in a general way and then return to them in chapter 3, where a detailed discussion of the process of validation is undertaken.

In order to describe the characteristics of the three practices, it is necessary to begin by considering the measurement-based concept of an assessment having an *intended use*. As will be shown, this concept is a fundamental consideration in the design and validation of large-scale assessments.

The Concept of an Intended Use

According to current measurement standards that guide the development and interpretation of educational and psychological assessments, assessments are typically designed to support one use, which is often referred to as the proposed or intended use (AERA et al., 1999; Joint Committee on Standards for Educational Evaluation [JCSEE], 2003; Joint Committee on Testing Practices [JCTP], 2004). For instance, an assessment developed to place students in a specific course or level of instruction would not be appropriate as a general measure of achievement in that content area. While a theoretical debate and numerous research initiatives regarding the feasibility of developing assessments suitable for more than one use are underway⁴, there is fairly widespread consensus that assessments that have been developed for one use cannot be assumed to provide valid information for other uses (see, for example, Herman, Baker, & Linn, 2004; Kane, 2006; Shepard, 1997; Wilson, 2007; Wolfe et al., 2004). For many researchers, developing assessments for more than one use creates insurmountable problems because the design decisions made for specific uses are considered to be mutually exclusive (Kane, 2006;

⁴ For example the Commission on Instructionally Supportive Assessment (2001) published a volume entitled *Building tests that support instruction and accountability: A guide for policymakers*. The commission was convened by five organizations (the American Association of School Administrators, National Association of Elementary School Principals, National Association of Secondary School Principals, National Education Association, National Middle School Association) and chaired by James Popham.

Koretz, Stecher, Klein & McCaffrey, 1994; Wiliam, 2007; Wolfe et al., 2004). Others take the view that designing assessments for more than one use is technically demanding and expensive (de Lange, 2007), but can be achieved (Baker & Linn, 2002; Popham, 2002; Wiliam, 2007). Implicit in both perspectives is the view that it is unlikely that an assessment that has been developed for one use will, by happenstance, provide valid information for other uses. And yet, despite this widely accepted principle, large-scale educational assessments are frequently used for purposes other than, or in addition to, their intended use.

Intended and Alternate Proposed Uses

The first of the three practices I describe is one that I refer to as *intended and alternate proposed uses*. This practice refers to the situation that frequently arises after an assessment developed for an intended use becomes widely available and individuals or groups begin to consider the value of the assessment for their particular needs. These individual users may decide to administer an assessment for a purpose which differs from its original intended use. For example, a college may decide that they would like to administer a test of English-language proficiency, originally designed to screen students for admission purposes, to make placement decisions for students taking English-as-a-second-language courses at their institution. While the assessment might be administered by other institutions for admissions purposes (i.e. the original intended use), in this location, the assessment would be administered and used for placement purposes (i.e. an alternate proposed use). A defining characteristic of this practice is that the assessment is administered separately for each alternate proposed use. In addition, alternate proposed uses may be carried out by individuals or groups outside of the agency that developed the assessment and may be

conducted with or without the sanction of that agency. In fact, the assessment developers may not be aware that the alternate proposed use is taking place.⁵

A Visual Representation of Intended and Alternate Proposed Uses

A representation of intended and alternate proposed uses is shown in Figure 1. Each circle in the diagram represents a separate administration of a large-scale assessment. The administration of the assessment for its original intended use is shown as the largest circle even though this may or may not be the most frequent use for the assessment. Additional administrations of the assessment for other uses are shown as smaller circles. Two alternate proposed uses are shown in this diagram but any number of uses may be proposed for an assessment. Each use is based on a separate administration of the assessment and takes place in a different setting. These differences are indicated in the diagram by representing each use with different shading. In addition, each alternate proposed use takes place at a different point in time. Some alternate uses may be proposed shortly after an assessment becomes available while others may be proposed months or years later. To indicate this characteristic, the passage of time is shown as an arrow along the bottom of the diagram. In this practice, the intended use and alternate proposed use(s) are assumed to be independent of one another because they occur in different locations and at different times. There is little reason to expect these uses to interact or have an impact on one another and for this reason no connections are shown between the circles.

⁵ Many individuals and agencies are involved in large-scale assessment practice: those who develop the test, those who publish the test, those who administer and score the test, those who use the test, those who interpret test results, those who take the test as well as the agencies that contract a test developer for a specific assessment. These seemingly separate roles are often combined. For example, the test developer may also be the publisher, administrator and scorer or, in other cases, the test administration and scoring may be carried out by a separate agency or by the test user. This “intermingling of roles” (AERA et.al.,1999, p.1) makes it difficult to ascribe a meaning to the term “test developer”. In this dissertation, I follow the convention of the current measurement standards (AERA et al., 1999) and refer in general terms to test developers, test users and test-takers even though I recognize that these individuals may have overlapping roles.

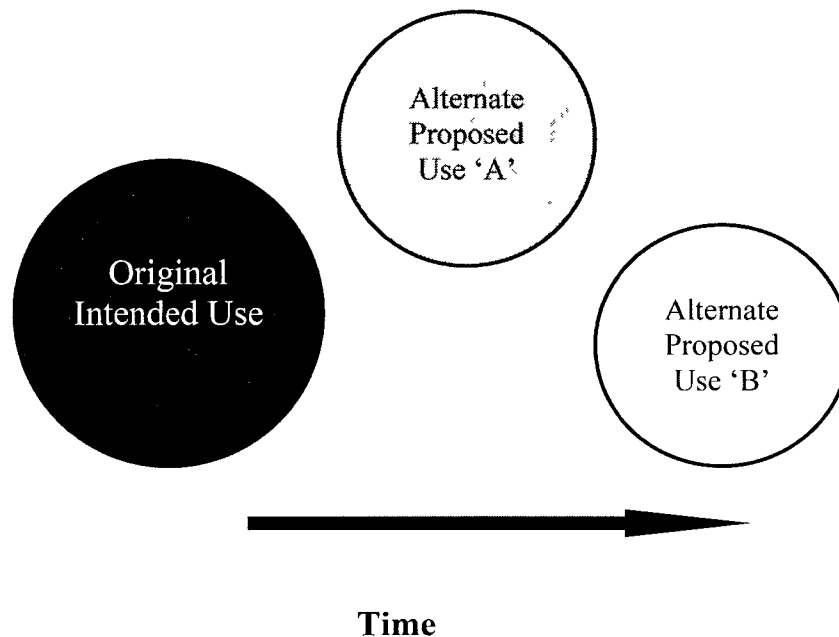


Figure 1. A visual representation of intended and alternate proposed uses.

Implications of Intended and Alternate Proposed Uses for Validation

The practice of intended and alternate proposed uses of large-scale assessments occurs quite frequently and is one reason that measurement standards explicitly state that evidence must be collected to demonstrate the validity of *each* of the uses which may be proposed for an assessment (AERA et al.,1999; JCSEE, 2003; JCTP, 2004). Validity evidence must be collected for the original use and then, as alternate uses for the assessment are proposed, validity evidence for those uses would need to be collected. For the most part, assessment developers have the principal responsibility for validation of the original intended use while those who propose an alternate use are seen as having the responsibility for the validation of that use (AERA et al.,1999; JCSEE, 2003; JCTP, 2004). While some

details regarding the kinds of validity evidence that need to be collected for each alternate proposed use may need to be established, current measurement theory seems to provide an adequate conceptual framework for undertaking the process of validation for this practice.

The Practice of Multiple Scoring

A second practice I want to consider is the multiple scoring of large-scale assessments. In this practice, a student's response to one item on a large-scale assessment is scored more than once and is used to make multiple inferences. According to Ercikan (2002), the multiple-scoring of large-scale assessments occurs quite frequently and most commonly involves the use of a student's answer to an extended response item as an indication of their performance on more than one "assessment unit" (p.8). Assessment units may be different content areas such as mathematics or science, or they may be processes within a content area such as problem solving and communication skills in mathematics. Ercikan (2002) identifies a variety of ways that multiple-scoring can be carried out. For example, a student's response may be scored the first time using a rubric or criteria focused on mathematics content and then scored a second time using a different rubric or criteria focused on science content. Ercikan refers to this approach as "multiple-rubric multiple scoring" (p.9). Ercikan also notes that one large scale assessment may include some items that are scored once and other items that are scored multiple times. Continuing with the example of mathematics and science, responses to some items would be scored and used to make inferences about students' mathematics proficiency, responses to other items would be scored and used to make inferences about students' science proficiency, and responses to a third group of items would be scored twice and used to make inferences about both content areas.

A number of examples of multiple scoring are described by Ercikan (2002) including the Maryland School Performance Assessment Program (MSPAP) and the New York Regents Comprehensive Examination in English. I draw on her discussion of the MSPAP to provide a better sense of how the practice of multiple scoring takes place. Ercikan indicates that the MSPAP consists of a series of performance tasks that are scored multiple times for different content areas and sub-skills. She describes an example from this assessment where each student's written response to one task is scored three times using three different scoring rubrics. Ercikan notes that the assessment criteria for each area are quite different and she includes a portion of each scoring rubric to illustrate the differences. As Ercikan explains, the practice of multiple scoring is problematic because it creates a condition known as local item dependence. That is, most assessments are created using psychometric models that assume that the contribution of each item to the overall test is independent from the other items. The scores which result from multiple scoring are not independent because they are related to the same item prompt and because they are derived from the same student response. The implications of multiple scoring and local item dependence for validation are discussed after a visual representation of this practice is presented.

A Visual Representation of the Practice of Multiple Scoring

A visual representation of the practice of multiple scoring is provided in Figure 2. Unlike the intended and alternate proposed uses shown in Figure 1, with the practice of multiple scoring only one administration of the assessment takes place. This single administration is represented in Figure 2 by the large outer circle. The various ways that student responses are scored are shown within the outer circle. One circle represents responses to items that are scored once for one content area. Another circle represents

responses to items that are scored once for a different content area. Because these responses are scored only once, they are assumed to be independent of one another and no connections are shown between them. Those responses that are scored multiple times using different criteria for each content area are represented as gears. The gears represent the local item dependence that exists between items that are scored more than once. Figure 2 shows items that are scored twice, but multiple scoring can involve scoring a student's response to one item many times. In theory, the degree of local item dependence would increase as the number of times a response is scored increases. In this sense, Figure 2 represents a simple case of multiple scoring. Two additional characteristics of multiple scoring are that each scoring procedure is under the jurisdiction of the assessment developer and that multiple scoring is part of the original design of the assessment. These features are represented in the diagram by the use of one colour for all the ways of scoring students' responses. It is also important to note that in multiple scoring all of the scoring activities take place within the same time frame; the assessment developer oversees each of the scoring activities and these activities take place immediately after the assessment has been administered. For this reason, the passage of time is not indicated in this diagram.

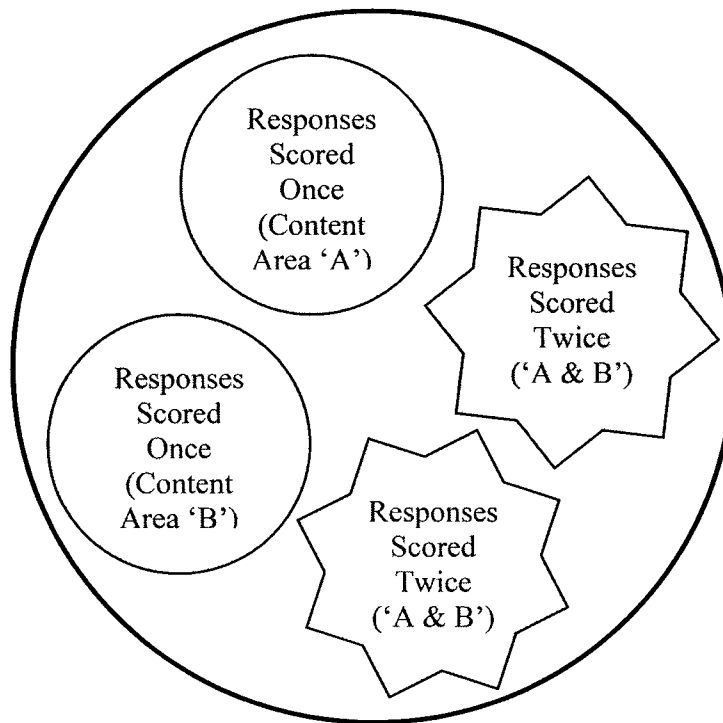


Figure 2. A visual representation of the practice of multiple scoring.

Implications of the Practice of Multiple Scoring for Validation

The practice of multiple scoring has a number of implications for the process of validation. Ercikan (2002) demonstrates that the local item dependence created by multiple scoring can contribute to the underestimation of measurement error, the overestimation of reliability, the potential for multidimensionality, and problems in scaling and score interpretation. While a detailed explanation of these psychometric issues is not necessary for this dissertation, Ercikan indicates that the impact of the dependence between items may be effectively minimized by making appropriate decisions with regard to the procedures for scoring and scaling the items during the design of the assessment.

From a theoretical perspective, current approaches to validation seem to be adequate for the practice of multiple scoring. In multiple scoring the assessment developers are responsible for the scoring procedures being used and they are in a position to investigate and minimize the impact of any local item dependence which may occur. They are also able to gather the necessary evidence to demonstrate the validity of each of the inferences which are made from the multiple scoring of student responses and to make changes to the scoring approach as needed. If some of the scoring was conducted by individuals outside their purview, the assessment developers might not be able to gather the required evidence for validation. Thus, the issues that emerge from the practice of multiple scoring seem to be reasonably well addressed through the application of psychometric techniques and this practice does not seem to raise issues that cannot be addressed using current conceptions of the process of validation. Further, while Ercikan expresses some concern that this practice should be more widely acknowledged and carefully addressed using effective psychometric tools, she does not suggest that multiple scoring requires a different approach to validation.

The Practice of Multiple-use

The third practice in the use of large-scale assessments which I discuss is multiple-use. As stated earlier in this chapter, each large-scale assessment is typically designed to support one intended use and an assessment developed for one use cannot be assumed to be valid for other uses. Despite this basic measurement principle, for many large-scale assessments, a single administration of the assessment may be used in a variety of ways. In this dissertation, I suggest the term *multiple-use* to refer to those instances where a single administration of an assessment is being used for its intended use and for one or more additional uses. A defining feature of multiple-use is that while each use may be suggested

and carried out by different individuals or groups, all of the uses are based on the same administration of the assessment. This is different than the practice of intended and alternate proposed uses where, as I noted earlier, each use is based on a separate administration of the assessment.

Before providing some examples of this practice, I will explain why I use the term *multiple-use*. For much of this inquiry, I was unable to locate any articles which discussed the multiple-use of large-scale assessments. I knew anecdotally that the practice was taking place with the EQAO Grade 9 Assessment of Mathematics as well as with a number of other large-scale assessments but I could not find a specific term referring to this practice in the literature. In an effort to locate an established term for this practice, I sent email inquiries to four experienced large-scale assessment researchers. Their replies suggested that no particular term was in use and as a result, I began to use the term *multiple-use* to refer to this practice. A few months before I completed this manuscript, I located a recent article which discusses a large-scale assessment which is used in multiple ways (Stobart, 2009). In this article, Stobart refers to an assessment that has *multiple purposes*. After reading his article, I was uncertain as to whether I should continue to use the term multiple-use or adopt Stobart's term of multiple purpose. In an effort to resolve this question, I considered how *purpose* and *use* are used in the AERA measurement standards and found that these terms are used interchangeably in most sections (AERA et al., 1999). However, neither *purpose* nor *use* is used in these standards in conjunction with the notion of multiple-use so this observation did not help me decide which term to use. In comparing these terms, it seems to me that *multiple-use* connotes activities which are actually taking place while *multiple purpose* suggests claims about the ways an assessment may be used but does not necessarily signify

uses that are occurring.⁶ In this inquiry I am less interested in claims of assessments with multiple purposes and more focused on the interactions that may occur among multiple-uses that are known to be taking place. For these reasons, I continue to use the term multiple-use to refer to this practice.

Moving beyond the terminology, I will describe some of the different ways that the practice of multiple-use can take place. There are two general categories of multiple-use that I want to identify. In some cases multiple-uses may be carried out by individuals or groups who are outside the purview of the assessment developers. In other instances, multiple-uses may be claimed and/or carried out by the assessment developers themselves. Each category is described to illustrate the issues each kind of multiple-use creates for the process of validation.

Multiple-uses Conducted by Other Individuals or Groups

Some instances of multiple-use involve the use of the results from a large-scale assessment by individuals and groups who are outside the realm of the assessment developers. One example of this kind of multiple-use would be when a researcher obtains the results from an assessment which was administered for its intended use and then uses these results to carry out a research study. Another example of this type of multiple-use would be when an independent organization creates a ranking system for schools based on large-scale assessment results which were gathered for another purpose. Each multiple-use occurs in addition to the use for which the assessment was administered. The assessment developer may or may not be aware of these additional uses and may or may not sanction them. Some of these uses might be viewed as inappropriate and labeled as misuses while others might be

⁶ To be clear, I am not suggesting that Stobart's (2009) use of *multiple purpose* signifies claims about the ways an assessment may potentially be used. Indeed, he seems to be referring to uses that are actually taking place.

considered to be valid uses. Despite these variations, as noted earlier, the defining feature of multiple-use is that the uses all relate to a single administration of the assessment.

Multiple-uses Claimed by Assessment Developers

For some large-scale assessments, a range of uses for the assessment is specified in the manual produced by the assessment developer. In some manuals the intended use is clearly identified and a list of additional uses for the assessment is offered. In other cases, a list of uses is provided but the intended use of the assessment, the one for which it was designed and developed, is not explicitly identified. In both cases, the assessment developers are claiming that the assessment has multiple-uses. A hypothetical example may help to illustrate this kind of multiple-use. Consider an accountability program where the results from a large-scale assessment are principally intended to be used to evaluate the performance of schools and districts. To meet this purpose, the scores are aggregated across schools and districts and compared from year-to-year. However, the assessment developers also claim that the assessment results are valid indicators of individual students' achievement in specific subject areas and provide individual student score reports. In addition, they indicate that another benefit of the assessment is that it models the assessment practices that policymakers want teachers to use and is an effective way to improve teachers' assessment practices. In this example one assessment is being used for three different purposes.

An example of multiple-uses claimed by an assessment developer is found in Stobart (2009) who considers the multiple-use of England's national curriculum assessments. As Stobart observes, the assessment agency responsible for these assessments has indicated 14 different purposes for which these assessments are used. These purposes include assessing

individual students' achievement in specific subject areas; raising standards by setting national targets based on achievement levels on the assessment; judging and comparing school performance; and monitoring changes in standards over time (p.166). Stobart provides evidence that the assessment does not serve all 14 uses equally well.⁷ He also demonstrates some of the challenges which these multiple-uses present for the process of validation. I return to Stobart's discussion of these challenges after I present a visual representation of multiple-use.

A Visual Representation of the Practice of Multiple-use

One way that the practice of multiple-use can be visually represented is shown in Figure 3. In multiple-use, as with multiple scoring, all of the uses are based on the same administration of the assessment. This single administration is represented by the large outer circle and the multiple-uses of the assessment are shown as smaller circles within the outer circle. The intended use is shown using a slightly larger circle than is used for the other uses even though the intended use may not be the most prevalent use for the assessment. In this diagram I have indicated three uses in addition to the intended use but, as Stobart (2009) has shown, many more uses may be associated with an assessment. While each use relates to the same administration of the assessment, some uses may occur at the time of the administration of the assessment while others may take place sometime later. For example, the ranking of schools by an independent organization may take place every year immediately after the assessment results are released to the public while a researcher may use these same assessment results months or years later. For this reason, I have shown the

⁷ There are a number of similarities between the uses of the national curriculum assessments described by Stobart and the uses of the EQAO Grade 9 Assessment of Mathematics. A detailed description of the multiple-uses associated with the EQAO assessment is provided in chapter 6 and I return to a discussion of these similarities in that section.

passage of time as an arrow at the bottom of the diagram. In addition, because an important feature of multiple-use is that some uses may be under the jurisdiction of the assessment developers while other uses involve individuals beyond the realm of the assessment developers, I have represented each of the uses with different shading.

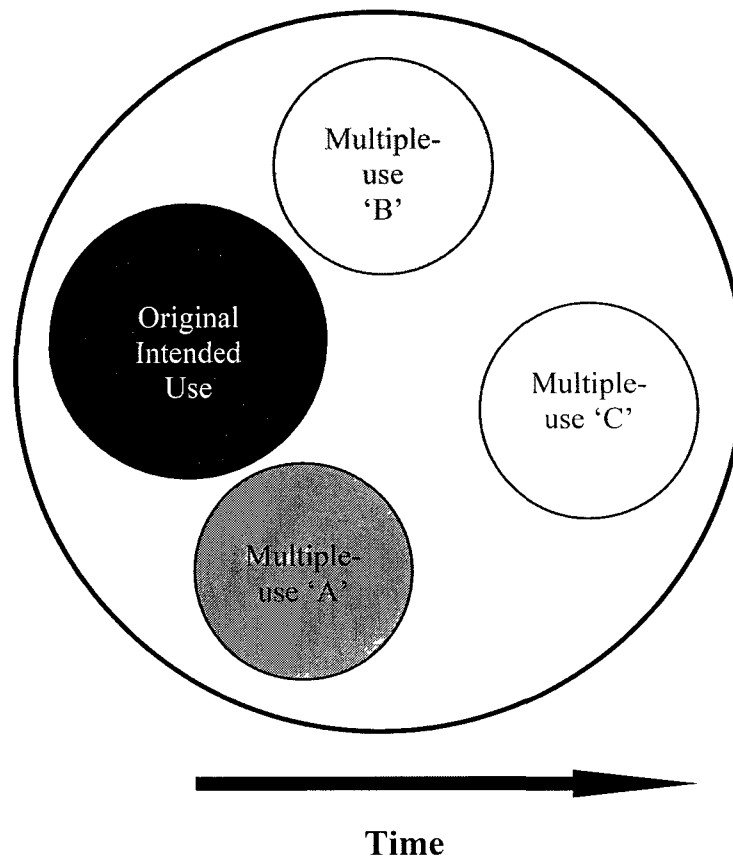


Figure 3. A visual representation of the practice of multiple-use showing four uses.

The Problems with Multiple-use

In this section I review existing research into the implications for validation which come from the multiple-use of large scale assessments. I then identify and describe two problems which emerge from the practice of multiple-use which have not been previously addressed. I argue that these two problems should not be overlooked in the validation of assessments where multiple-use is known to occur.

Existing Literature on the Implications of Multiple-use for Validation

Despite an extensive literature search, I was only able to locate two articles which focus on the implications of the multiple-use of large-scale assessments for validation. The papers I located are Haertel (1999) and, as previously mentioned, Stobart (2009). A summary of each paper is provided in this section.

Haertel (1999) discusses validation for large-scale assessments that have multiple-uses by describing, from a theoretical perspective, how the process of validation might proceed for a fictitious assessment that is used for three purposes.⁸ While he implies that multiple-use is not an uncommon occurrence, he does not identify any specific examples of large-scale assessments where multiple-use takes place. In his example, Haertel draws on an argument-based approach to validation and begins by considering some of the assumptions which would be common to all three uses as well as some of the evidence that might be used to support these assumptions. Haertel then considers the additional assumptions which might be associated with each individual use and suggests some evidence that might need to be collected to support these assumptions. Using this hypothetical example, Haertel effectively demonstrates the plethora of assumptions that would need to be considered as part of the

⁸ Haertel (1999) refers to these assessments as having more than one purpose but he does not use either the term *multiple purpose* or the term *multiple-use*.

process of validation for an assessment with multiple-uses. Haertel suggests that the validation of any large-scale assessment should consider each of the purposes for which the assessment is being used. While he alludes to the challenges that would be inherent in this process, Haertel does not explore the topic any further. In particular, he does not consider the possibility for the existence of interactions between the multiple-uses and he seems to be making a tacit assumption that multiple-uses are independent of one another.

The second paper I located which discusses the implications of the multiple-use of large-scale assessments for validation is Stobart (2009). As noted in the previous section, Stobart (2009) discusses the multiple-use of the national curriculum assessments in England. He describes 14 purposes for which this assessment has been used and argues that given the high-stakes associated with many of these purposes, there is a need for a more extensive validation effort. Stobart maintains that existing approaches to validation do not adequately address the complex situation which is created by multiple-use and he proposes a framework to structure the process of validation for assessments with multiple-uses. Stobart's framework has five main elements: purpose, fitness-for-purpose, reliability, result interpretation, and impact/consequence. In this framework, the validation process begins with identifying all of the purposes for which an assessment is used. After the purposes have been identified, the fitness-for-purpose is evaluated for each purpose by considering what is being assessed and whether the assessment does what it claims to do. The reliability of the assessment system is evaluated followed by a consideration of the interpretation and impact of the assessment results. For both the interpretation of results and the consideration of the impact of the assessment, Stobart indicates that the multiple purposes for the assessment must be considered. He suggests that the validation process include questions such as "How

dependable are the results in relation to the purposes for which they are used?” and “How effectively have the purposes been met?” (p.165). Stobart describes his attempt to use this framework for the national curriculum assessments. His analysis of the validation of these assessments is helpful in showing some challenges for the process of validation arising from multiple-use and his framework provides a means for making the multiple-uses of an assessment more central in the process of validation. However, like Haertel (1999), Stobart seems to assume that multiple-uses are independent of one another. He does not discuss the possibility of interactions between the uses taking place. In reading about Stobart’s framework, it was not clear to me how his approach might be used for large-scale assessments where interactions between multiple-uses are known to exist.

The Issue of Increased Stakes

One issue that emerges from the practice of multiple-use which is not explored by Haertel (1999) or Stobart (2009) is the impact that this practice can have on the stakes associated with an assessment. I contend that the practice of multiple-use effectively raises the stakes associated with an assessment to the extent that assessments which were formerly low-stakes may become high-stakes. That is, as the number of uses associated with a single administration of an assessment increases the stakes the assessment has for students, teachers and other educators may also increase. Koretz and Hamilton (2006) demonstrate that the validation of assessments with high-stakes may need to be approached in a different manner than the validation of assessments with lower stakes. Thus, the practice of multiple-use may result in assessments becoming high-stakes and this may have implications for how the process of validation should be undertaken. While the increased stakes that result from

multiple-use are not the central concern of this dissertation, I see a need for theoretical discussion and empirical research regarding this issue.

The Potential for Interactions among Multiple-uses

A second significant issue that emerges from the practice of multiple-use which has not been identified in the literature is the potential for interactions among the various uses to occur. Given that each multiple-use is based on the same administration of an assessment, I suggest that some uses have the potential to affect other uses. I use the term interaction to refer to the situation which is created when a practice associated with one of the multiple-uses has an impact on the validity of the interpretations which can be made for the other use. I contend that in the multiple-use of assessments the separate validation of each use may not be adequate because intended and additional uses may interact. Validating each of the uses individually, as though they were unrelated, may not be sufficient. For example, consider a large-scale assessment where the results from one administration of an assessment are used to measure students' mastery of course content in mathematics. In addition, the same assessment results are used by some school districts to evaluate the effectiveness of mathematics teachers. I suggest that not only should we be concerned with the appropriateness of the assessment for each of these uses, we also need to be concerned with the impact of the additional use, measuring teachers' effectiveness, on the validation of the original intended use, measuring students' mastery of course content. As part of the validation process we need to ask questions such as: Is there evidence of any interaction between the uses? If so, what is the nature and impact of those interactions on the interpretations which are being made for each use?

The existence of interactions between multiple-uses may relate to a number of factors. Time is one key factor. In multiple-use all of the uses are from a single administration of an assessment but some uses may occur at the same or nearly the same time while other uses may take place after some time has elapsed. As the time between the various uses increases, the potential for interactions among the uses may decrease even though all of the uses relate to the same administration. In addition, the potential for interactions between uses may differ in situations where the test-takers and test users know that the multiple-use will take place at the time the assessment is administered as compared with situations where the multiple-use does not become apparent until sometime after the assessment has been administered. Another factor which may need to be recognized is that *interactions among uses may be important not only in cases where the additional use is considered to be an appropriate use but also in cases where the additional use is characterized as a misuse.* Where multiple-use is known to occur, the misuse of an assessment cannot be assumed to be an isolated practice because the misuse may have a considerable impact on other more legitimate uses of the assessment that are associated with the same test administration. Moreover, I suggest that interactions between uses are likely to become considerably more complex where more than two multiple-uses occur.

Thus, I maintain that multiple-use is a practice that challenges current conceptions of validity not only because of the plethora of assumptions that must be validated for each purpose or use, as noted by Stobart (2009) and Haertel (1999), but because the stakes associated with an assessment increase and because multiple-uses cannot be assumed to be independent of one another. Interactions may occur between any of the uses which are taking place such that one use may have a substantial impact on other uses.

A Visual Representation of the Interaction of Multiple-uses

Figure 4 shows one way that the interactions among multiple-uses may be represented. The large outer circle in Figure 4 represents a single administration of a large-scale assessment. Within this outer circle, the original intended use of the assessment is shown as the largest gear and four additional uses are shown as smaller gears and circles. Some uses may take place at or near the time of the test administration while others may take place considerably later. For this reason, I have shown the passage of time as an arrow at the bottom of the diagram. The potential interactions among the multiple-uses are suggested in Figure 4 by the use of gears rather than circles and by the addition of arrows to indicate the inter-related motion of those gears. In Figure 4 multiple-uses 'A' and 'B' interact with the intended use but not with any of the other uses whereas multiple-uses 'C' and 'D' do not interact with any other uses. In an actual instance of multiple-use, neither the extent nor the nature of the interactions between the uses would be known. For example, one additional use may have minimal impact on the intended use but strongly influence another additional use. In this diagram I have shown a total of five uses but other uses could be added that might interact with any or all of the uses shown. The extent and significance of the interactions among the multiple-uses of an assessment would only become apparent through empirical study. An additional feature of the problem of multiple-use is that each use may be conducted by different individuals or groups; some under the jurisdiction of the assessment developers and others outside their realm. The potential involvement of a variety of individuals and groups is represented in this diagram by using different shadings for each use.

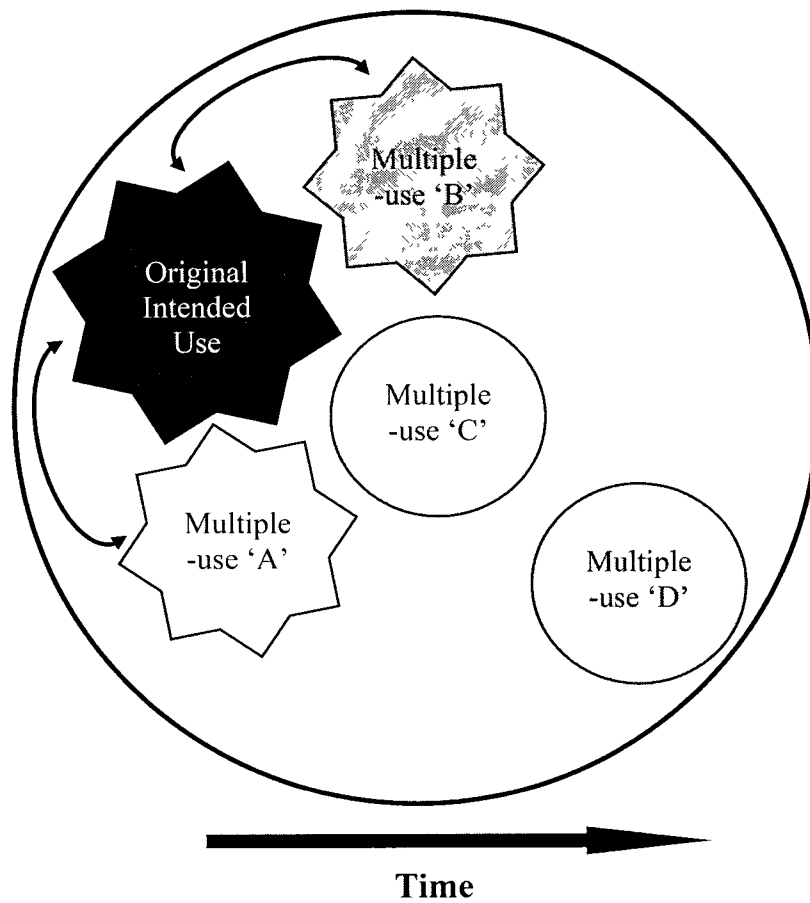


Figure 4. A visual representation of the practice of multiple-use showing interactions.

Overlapping Practices: A Further Complication

In this chapter I have presented three practices in the scoring and use of large-scale assessments. I have also described the problem of the interactions which can occur when there are multiple-uses for an assessment. These practices are an indication of the complex ways that large-scale assessments are being used. However, for any large-scale assessment the pattern of uses may become even more complex because a combination of any of these three practices may occur. Thus, a single large-scale assessment may have intended and alternate proposed uses (as shown in Figure 1), include multiple scoring practices (as shown in Figure 2), be used for multiple-uses (as shown in Figure 3) and evidence of interactions among the uses may be found (as shown in Figure 4).

Using the elements of the previous diagrams, this situation is represented in Figure 5. In this diagram the large outer circle represents a single administration of the assessment and the two smaller circles on the right hand side represent separate administrations of the assessment for alternate proposed uses. Within the single administration of the assessment represented by the large outer circle, the original intended use is shown as the largest gear and the practice of multiple scoring is shown as taking place as part of the original intended use. Four additional multiple-uses from the single administration of the assessment are shown. The interactions among these uses are indicated using gears and arrows in the same manner as in Figure 4. The diagram may seem to suggest a worst-case scenario. However, as will be shown in chapter 6, several of these characteristics are found in the pattern of uses associated with the EQAO Grade 9 Assessment of Mathematics.

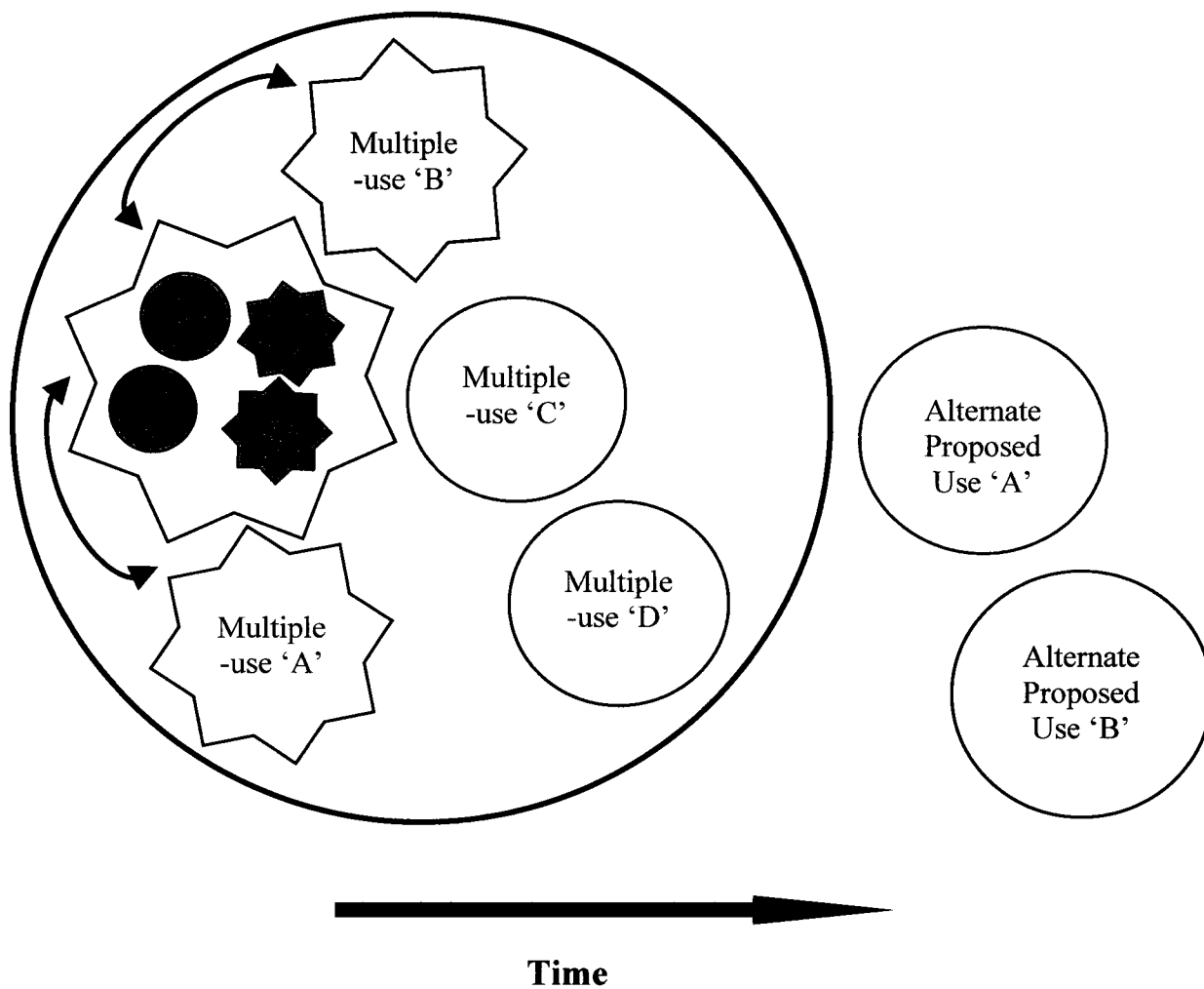


Figure 5. A visual representation of intended and alternate proposed uses, multiple scoring and multiple-use practices including interactions among some uses.

The Prevalence of Multiple-use

To demonstrate the significance of the problem of multiple-use, in this section I summarize my findings regarding the prevalence of the multiple-use of large-scale assessments. Identifying examples of multiple-use in the assessment literature is not a straightforward process. While many published articles mention situations where an assessment is or has been used for more than one purpose, most of these articles provide insufficient detail about whether or not the additional use is based on the same administration of the assessment as the intended use. As a result, it is difficult to establish, with certainty, whether a particular account refers to an alternate proposed use or is an example of multiple-use. In spite of this difficulty, I located several examples that, in all likelihood, refer to the practice of multiple-use.

The earliest mention of multiple-use I located refers to an instance in the mid-1900s when, despite the objections of teachers, high school entrance examinations designed to screen students for admission were also used to monitor teachers' performance in several American states (Tyack, as cited in Koretz & Hamilton, 2006). Teachers were concerned that the sample of students who wrote these assessments might not adequately represent all the students they taught and they argued that the assessments should not be used to compare teachers or schools. These concerns suggest that the assessment had not been administered a second time for the purpose of measuring teachers' performance but that the results from the original administration of the assessment to students entering high schools were being used for an additional purpose. For this reason, I suspect that this is an early example of the practice of multiple-use.

More recently, Phelps (2000) indicates that the results from assessments administered to provide information for international comparative studies are also used within some countries to monitor and diagnose the performance of their schools. Phelps provides examples of this practice occurring with results from the Programme for International Student Assessment (PISA) and from the Third International Mathematics and Science Study (TIMSS). Because these international comparative assessments are regulated by organizations outside the purview of the educational systems of individual countries, it is unlikely that the additional use, to monitor and diagnose school performance, was from a separate administration of these assessments. Thus, these assessments provide additional examples of multiple-use.

Several other examples can be found in the assessment literature of the past decade. Wilson (2007) indicates that results from assessments administered as part of accountability programs and designed to provide information about student learning at the aggregate level are frequently used as sources of information about specific classrooms or to provide information to guide the instruction of individual students. Paradoxically, results from achievement tests designed to provide information about individual students are also combined and used as measures of school effectiveness in some American states (Popham, 2002). In a comparable instance, de Lange (2007) describes how results from high school exit exams administered in many countries to determine students' mastery of course content and eligibility for continued studies are also used to assess individual teachers and to compare and monitor the progress of schools. Most recently, as noted, Stobart (2009) describes the multiple-uses associated with England's national curriculum assessments.

These examples suggest that the multiple-use of large-scale assessments is neither a recent development nor a rare occurrence.

Factors Which May Encourage Multiple-use

A number of factors may enable the multiple-use of large-scale assessments to occur and increase the frequency with which this practice takes place. In this section three factors are considered: the availability of large-scale assessment data, the high cost of large-scale assessments, and the vague indication of intended interpretations and uses for many assessments. These factors are discussed to further illustrate the importance of investigating multiple-use.

Availability of Large-scale Assessment Data

The results from large-scale assessments, particularly those associated with accountability programs, are often widely disseminated through print media and the internet (Koretz & Hamilton, 2006; Sicolý, 2002; Volante, 2007). Results are readily available for individual students, classrooms, schools, and districts.⁹ The availability of this data means that teachers, administrators, researchers, the media, policymakers and others are able to use assessment results in a variety of ways. For example, Sicolý (2002) was able to collect school-level data for 13,000 schools from 14 large-scale assessment programs administered in various Canadian provinces and American states. He analysed this data using multiple-regression techniques to study the degree of common variance among mathematics, reading and writing scores on these assessments. Sicolý indicates that, in most cases, the data was available in downloadable digital files easily accessed on the department of education website for the province or state. While the widespread availability of large-scale assessment

⁹ The identity of individual students is typically not revealed but anonymous individual student results are often available as data sets.

results facilitates potentially valuable research regarding these assessments, this accessibility can also encourage their multiple-use.

Another example of the use of publicly released assessment data is the use of Ontario's province-wide assessment results by the Fraser Institute in their annual publications which rank all elementary and secondary schools in the province (Cowley & Easton, 2008a; 2008b). With regard to data sources, Cowley and Easton state "we base our overall rating out of 10 on the school's performance on six indicators, *all of which* [emphasis added] are derived from province-wide tests of literacy and mathematics skills that are administered by the province's Education Quality and Accountability Office" (2008b, p.5). Thus, the ranking system developed by the Fraser Institute is entirely based on publicly available EQAO assessment results. And yet, EQAO does not sanction the use of their assessments for ranking purposes. In fact, they explicitly state "EQAO is opposed to the ranking of schools or school boards" (EQAO, 2009b) and describe their concerns with the use of their assessments for this purpose. Others have also expressed their concern with the Fraser Institute's use of the EQAO assessments for ranking purposes¹⁰ (Canadian Psychological Association Working Group on Test Misuse, 2000; Earl and Torrance, 2000). Nonetheless, because the assessment results are released to the public, the Fraser Institute has been able to continue with this practice for several years. While it may be tempting to classify the Fraser Institute's use of these assessment results for ranking as a misuse of the assessment and discard the practice as inappropriate, I suggest that the impact of this

¹⁰ The Canadian Psychological Association and the Canadian Association of School Psychologists released a strongly worded joint position statement expressing their concerns with the use of province-wide assessments for ranking purposes and with the reporting of these rankings by the media (Canadian Psychological Association Working Group on Test Misuse, 2000). In addition, Earl and Torrance (2000) note that the media often reports these rankings even though EQAO had a public awareness campaign to encourage appropriate interpretations and uses of their assessment and to discourage the use for ranking purposes.

additional use on other uses of the EQAO assessment may be considerable and should not be ignored. The Fraser Institute's ranking practice may increase the stakes associated with the EQAO assessment and may have an impact on the way other uses of the assessment are interpreted and understood. However, no empirical studies of the impact of this use on other uses of the EQAO assessment were located. I return to a discussion of the Fraser Institute's use of the EQAO Grade 9 Assessment of Mathematics in chapter 6.

Costs of Large-scale Assessment Programs

The development, administration, scoring and dissemination of results for large-scale assessments represent a considerable expense in dollars and in lost instructional time. These high costs are another factor which may encourage the tendency toward multiple-use. Cronbach notes that "a test with multiple uses can repay a cost that would be unreasonable if a test were to be used once and forgotten" and observes that "in educational testing the information is usually placed in the student's file and used for many later decisions: admissions, choice of major, selection of courses, career planning, etc." (1971, p. 496). Cronbach points out the pragmatic merits of using a single administration of an assessment for multiple purposes but he does not address the implications for validation that may come from these multiple-uses. The tendency to offset the high costs of large-scale assessments by making an assessment serve multiple purposes has also been articulated by the National Council of Teachers of Mathematics (NCTM, 1995) with regard to mathematics assessment. NCTM cautions that "the pressure to make a single assessment serve multiple purposes is likely to increase. Consequently, special vigilance may be needed to assure that all the uses to which assessment information is being put are in harmony with the purposes of the assessment" (1995, p. 21). Thus, the connection between the high costs of large-scale

assessments and the multiple-use of results from those assessments seems likely but the implications of these multiple-uses are not known.

Ambiguous Intended Interpretations, Purposes and Uses

A third factor which may increase the frequency of multiple-use practices is that the intended interpretations, purposes and uses of large-scale assessments are not always clearly discernable. This lack of clarity has been frequently observed in the assessment literature. Ambiguity regarding the appropriate inferences and intended uses for a large-scale assessment can arise from the name or phrase that is provided to students, parents and others to identify what an assessment measures (Cronbach, 1971; Kane, 2006; Shepard, 1993). In addition to the misconceptions associated with the name of an assessment, a number of researchers have identified the tendency of assessment developers to only vaguely indicate the intended interpretations and uses for their assessments in manuals and technical documents. As Koretz and Hamilton (2006) observe “performance standards and test guidelines leave considerable room for uncertainty and disagreement about intended or appropriate inferences” (p. 546). Many researchers have suggested that assessment developers should be obligated to explain the precise meaning of the assessment data they release in score reports produced for schools and districts and when results are released to the public (Baker & Linn, 2002; Nagy, 2000). Baker and Linn maintain that assessment programs that make results available to the public should be required to provide “clear explanations of legitimate and potential illegitimate interpretations of results” and to ensure that media personnel are informed of these limitations (2002, p.23). They also contend that the media has a responsibility to communicate information about the appropriate interpretations and uses of an assessment to the public in an effective manner.

Further ambiguity in the intended uses for a large-scale assessment can be created if the intended purpose for an assessment changes after the assessment is developed. The tendency for intended uses to shift and, in many cases, to expand to encompass additional uses has been observed for a number of large-scale assessment programs. For example, Klinger et al. (2008) consider the purposes for provincial large-scale assessment programs across Canada. The authors identify the four main purposes for these programs as accountability, gatekeeping, instructional diagnosis and monitoring student achievement over time. They observe that many of the provincial assessments were initially developed for one of these purposes but additional purposes are often added after the assessments are developed. Further, their review of these assessment programs indicates that the assessment items and formats have not been substantively changed to address the additional purposes (2008, p.28). Other researchers have made similar observations for assessments in other contexts (Nagy, 2000; NCTM, 1995) and the shifting and expansion of assessment purposes was identified as an important concern for the EQAO assessment program (Wolfe et al., 2004). As the intended use for an assessment shifts or expands, the incidence of multiple-use for that assessment is also likely to increase. The significance of the expanding intended uses of EQAO's assessments is described in greater detail in chapter 6.

The Need for Research into Multiple-use

Given the incidence of multiple-use found in the assessment literature, the factors which suggest that multiple-use may increase, the limited existing research into the implications of this practice for validation, the increased stakes which result from multiple-use and the potential for interactions among multiple-uses, I suggest that there is a need for additional research in this area. As noted, I was able to locate only two articles, Haertel

(1999) and Stobart (2009), which consider the implications of multiple-use for the process of validation and neither article mentions the possibility of interactions among the uses. To address this gap in the literature, empirical studies of instances where the multiple-use of a large-scale assessment is known to occur are needed. Through such studies, the nature and extent of the interactions which may be occurring between multiple-uses may be better understood. Continued theoretical exploration of this issue is also needed to better understand various aspects of the problem of multiple-use and to suggest techniques for analyzing the complex interactions which may occur. As noted in the introduction, this dissertation is an initial foray into both the empirical and theoretical aspects of this issue.

CHAPTER 3

THEORETICAL FRAMEWORKS FOR THE PROCESS OF VALIDATION

In this chapter, I present the two theoretical frameworks for the process of validation which I draw on in this inquiry. My thinking about validity theory and the process of validation was initially rooted in traditional measurement-based views. I was introduced to this view in an undergraduate psychometrics course and became much more familiar with it during my studies in educational measurement at the Masters level. After completing my Masters' degree, I continued to follow the dialogue on validity in educational and psychological measurement and, for a number of years my research was based on this perspective. Accordingly, as I began this inquiry into the process of validation in the context of multiple-use, my thinking was strongly influenced by measurement-based views. To explain this theoretical perspective, I present some basic tenets of measurement-based approaches and then focus on Kane's argument-based model of validation (Kane 1992; 2006), which is one of the models I apply to the multiple-use of the EQAO Grade 9 Assessment of Mathematics in chapter 7.

During my doctoral program, my view of validity has expanded. In particular, I have been influenced by validity theorists who value the use of locally situated studies as a means of reconceptualizing validity (Kane, 2006; Messick, 1989; Moss, 2007) and who call for the expansion of measurement-based approaches by drawing on other disciplines (Messick, 1989; Mislevy, 2007; Mislevy, Moss & Gee, 2009; Moss 2003, 2007). From among these validity theorists, I draw mostly on the work of Moss who proposes opening the question of validation to the wider interpretive scope of philosophical hermeneutics and sociocultural studies (2003; 2007; Moss et al., 2006). Within this emphasis on interpretive approaches to

validity, Moss suggests that the concept of boundary objects, developed in sociology and information systems theory, might be a useful analytic technique for the process of validation. I describe Moss's perspective in this chapter and in chapter 7, I apply the concepts of boundary objects (Star & Griesemer, 1998; Bowker & Star, 1999) and boundary encounters (Cobb et al., 2003) to my observations regarding the multiple-use of the EQAO Grade 9 Assessment of Mathematics.

Clarifying Validity-related Terms

Before presenting these two theoretical frameworks, in this section I will clarify my use of the terms *validity*, the *process of validation* and *validity theory*. Tittle (1989) states "validity is a construction based on the frameworks of researchers, philosophers, test publishers (test developers), and of the teacher and learner in the classroom" (p.6). I interpret this statement as an expression of the contingent nature of the concept of validity, and while I see this contingency as precluding an unequivocal definition of validity, I have attempted to arrive at a working definition.

One means of defining the concept of validity is to find what Toulmin (2003; 1958) refers to as a "tangible counterpart" or referent for the term (p.63). This approach is often used in educational and psychological measurement. For example, in the 1985 measurement standards the terms *appropriateness*, *meaningfulness* and *usefulness* are used to define validity (AERA et al.). Other validity theorists in measurement refer to validity as the *soundness* of interpretations made from measurements (Cronbach, 1971, p.443; Moss et al., 2006, p. 109) or to the *adequacy and appropriateness* of inferences based on measurements (Messick, 1989). A different, but related, approach to defining validity is to suggest alternate terms. For example, in discussing research methodology, some social science researchers

take the view that *validity* has strong positivist associations. These researchers prefer not to refer to validity and suggest various alternate terms such as *balance, fairness, completeness and sensitivity* (Wolcott, 1990, p.122), *trustworthiness* (Mishler, 1990), *goodness* (Schwandt, 1996) or *credibility, plausibility and integrity* (Erickson, 1986). For me, using either explanatory or alternate terms as a means of defining validity is not particularly satisfactory in that it leads to an infinite regress of continually seeking new words to clarify or explicate other words.

Toulmin (2003; 1958) articulates the problem of defining concepts such as validity in a manner which I find useful. He includes validity along with other abstract nouns such as truth, knowledge, and goodness and maintains that finding tangible counterparts for such words is misguided (2003; 1958, p. 63). In his view, establishing what abstract nouns signify is an endless process with attempts to articulate a meaning inevitably leading to a circumscribed view of the concept. However, Toulmin does not suggest that the matter of defining abstract nouns be left completely unresolved. Rather, he suggests that the meaning of such words may be more effectively derived from the way they are used in practice. Toulmin notes that practical ideas are often better expressed using verbs, adverbs and adjectives and he contends that abstract concepts may be best understood through the verbs, adverbs and adjectives used in reference to them. Toulmin illustrates this point using the example of the abstract concept of *time*. He notes that we can measure the amount of time it takes to do something with varying degrees of precision and that we can communicate this information clearly. These actions help us to understand the way the concept of time is used in everyday contexts even though they contribute little or nothing to understanding or defining the concept of time itself (2003; 1958, p.64). Stated more simply, while we don't

know what *time* is, one way we can get a sense of this concept is by considering the way it is used in practical situations. In a similar vein, rather than trying to settle on a definition of what validity *is*, it may be more useful to focus on the ways this concept is used.

Drawing on both Tittle (1989) and Toulmin (2003; 1958), I think of validity as an elusive and abstract concept with shifting meanings that are contingent on the where the term is being used. Although I continue to think about possible meanings for validity and to consider the merits of various comparable terms, my focus is less on defining the concept of validity and more on the process of *validating* or *doing validity*; of understanding the ways that the process of validation takes place in diverse assessment settings. In my writing, I try to avoid using the noun *validity* and would prefer to use verbs such as *validating* or *doing validity*. However, while validating and doing validity are expressions that are closer to my meaning, these expressions frequently result in awkward sentence structures. Accordingly, I have settled on using the expression *the process of validation* as a way of avoiding the abstract noun *validity* as well as the awkwardness of terms such as validating and doing validity.

Indeed, preference for the use of either validity or validation has oscillated within the measurement literature of the past few decades. One way to look at the use of these terms is to consider the title of the chapter concerned with validity that has appeared in successive editions of *Educational Measurement*; a volume which is widely recognized as seminal in the field of measurement. In the first edition of this book, Cureton (1951) used the title “Validity”; in the second edition, Cronbach (1971) titled his chapter “Test Validation”; Messick (1989) wrote the “validity chapter” in the third edition and returned to “Validity” for his title, while Kane (2006) titles his chapter “Validation” in the current edition of this

volume. The use of these terms in measurement standards can also be considered. The first standards which were published in 1954 by the American Psychological Association used the term validity and the most recent standards document (AERA et al., 1999) also refers to validity rather than validation. Given the varied use of these terms in the literature, in this dissertation, I use the term validity when I am referring to ideas from the standards or when I am citing authors who refer to validity but, in writing about my own ideas, as noted, I refer to the process of validation.

My thinking with regard to the process of validation in educational assessment has also been influenced by the ways that validation is characterized in social science research methodology where it is often viewed as a means of legitimating knowledge and judging the quality of research (Kaplan, 1997; Lather, 1993, 2001; Lincoln, 1995; Lincoln & Guba, 2000). In addition, some social science methodologists consider validation as a way of critically reflecting on the inquiry process and of increasing the transparency of research methods (Erickson, 1986; Moss, 1998). I see these perspectives as a means of complementing and extending traditional measurement-based conceptions of validation. Thus, for me, the process of validation is a means of legitimating and judging the quality of information that comes from an assessment. Further, I think that critically reflecting on the assessment process and increasing the transparency of assessment methods are worthwhile ways of conceptualizing the process of validation.

I also want to clarify my use of the term *validity theory*. Delandshere (2002) suggests that theories are not static truths but “frames for temporary understanding” (p.1469) and Moss (2003) describes theory as “an intellectual framework or set of conceptual tools that shape both our understanding and our actions” (p.13). Drawing on these ways of

characterizing theory, I see validity theory less as an attempt to define or circumscribe the concept of validity and more as a framework to guide my thinking about this concept and about the process of validation. I do not seek to unequivocally define the concept of validity nor do I advocate an approach where validity must be wholly redefined for each assessment context. The approach I take in this inquiry is to begin with some generally accepted principles from validity theory and then to consider how these principles might be revised or expanded given the specific assessment uses I am investigating. I see validity theory as dynamic and continually emerging over time and in response to various circumstances with this dissertation participating in and contributing to that process.

Measurement-based Conceptions of the Process of Validation

Having traced my emerging view of the process of validation and clarified my use of validity-related terms, I now describe some basic tenets of measurement-based conceptions of validation. As noted earlier, these measurement-based approaches are the theoretical starting point for my inquiry because they are the perspective of validation I was most familiar with as I began this research.

Centrality of Validity in Measurement

Validity is a central concept in educational measurement; described as “the most fundamental consideration in developing and evaluating tests” (AERA et al., 1999, p. 9) and as “the single most important issue in student evaluation” (JCSEE, 2003, p. 128). Evidence of the prominence of validity in measurement can be seen in the chapters devoted to this topic in standards documents (AERA et al., 1999; JCSEE, 2003; JCTP, 2004), in the discussion of validity in key measurement texts (e.g. the four editions of *Educational Measurement* described in the previous section) and in the special issues of journals which

regularly appear on the topic (e.g., *Educational Measurement: Issues & Practice*, 1997; *Social Indicators Research*, Nov 1998; *Educational Researcher*, Nov 2007 etc.).

Conceptions of validity and validation have continually evolved over the course of the 20th century through an ongoing dialogue among researchers and practitioners.¹¹ Cronbach and Meehl (1955) capture an early stage of this dialogue noting that “writers on validity during the preceding decade had shown a great deal of dissatisfaction with conventional notions of validity, and introduced new terms and ideas, but the resulting aggregation of types of validity seems only to have stirred the muddy waters” (p.281). In many ways, Cronbach and Meehl’s comment would have been apt at any point between 1955 and the present for, just as they observed over fifty years ago, there are currently many distinct conceptions of validation and new terms and approaches are continually being suggested. Some recent papers that demonstrate the range of perspectives within measurement include Borsboom, Mellenbergh and van Heerden (2004) who define validity using a causal theory of measurement, Slomp and Fuite (2005) who employ mathematics and information theory in their discussion of validity, Lissitz and Samuelsen (2007) who call for changes in terminology and an increased emphasis on content in the process of validation, and Mislevy (2009) who looks at validity from the perspective of model-based reasoning. A comprehensive description focusing on the distinctions among measurement-based approaches to validation is neither feasible nor necessary for this dissertation. Rather, I provide a description of some widely accepted characteristics of measurement-based validity by drawing on seminal articles (notably Cronbach, 1971; Kane, 1992, 2006; Messick, 1989;

¹¹ An appreciation of the historical development of validity theory is helpful for understanding current approaches but is beyond the scope of this dissertation. Many interesting accounts are available (see, for example, Jonson & Plake, 1998; Kane, 2006; Lissitz & Samuelsen, 2007; Messick, 1989; Moss et al., 2006; Shepard, 1993; Sireci, 2009).

Moss, 1992; Moss et al., 2006; Shepard, 1993) and by considering current editions of measurement standards (AERA et al., 1999; JCSEE, 2003; JCTP, 2004).

Common Characteristics of Measurement-based Conceptions of Validation

In this section I describe several characteristics which are common to most measurement-based conceptions of validation: the validation of inferences; validation as an ongoing process; validation as the integration of multiple sources of evidence; validation as a value-laden process; the role of consequences in validation and validation as a shared responsibility. While other characteristics of measurement-based conceptions of validation might be included in this chapter, I focus on these six characteristics because each has relevance to the practice of multiple-use. I provide a brief explanation and supporting references for each characteristic and then describe how it relates to multiple-use.

The Validation of Inferences

A central idea in measurement-based conceptions of validation is that the process of validation pertains to the inferences which are made from an assessment rather than to the assessment itself. This distinction is made in measurement standards (AERA et al., 1999; JCSEE, 2003; JCTP, 2004) and has been reiterated by many validity theorists (see, for example, Cronbach, 1971; Kane, 2001; Messick, 1989; Moss, 1994; Shepard, 1993; Sireci, 2007; Zumbo, 2007). Shepard (1993) notes that this distinction was first expressed in the early 1900s and Kane (2006) describes how validity practitioners gradually shifted from the notion of validating a test to validating the interpretations and uses of a test score over the period from 1955 to 1989 (p.22). Accordingly, for most measurement-based validity theorists, validity is not viewed as a quality inherent in a test. Rather, the process of validation is viewed as a means of evaluating the inferences which are made from an

assessment. The validation of inferences rather than tests is important because it underpins the notion that validity evidence must be collected for *each* proposed inference or use of an assessment (AERA et al., 1999; JCSEE, 2003; JCTP, 2004). As Zumbo notes, “the validity of the inferences one makes from test scores is bounded by place, time, and use of the scores resulting from a measurement” (2007, p. 48).

While the view that it is the inferences that are being validated is fairly widely held among measurement-based validity theorists, it is not unanimous. In a special edition of *Educational Researcher* (Nov, 2007), Lissitz and Samuelson argue that validation should focus principally on the internal psychometric characteristics of an assessment such as the test definition and development process, test content and test stability. They take the view that an evaluation of these characteristics should form the principal justification for an assessment and maintain that these characteristics “exist independent of, or regardless of, the application of the test or the use of the test in some theoretical formulation” (2007, p.446). The remaining papers in this special edition respond to the ideas put forward by Lissitz and Samuelson. While the responding authors indicate that they agree with some points made by Lissitz and Samuelson, they do not support the idea that validity can reside in the test itself (see for example, Embretson, 2007; Gorin, 2007; Moss, 2007; Sireci, 2007). Their lack of agreement with Lissitz and Samuelson and the reasons they articulate for the validation of inferences being important are further evidence that the validation of inferences is a characteristic of most measurement-based views of validation.

The emphasis on the validation of inferences rather than tests themselves is very relevant for the process of validation in the context of multiple-use. In multiple-use, a single administration of an assessment is used for the intended use and for one or more additional

uses. Rather than attempting to validate the assessment itself, it is the inferences which are being made for the various uses that must be considered in the process of validation and, as described in chapter 2, validating each use independently as though the other uses were not taking place may not be adequate because interactions between the uses may not be accounted for.

Validation as an Ongoing Process

In most measurement-based conceptions, validation is viewed as an ongoing process involving evaluative judgments made on the basis of accumulated evidence (Cronbach, 1988; Kane, 2006; Messick, 1989). Validity is not a fixed entity that can be established once and left to stand. The evaluative judgments that are the basis of validation may change over time as new evidence is gathered or as the context where the assessment is used changes.¹² In addition, validity is expressed as a matter of degree rather than as a judgment of valid or invalid. Given these characteristics, adequate validation evidence for an assessment cannot be provided in one technical guide or a single published paper. Rather, validation evidence is gathered as part of an ongoing process that may be likened to a research program or inquiry. Ideally, the accumulation of evidence begins as an assessment is being developed and continues as long as the assessment continues to be used.

Kane (2006) distinguishes between two stages in the process of validation referring to them as the development stage and the appraisal stage (p.25). In the development stage evidence to support the intended interpretations and uses of the assessment scores is established. Much of the evidence in the development stage is gathered by the assessment developers. The appraisal stage of validation begins when the development process is

¹² This is an example of the use of context to signify a bounded setting, which as I observe in chapter 1, seems to be the sense in which context is most often used in measurement literature.

considered complete. Individuals or groups external to the assessment development process critically evaluate the proposed interpretations and uses of the assessment. Kane emphasizes the need for both stages of validation and observes, along with other validity theorists, that validation efforts beyond the development stage have been inadequate for many large-scale assessments (Shepard, 1993; Tittle, 1989). I refer to Kane's development and appraisal stages of validation in chapter 4 when I discuss existing validity evidence for the EQAO Grade 9 Assessment of Mathematics and in the closing chapter when I consider characteristics of the process of validation that should be reconsidered to better address the practice of multiple-use.

Considering validation to be an ongoing process is clearly appropriate for the practice of multiple-use. During the development stage assessment developers may gather validity evidence related to the intended use of an assessment. However, it is during the appraisal stage that multiple-uses become evident. Further, because of the potential for interactions among the multiple-uses, as the pattern of uses for an assessment changes, judgments about the appropriateness of the inferences made for either the intended use or the additional proposed uses may also change.

Validation Requires the Integration of Multiple Sources of Evidence

A third characteristic common to most measurement-based views of validity is that validation requires the integration of multiple sources of evidence. This view has been expressed by many validity theorists (Cronbach, 1971; Messick, 1989; Shepard, 1993; Kane, 2006). For example, Kane states "the validity argument as a whole requires the integration of different kinds of evidence from different sources" (2006, p.23). The kinds of evidence required for a validation program differ depending on the interpretations or uses that are

being made from the assessment and both logical argument and empirical studies may be included. Some examples of these sources of evidence include correlational and experimental studies, expert judgment, and the results of previous research (Shepard, 1993, Kane, 2006). Both Tittle (1989) and Haertel (1999) maintain that another source of evidence that should be gathered as part of the validation process for many large-scale assessments is information about how teachers and students interpret and use assessment results. In addition, Haertel suggests that focus groups should be conducted with the public to learn how large-scale assessment scores are being interpreted and evidence from these focus groups should be included in the process of validation.

With regard to the practice of multiple-use, the integration of multiple sources of evidence is of critical importance. Given that the kinds of evidence that need to be collected in the process of validation depend on the uses and interpretations being made from an assessment, both the intended use and the additional uses would need to be considered in determining what evidence should be collected. In addition, I argue that where interactions among the multiple-uses are identified, specific evidence relating to those interactions also needs to be collected. As will be shown, the importance of considering the perspectives of many different individuals and groups when gathering and integrating multiple sources of evidence is critical in the process of validation for assessments where multiple-use is known to occur.

Validation is a Value-laden Process

There is fairly broad agreement among measurement-based validity theorists that values play a central role in the process of validation. Messick (1989) illustrates the central role of values in the assessment process using the example of assessments used to screen

students for admission into post-secondary studies. He points out that these assessments may be designed to select students on the basis of their ability, their effort, their prior accomplishments and achievements or their educational or economic needs. Each of these value judgments would result in the creation of a different assessment and these value judgments need to be considered as part of the process of validation. Many validity theorists argue that value judgments must be made explicit as part of the process of validation (Cronbach, 1988; Messick, 1975, 1989; Moss, 1998; Shepard, 1993).

For multiple-use, value judgments are associated with the intended use as well as with each additional use. The value judgments that are inherent in each use must be considered in the process of validation as well as the potential for tensions to arise between the value judgments inherent in one use and those that are made for another use.

Considering Consequences in Validation

Most current measurement-based views acknowledge that the consequences of the use of an assessment should be considered as part of the practice of validation. Messick (1989) is often credited with drawing attention to the importance of considering both the intended and unintended consequences that come from the interpretations or uses of assessments. Examples of the kinds of consequences that might be investigated as part of the practice of validation for large-scale assessments are ‘Does the inclusion of a writing sample in an admissions test for post-secondary studies increase attention to the teaching of writing in some high school courses?’ or ‘What changes to teaching and learning might follow from an increased reliance on multiple-choice items on achievement tests?’ (Shepard, 1993, p. 426). For some validity theorists intended and unintended consequences are considered to be an integral part of validity (Messick, 1989, 1994, 1998; Shepard, 1997; Wiliam, 2004). For

others, these consequences are seen as an important consideration in the process of validation but not as a defining characteristic of validity (i.e. Mehrens, 1997; Popham, 1997). Moss (2003) and Zumbo (2007) point out that the precise role of consequences in the concept of validity remains controversial. Despite this debate, there is fairly broad consensus that consequences must, at the very least, be considered as a factor in the process of validation.

Considering intended and unintended consequences in the process of validation has relevance to multiple-use in two distinct ways. First, the impact which one use may have on another use, which I see as a significant concern arising from multiple-use, might be characterized as an unintended consequence. For example, in chapter 2, I describe an instance of multiple-use where results from high school exit examinations administered to determine students' mastery of course content and eligibility for continued studies are also used to assess the performance of individual teachers and to compare and monitor the progress of schools (de Lange, 2007). In this instance of multiple-use there is an intended use (students' eligibility for continued studies) and two additional uses (assessing teacher performance, comparing and monitoring school progress). One way to think about the interaction among these three uses is to consider the interactions to be unintended consequences which each use may have on the other uses.

There is another sense in which the concept of consequences relates to multiple-use. Some of the interpretations and uses associated with a large-scale assessment might be described as consequences rather than uses. For example, while I consider the use of assessment results by an external agency for ranking purposes to be a use, I think that the claim made by assessment developers that an assessment is an effective way to improve

teachers' assessment practices might be better described as an intended consequence rather than as a use. Similarly, I see using large-scale assessment results as the basis for selecting schools to be included in a new educational initiative as a use for an assessment while encouraging teachers to implement specific aspects of the curriculum by including them on the large-scale assessment may be an intended consequence.

I return to a discussion of intended and unintended consequences and their relevance to the practice of multiple-use when I discuss the pattern of uses associated with the EQAO Grade 9 Assessment of Mathematics in chapter 6.

Validation is a Shared Responsibility

Although there are debates about what aspects of validation fall to each group, there is fairly broad agreement among measurement-based validity theorists that large-scale assessment developers, those who use large-scale assessments, as well as other stakeholders in education must share the responsibility for the process of validation in an effort to ensure that the interpretations that are being made can be supported (Cronbach, 1971; Kane, 2006, 2008; Messick, 1989; Shepard, 1997). Assessment developers are expected to indicate the intended purpose for an assessment and to provide “an explicit, coherent interpretive argument linking test performances to the proposed interpretations and uses” (Kane, 2006, p.26). However, Shepard notes that “when users appropriate tests for purposes not sanctioned and studied by the test developers, users become responsible for conducting the needed validity investigation” (1997, p.8). In an earlier paper she also cautions “often, users of tests are not qualified, or lack the necessary resources, to conduct validity investigations” (Shepard, 1993, p.445). The responsibility of test users to conduct validation studies when they use an assessment for purposes other than those indicated by the test developers is also

identified in the measurement standards (AERA et al., 1999) and a similar concern for these test users ability to conduct these studies is expressed. Thus, while responsibility for validation may be shared, not all test users may be in a position to effectively undertake the process of validation or believe it to be their responsibility.

For the practice of multiple-use, the shared responsibility for validation and the limited capacity or commitment of some assessment users to conduct validity investigations have many implications. While the assessment developer should probably accept much of the responsibility for validation of the intended use, how responsibility for the validation of additional uses should be allocated is not clear especially in cases where there is evidence of interactions between the multiple-uses. In such cases, neither the assessment developer nor the assessment users acting in isolation may be in a position to uncover the interactions among uses which may be occurring or to gather the information needed to fully appreciate the implications of those interactions.

Validation for Large-scale and Classroom Assessments

The six characteristics of measurement-based conceptions of validation described in the previous section are more often considered with reference to large-scale assessment than with reference to classroom assessment activities. A considerable body of research about the influence of large-scale assessments on classroom assessment practices has been published (Abu-Alhija, 2007; Barnes, Clarke & Stephens, 2000; Koretz et al., 1994; Lock, 2001; Tierney, 2006; Wiliam, 2007; Wilson, 2007). In this research, these two realms of assessment are typically seen as discrete endeavours requiring different conceptions of validation. For example, Tittle (1989) notes that the language and activities of the measurement community are not the same as those of teachers and students in classrooms

and Moss (2003) identifies several ways that the process of validation for classroom assessment may differ from validation for large-scale assessments. Moss points out that the interpretations and uses of classroom assessments are often temporary as students' learning is ongoing and she observes that in classroom assessment practice teachers often draw on multiple sources of evidence. Given these characteristics, Moss contends that the validation of classroom assessments is primarily concerned with how various assessment practices fit together into a coherent interpretation. However, she observes that this focus is not a key element in the validation of large-scale assessments.

The difference in conceptions of validation for large-scale and classroom assessment practices presents a challenge for the instance of multiple-use examined in this dissertation. The multiple-use of the EQAO Grade 9 Assessment of Mathematics includes the use of this large-scale assessment for accountability purposes as well as for classroom assessment purposes as part of students' grades. In this particular instance of multiple-use, classroom and large-scale assessment practices may not only influence one another (as suggested by the literature cited in the previous paragraph), they are to some extent one and the same activity. Accordingly, theoretical approaches to the process of validation that are relevant to only one of these assessment categories are of rather limited value for this instance of multiple-use. This concern is re-visited and further elaborated in my discussion of the validation of the multiple-uses of the EQAO Grade 9 assessment in chapters 7 and 8.

Kane's Model of Argument-based Validation

As part of this inquiry I wanted to consider the adequacy of a measurement-based conception of validation for the multiple-use of the EQAO Grade 9 Assessment of Mathematics. To do this, I needed to select one measurement-based validation model from

the range of measurement-based conceptions of validation suggested in the literature (some of which are mentioned earlier in this chapter). After extensive reading, attending a conference on current conceptions of validity¹³ and some deliberation, I decided to focus on *validity as argument* (Cronbach, 1988; Kane, 1992, 2001, 2006; Shepard, 1993; Wiliam, 2004). Validity as argument has been discussed by a number of validity theorists but its origin and development are principally attributed to Cronbach (1971, 1988) and Kane (1992, 2006).¹⁴ The most fully articulated description of this approach is Kane's model of argument-based validation (2006). In this section, I provide a detailed description and visual representation of Kane's model. My reasons for choosing this model for validating the multiple-use of the EQAO Grade 9 assessment are explained in chapter 7 in the section where I attempt to apply Kane's model to this instance of multiple-use

An Overview of the Model

In his 2006 chapter, Kane traces the development of validity as argument acknowledging the earlier work of Cronbach and others. He offers a theoretical rationale for his approach by drawing on Toulmin's model of inference. Kane notes that validity theory in the measurement field has often provided a conceptual framework for validation but has typically fallen short of including explicit guidance for validation practice. With his model, Kane intends to offer a "pragmatic approach to validation" (p.18).¹⁵ Kane considers validation to be the process of evaluating the plausibility of the proposed interpretations and

¹³ In October 2008, I attended a conference entitled "The Concept of Validity Revisions, New Directions and Applications" which was held at the University of Maryland. At the conference a number of validity theorists discussed their work and engaged in a critical dialogue on validity. Subsequently, proceedings from the conference were published (Lissitz, 2009).

¹⁴ Validity as argument is reflected in the AERA et al. measurement standards where validity is described as a scientific argument. Alternatively, Shepard (1993), drawing on the work of Cronbach, refers to validity as evaluative argument.

¹⁵ As I want to be careful to represent Kane's model as accurately as possible, throughout this section I frequently paraphrase Kane (2006). These passages have been identified by indicating page numbers.

uses of an assessment (p. 17). In his model, the validation process consists of two phases: an *interpretive argument* and a *validity argument*. In the interpretive argument, the proposed uses and interpretations of an assessment are specified and the inferences and assumptions being made are explicitly identified. In the validity argument, logical and empirical evidence is gathered to evaluate the plausibility of these inferences and assumptions and of the reasoning that connects these elements. The interpretive argument and the validity argument are the main elements of Kane's model and are described in more detail as follows.

Interpretive Arguments

According to Kane, validation begins with an explicit statement of the proposed interpretation of test scores. For Kane, interpretation refers to the meaning which is assigned to the test score (p.18). A test score may have several legitimate interpretations and may be used to make different kinds of decisions. However, an interpretive argument must be developed for each interpretation and a rationale for the relevance of the test-score interpretation to the proposed use must be provided (p.29). The importance of this initial step is clearly communicated by Kane when he notes that "a failure to state the proposed interpretations and uses clearly and in some detail makes a fully adequate validation essentially impossible" (p.60).

Once the proposed interpretation(s) and associated use(s) have been established, the development of the interpretive argument can proceed. In the interpretive argument, all of the inferences which lead from the observations (e.g. performance on the assessment items) to the conclusions or decisions based on these observations are identified. An inference typically takes the form of an *if-then* rule such as *if* the performance has certain characteristics *then* the observed score should have a certain value or *if* the score is above a

certain cut-point *then* the student is admitted (p.23). Each interpretive argument involves a number of inferences some of which may be fairly apparent while others may be difficult to identify. Further, these inferences may be related to one another in the sense that conclusions of earlier inferences may serve as starting points for later inferences (p.27). Kane notes that a failure of any one of the inferences undermines the interpretive argument as a whole even if convincing evidence for the other inferences is provided (p.34).

A number of assumptions are inherent in each inference. Some assumptions will be readily apparent while others will be largely hidden or implicit in the inference. In the process of validation each of these assumptions must be clearly articulated. In addition, for each assumption within an inference, the *warrant* or basis of the assumption and the *backing* or evidence for that warrant must be provided. Warrants for each inference in the interpretive argument are likely to be different from one another and may take many forms. For example, Kane suggests that one of the warrants for a scoring inference may take the form of a scoring rubric and the backing for that warrant might include expert judgment regarding the rubric (p.34). In addition, conclusions from one inference may provide the warrant for another inference in the interpretive argument thereby forming a chain of reasoning within the interpretive argument.

In Kane's discussion of the nature of interpretive arguments, he describes two characteristics which I believe have particular relevance for the practice of multiple-use: (i) interpretive arguments are tentative and defeasible and (ii) interpretive arguments are subject to the assumption of *ceteris paribus*. Each characteristic is explained as follows.

Interpretive Arguments are Tentative and Defeasible

Drawing on Toulmin's model of inference, Kane indicates that interpretive arguments rely on formal reasoning as well as on informal or presumptive reasoning (2006, p. 27). In formal reasoning a proposition is considered valid if the conclusions can be generated from the premises according to the rules of inference. For Kane, formal reasoning plays an essential but limited role in interpretive arguments. Informal or presumptive reasoning focuses on coherence and the plausibility of inferences and assumptions. These inferences and assumptions must be supported with empirical evidence. Since interpretive arguments rely more strongly on presumptive reasoning, these arguments are always subject to empirical challenge and are always somewhat tentative, often explicitly so. For example, Kane notes that for large-scale assessments these indications of uncertainty can take the form of standard errors of measurement or confidence intervals (p. 27). Interpretive arguments are also described as *defeasible* meaning that even when they are accepted in general they can be overturned for a particular case (p. 27). Kane notes that interpretive arguments cannot be unequivocally proven, rather, they are "rigorously evaluated" (p. 29).

The Assumption of Ceteris Paribus

Kane states that "every interpretive argument includes a general assumption to the effect that nothing has interfered with the proposed interpretation" (p.28). He associates this assumption with the concept of *ceteris paribus* or "all else being equal" from the philosophy of science. This assumption of "all else being equal" is made for the interpretive argument as a whole. According to Kane, when this assumption is not met, such as when unusual circumstances interfere with the proposed interpretation, an exception to the interpretive argument needs to be stated. I return to a discussion of the significance of interpretive

arguments as tentative and defeasible and the assumption of *ceteris paribus* in chapter 7 when I discuss my attempt to apply Kane's model to the multiple-use of the EQAO Grade 9 assessment.

Kane's Examples of Interpretive Arguments

A more complete understanding of interpretive arguments can be gained by considering the examples that Kane (2006) includes in his chapter. Among the four examples he includes are two that are particularly relevant here: an interpretive argument for an accountability assessment and an interpretive argument for classroom assessment activities.

An Interpretive Argument for an Accountability Assessment

Kane (2006) suggests four components that might be included in the development of an interpretive argument for an accountability assessment. In the first component of the argument an initial *semantic interpretation* of each student's performance is made. By semantic interpretation, Kane is referring to the meaning which is assigned to the assessment results in terms of the standards or curriculum objectives on which the assessment is based (p. 51). He outlines a number of the inferences and assumptions which might be inherent in this semantic interpretation. Once the semantic interpretation is established, the conversion of scores to achievement levels must be considered and this is the second component of the interpretive argument. For many accountability assessments three or four achievement levels are established and referred to using terms such as 'basic', 'proficient' and 'advanced'. The inferences and assumptions which are made when converting scores to these achievement levels must be identified. A third component of the interpretive argument for an accountability assessment is to identify the inferences which are made in computing

percentages for each achievement level in each of the grades being assessed. These are the inferences which must be made in order to claim that, for instance, 54% of students in Grade 4 at a given school or in a given school district achieved a level of 'proficient'. The fourth component of the interpretive argument relates to the decisions which are made based on the percentages at each achievement level. Kane's example refers specifically to the accountability assessments which are implemented as part of the No Child Left Behind policy currently in place in the United States. In this context, decisions are made in the form of sanctions or rewards based on a school's results (p.53). The inferences required for making these decisions as well as the assumptions, warrants and backing for those inferences must be made explicit in the interpretive argument.

Interpretive Arguments for Classroom Assessments

An interpretive argument for classroom assessment activities differs in a number of ways from an interpretive argument developed to validate the use of a large-scale assessment. Kane (2006) maintains that teachers do not begin with a semantic interpretation or an operationally defined scoring inference, as is the case for an accountability assessment. Rather, he suggests that teachers use their own conceptual frameworks and experience to integrate available evidence about their students into a coherent view of each student (p. 51). For Kane, a teacher's conceptual framework includes their views concerning the subject area, appropriate pedagogical theories and effective instructional techniques. Kane suggests that this process can be made explicit as an iterative interpretive argument which consists of three stages (p.48). In the first stage, the teacher develops an initial view of their students based on their conceptual framework and on the evidence they gather as they interact with students during instruction and assessment activities. In the second stage there is a

refinement of the teachers' view of the student based on ongoing interactions. The third stage involves the extension of the teacher's view of the student to new contexts and decisions such as granting credit in a course so that a student may continue on to a subsequent course. For Kane, in the interpretive argument for a classroom assessment, the teacher's conceptual framework provides the warrants for the inferences in each of these three stages (p. 49). The backing for these warrants comes from the teacher's training, qualifications and experience as well as from the ongoing dialogue the teacher has with other teachers, parents and the students themselves (p. 49). Thus, while interpretive arguments for classroom assessment differ considerably from interpretive arguments for large-scale assessments, in Kane's view an argument-based model of validation may be used for both types of assessment.

Validity Arguments

Once the interpretive argument is established, the development of the validity argument begins. The validity argument is a critical appraisal of the interpretive argument (Kane, 2006, p. 29). The validity argument begins with an evaluation of the clarity and coherence of the interpretive argument as a whole. Once this has been established each inference in the interpretive argument is evaluated to determine if it is reasonable and each assumption is examined to assess its plausibility. The warrants and the backing for the inferences and assumptions are also reviewed. Multiple types of evidence, both analytic and empirical, may be collected in the evaluation of each of the inferences and assumptions in the interpretive argument. The types of evidence used in a validity argument depend on the specific inferences, assumptions and warrants being made and may include expert judgment, empirical studies, results of previous research and the identification of value judgments (p.

25). Kane also indicates that ruling out alternative hypotheses is often an appropriate part of the validity argument. While the interpretive argument sets the parameters for what needs to be addressed in the validity argument, the validity argument evaluates the interpretive argument for a particular context (p. 29).¹⁶

A Visual Representation of Kane's Model of Argument-based Validation

One way to visually represent Kane's model of argument-based validation, is shown in Figure 6. At the top of this figure the proposed interpretation and the proposed use are shown. These two elements are connected with an arrow to indicate that the relevance of the interpretation to the use must be stated at the beginning of the validation process. In the interpretive argument, the relevance of the interpretation to the use is made more explicit by specifying the inferences which are being made. In Figure 6, I have shown three inferences but any number may be made for a given interpretation or use. The assumptions inherent in each inference must also be made explicit. I have shown three assumptions for each inference but, in practice, the number of assumptions on which an inference is based may vary. Curved arrows are included along the left side of the figure to show that inferences and assumptions may be connected to one another. Thus, a chain of reasoning among inferences and assumptions may be created in the interpretive argument.

The second phase of the model, the validity argument, is shown on the right side of Figure 6. A large bracket is included in the figure to indicate that the validity argument involves a critical evaluation of all the elements in the interpretive argument. The activities involved in developing a validity argument are summarized in Figure 6 with the three brief statements on the right-hand side of the diagram. This representation of the validity

¹⁶ In his argument-based model of validation, Kane repeatedly emphasizes that validity arguments are context specific. In my view, his use of the notion of context is as a bounded setting or backdrop for the assessment activity that has taken place (see chapter 1).

argument may be somewhat misleading in that in practice the development of a validity argument is a substantial undertaking which requires the sustained efforts of assessment developers, researchers and assessment users (Kane, 2006). The elements of the validity argument could be represented in this diagram in greater detail. However, as will become apparent in chapter 7 when I attempt to apply Kane's model to the instance of multiple-use I investigate, it is the components of the interpretive argument that have the most relevance for validating the multiple-use of the EQAO Grade 9 Assessment of Mathematics. For this reason, I have simplified the representation of the validity argument in Figure 6.

As noted earlier, measurement-based conceptions of the process of validation are the theoretical starting point for this inquiry and are the basis from which I established the purpose of this research. Indeed, from the outset of this project I have been interested in querying these measurement-based conceptions of validity and in considering how adequate they may be for the process of validation where the multiple-use of large-scale assessments is known to occur. As noted, Kane's argument-based model of validation is the measurement-based approach I consider for validating the EQAO Grade 9 Assessment of Mathematics and I present my findings with regard to this approach in chapter 7. However, as the inquiry proceeded I became increasingly aware of various calls that have been made to re-conceptualize validity theory. In the next section, I summarize the calls to re-conceptualize validity theory that have influenced my thinking about the process of validation and have been taken up in this inquiry into the implications of multiple-use.

Phase 1: Interpretive Argument

Proposed Interpretation → Proposed Use

Inference 1

- Assumption 1
 - warrant(s) & backing
- Assumption 2
 - warrant(s) & backing
- Assumption 3
 - warrant(s) & backing

Inference 2

- Assumption 1
 - warrant(s) & backing
- Assumption 2
 - warrant(s) & backing
- Assumption 3
 - warrant(s) & backing

Inference 3

- Assumption 1
 - warrant(s) & backing
- Assumption 2
 - warrant(s) & backing
- Assumption 3
 - warrant(s) & backing

Phase 2: Validity Argument

1. Evaluate clarity & coherence of interpretive argument
2. Collect evidence for warrants & backing
3. Rule out plausible alternative hypotheses

Figure 6. A visual representation of Kane's argument-based model of validation.

Re-conceptualizing Validity Theory

Over the past few decades a number of researchers have perceived a need for validity theory to be re-conceptualized to provide more adequate guidance for the practice of validation (Jonson & Plake, 1998; Lissitz & Samuelson, 2007; Messick, 1989; Moss, 2007). More specifically, these researchers have observed that while validity theory provides a conceptual framework that is useful for the development of measurement standards, it does not identify the amount and kinds of evidence that need to be gathered in the process of validation for an assessment. This limitation has been identified as being particularly problematic for a range of specific assessment practices such as high-stakes assessments (Koretz & Hamilton, 2006), locally situated uses of large-scale assessments (Moss et al., 2006), assessments with multiple purposes (Stobart, 2009) and classroom assessment activities (Brookhart, 2003, 2004; Moss, 2003; Shepard, 1993). In this section I consider two approaches which have been suggested for re-conceptualizing validity theory: the value of locally situated studies of large-scale assessment practices and the value of drawing on disciplines beyond educational measurement such as sociocultural studies. Both suggestions have influenced my view of the process of validation and are key elements of this dissertation.

Value of Locally Situated Studies

A number of researchers have identified the value of locally situated studies of large-scale assessment practices in contributing to the re-conceptualization of validity theory (Kane, 2006; Messick, 1989; Moss, 2007). While theoretical discussions of large-scale assessment practice are necessary, valuable insights for validity theory can also be gained through empirical studies of the ways that large-scale assessments are actually administered,

interpreted and used in specific settings. As mentioned in the introductory chapter, Shepard effectively expresses this idea when she states that research that is “embedded in the dilemmas of practice” (2000, p. 13) may be particularly useful for learning about and understanding complex issues. In a similar vein, Moss suggests that “richly contextualized cases of practice” (Moss, 2003, p.14) can be helpful in expanding existing conceptual frameworks. She maintains that by illustrating how validity theory may be instantiated in practice, locally situated studies can reveal the limitations of established principles, provide opportunities to critique these principles and suggest new principles to expand existing theory (p. 21).

I concur with these researchers in terms of the value of locally situated studies for expanding conceptions of validity theory. As I see it, the multiple-use of the EQAO Grade 9 Assessment of Mathematics creates a dilemma of practice that raises challenging issues for the process of validation. Conducting a locally situated study of this instance of multiple-use may provide insights that will contribute to a better understanding of the process of validation for other instances of multiple-use and may suggest ways of building on existing conceptions of validation to better address the practice of multiple-use.

Value of Drawing on Disciplines beyond Educational Measurement

Another approach which has been suggested as worthwhile in the re-conceptualization of validity theory is to consider perspectives and incorporate ideas from disciplines beyond educational measurement. This approach has been suggested by a number of researchers including Messick (1989), Mislevy (2007; Mislevy, Moss & Gee, 2009) and Moss (2003; 2007). Mislevy indicates that he adopts a cross-disciplinary perspective in his research because “familiar test-development and analytic methodologies were falling short

for the challenges posed by new forms of assessment” (Mislevy et al., 2009, p.69). Mislevy indicates that disciplines such as cognitive science, psychology, statistics, and philosophy inform his thinking about validity theory and build on the perspective of validity he has drawn from measurement-based psychometrics. For Moss, valuable insights for re-conceptualizing validity theory can be found in interpretive social science, hermeneutics, and sociocultural studies (Moss, 2003, 2007; Moss et al. 2006). Indeed, Moss has been working for some time toward developing what she refers to as “multi-discourse” practices in educational assessment (Mislevy et al., 2009, p.79).¹⁷ Moss’ view of the value of drawing on these disciplines and some of the analytic tools she suggests have substantially influenced this dissertation. The key points I have drawn from her writing with regard to re-conceptualizing validity are summarized in the next section.

The Value of Sociocultural Studies in Reframing Validity Theory

Moss maintains that measurement-based approaches to validity, including Kane’s argument-based model, are appropriate and effective for validating the intended uses of a large-scale assessment across a broad range of educational settings. However, she argues that these approaches are less effective and relevant when validating the specific local uses of large-scale assessments (Moss et al., 2006). In re-conceptualizing validity theory, Moss identifies “the need for an expanded conception of validity to support the interpretive work in which teachers and other education professionals routinely engage” (p.128). She envisions a theory of validity that is flexible and dynamic and argues that hermeneutics and sociocultural studies may contribute substantially to developing such an approach.

¹⁷ Although this citation may appear incorrect, the Mislevy et al. paper is written as a conversation among Mislevy, Moss and Gee. Each author takes clearly identified turns in the conversation and these turns appear as separate passages in the text. Thus, the quotation is properly attributed to Moss.

In this dissertation I focus primarily on Moss's suggestions with regard to the value of sociocultural studies in reframing validity theory. However, the benefits of adopting a hermeneutic perspective in the process of validation which Moss describes also seem to be a promising means of addressing many of the problems which multiple-use creates for the process of validation. For this reason, in this section I summarize Moss's ideas with regard to hermeneutics and the process of validation before presenting her discussion of the value of sociocultural studies in reframing validity theory.

For Moss, adopting a hermeneutic perspective, which is broadly defined as the theory and practice of interpretation, is a "theoretical resource" (Moss, 2003, p. 18) that can contribute to developing an interpretive approach to validity. Drawing on Gadamer and others, Moss indicates that being aware of one's bias, looking for coherence in what others are saying and believing that you have something to learn from others are prerequisites for understanding in philosophical hermeneutics (Moss et al., 2006, p.130). As Moss explains, the focus is not on arguing against the other person but on questioning to bring out the strengths in the other's argument and the aim of philosophical hermeneutics is to understand and learn from differences rather than to arrive at consensus. Drawing on a long philosophical lineage, Moss describes the hermeneutic circle:

Hermeneutics supports a holistic and integrative approach to interpretation of human phenomena, which seeks to understand the whole in light of its parts, repeatedly testing interpretations against the available evidence, until each of the parts can be accounted for in a coherent interpretation of the whole.

(Moss, 2003, p.14)

In this way, interpretations that are appropriate to the available evidence and are situated within a local setting can be made and the process through which the interpretations were

created can be described along with any remaining questions and insufficiencies in evidence. Such an approach allows others to evaluate the validity of the interpretations for themselves. Moss et al. also describe Habermas' *depth hermeneutics* where the goal is to explain a text and various interpretations of that text in terms of the conditions that produced the interpretations; to "seek to connect meaning/interpretation to the social structures that shaped it" (2006, p.133). Depth hermeneutics not only helps to illuminate the social influences that shape an interpretation, it also acknowledges the need to consider the consequences of people's interpretations.

Moss argues that these aspects of hermeneutics have direct relevance to validity theory and the process of validation. She sees hermeneutics as particularly useful in providing a framework for integrating a variety of types of evidence when developing an interpretation, in appreciating multiple interpretations, in making those interpretations mutually comprehensible to the various individuals involved and in illuminating the social forces that shape interpretations of assessments. For Moss, hermeneutics is a means for better understanding the multiple interpretations inherent in educational assessment practice and can contribute to the development of a more interpretivist approach to validity theory. Moss maintains that a more interpretivist approach is a worthwhile way to build on views of validity that have been conceptualized in traditional educational measurement approaches.

From my perspective, adopting a hermeneutic approach may be quite worthwhile for the process of validation in the context of multiple-use. As I explained in chapter 2, the multiple-use of large-scale assessments takes place in various settings and involves multiple interpretations of a single administration of an assessment by a range of individuals or groups. Given the potential for interactions to occur between the multiple-uses, I suggest that

the consequences of these different interpretations must be considered as part of the process of validation. In addition, the social forces that shape these interpretations must be made clear as the validation proceeds.

Although I consider hermeneutics as a worthwhile area to pursue with regard to the implications of multiple-use for the process of validation, early in this inquiry process as I was just beginning to read about hermeneutics it was not readily apparent to me how I might apply hermeneutic concepts to this instance of multiple-use. I understood hermeneutics more as a philosophical perspective or epistemological stance and less as an analytic tool. Accordingly, in this dissertation I decided to focus on the use of some specific analytic tools from sociocultural theory which might be used in the process of validation. My understanding of hermeneutics has expanded since that time and while hermeneutics did not become a major focus of this dissertation, I do return to this topic in the final chapter where I suggest some of the ways hermeneutics may contribute to the process of validation for multiple-use situations.

As noted earlier in this chapter, Moss identifies sociocultural studies, in addition to hermeneutics, as a valuable theoretical perspective for developing an interpretive approach to validity (Moss et al., 2006; Mislevy et al., 2009). With the term sociocultural, Moss encompasses the sociocultural theory of mediated action (drawing primarily on the work of Wertsch, 1995, 1998), cultural-historical activity theory (drawing primarily on the work of Engeström, 1993, 1999, 2001), and the situated theory of learning (drawing on the work of Lave and Wenger, 1991). Moss sees ideas from these areas within sociocultural studies as useful both for suggesting the types of evidence that may need to be gathered in the process

of validation for educational assessments and for providing ways to analyze the complex environments where educational assessment takes place (Moss et al., p.111).¹⁸

To explain how sociocultural studies can contribute to validity theory, Moss begins by drawing on the work of Wertsch and describing the idea of mediated action. She explains that mediated actions are actions which are carried out by individuals or groups using a variety of mediational means (Moss et al., 2006, p.138). These mediational means may be physical (i.e. texts, tools etc.) or symbolic (i.e. concepts, language systems etc.) and include both the cultural tools that a community inherits and new cultural tools that a community creates. Cultural tools link the actions which are carried out by individuals and groups in a community to the cultural, institutional and historical settings where those actions take place (p.138). Moss suggests that artefacts of assessment such as classroom assessment activities and large-scale assessments can be viewed as cultural tools which mediate actions within educational settings.

Moss et al. (2006) consider mediated activity to be an integral aspect of the concept of a community of practice, suggested by Lave and Wenger (1991), as well as the concept of an activity system, as conceived by Engeström (1993, 1999) and they explore the possibilities of looking at assessment as taking place within a complex activity system or within a community of practice. Moss et al. do not equate activity systems with communities of practice, but consider how each might contribute to understanding educational assessment. In particular, Moss et al. suggest that the tools and techniques used to analyse activity systems as well as some approaches used in the analysis of a community of practice might also be useful in the process of validating the interpretations, decisions and actions

¹⁸ These three aspects of sociocultural studies are distinct and fully developed discourses which, in another context, would merit a much more extensive discussion. However, here, my goal is to explain how Moss draws on these ideas to re-conceptualize measurement-based approaches to validity theory.

that are made on the basis of large-scale educational assessments. These tools and techniques may suggest the kinds of questions which need to be asked as well as appropriate ways of gathering evidence in the process of validation. For example, Moss et al. state “the full characterization of any particular activity system is typically constructed through empirical work involving observations, interviews, artefact analysis, and so on” (p. 141). Moss et al. maintain that assessment activities can be analyzed in a similar manner and in this way these approaches which are rooted in sociocultural studies can contribute to the process of validation.

Large-scale Assessments as Boundary Objects

In re-conceptualizing validity theory, Moss et al. consider assessment activities that take place within an activity system or community of practice but also acknowledge that many large-scale assessment activities cross from one community of practice to another (Moss et al., 2006). That is, a large-scale assessment may be used by classroom teachers at the same time as it is used by school districts or state-level educational policymakers. Moss et al. suggest that the analytic concept of “boundary objects” (Bowker & Star, 1999; Star & Griesemer, 1989) may be particularly useful for understanding assessments that are used in more than one community. Moss et al. describe how boundary objects inhabit different activity systems and enable communication and cooperation across these systems: “a boundary object is a particular kind of cultural tool that not only crosses boundaries of activity systems, such as mandated assessments, but also is plastic enough to adapt to local needs while maintaining a common identity across sites” (Moss et al., p.146). Boundary objects can mean different things in each realm but the different meanings can be reconciled. Moss et al. suggest that considering large-scale assessments as boundary objects may

provide additional insights for validity theory and the process of validation. Moss states “a large-scale assessment would function as a boundary object when actors in the local context are able to cooperate in providing necessary information to outsiders while maintaining a productive level of authority and agency over their own practice” (Mislevy et al., 2009, p.97).

As I read Moss’ discussion of large-scale assessments which are used in more than one community of practice and her idea of looking at a large-scale assessment as a boundary object, I was struck by how relevant this approach might be to my observations of the multiple-use of large-scale assessments and to the multiple-use of the EQAO Grade 9 assessment in particular. Thus, the Moss et al. (2006) article provided the impetus for my decision to look at the multiple-use of the EQAO Grade 9 assessment in terms of the concept of boundary objects. Once introduced to the idea, I sought out more information beginning with Star and Griesemer (1989), as cited by Moss et al., and then looking more broadly at references to boundary concepts in sociocultural theory. A more detailed description of boundary objects and the related concept of boundary encounters (Cobb et al., 2003) is included in chapter 7 (where I apply these concepts to the multiple-use of the EQAO Grade 9 assessment) as these boundary concepts are a separate topic from theoretical conceptions of the process of validation which is the focus of the current chapter.

CHAPTER 4

CONTEXTUALIZATION OF THE CASE STUDY

Moving away from a theoretical discussion of multiple-use, I now focus on the empirical study of multiple-use which I conducted. As described in chapter 3, in-depth studies of local practice have been suggested as an effective means of further developing validity theory. In keeping with this view, my case study of the multiple-use of the EQAO Grade 9 Assessment of Mathematics is intended to provide insights into the implications of the practice of multiple-use for the process of validation. In this chapter, I describe my efforts to contextualize this study by providing information about Grade 9 mathematics instruction and assessment in Ontario as well as a summary of the development, administration and validation of the EQAO Grade 9 assessment. In addition, I review those studies which mention teachers' use of the EQAO Grade 9 assessment as part of students' grades which is one of the two multiple-uses I focus on in this study.

The information presented in this chapter not only informed my research design, it emerged from and continually influenced the process of data construction in the study. For example, knowledge of current assessment policies and practices for Grade 9 mathematics helped me develop questions for the interviews I conducted with teachers and with EQAO personnel. At the same time, conducting those interviews with teachers and EQAO personnel expanded my understanding of current assessment policies and practices for Grade 9 mathematics. I view my contextualization of this case study less as setting the stage or creating a static backdrop for the study and more as a process of gaining a better understanding of the assessment practices that are taking place as part of the multiple-use of the EQAO Grade 9 assessment. With reference to the discussion of the concept of context in

chapter 1, the contextualization of this case study is a dynamic, emergent and ongoing process rather than an initial, a priori aspect of the research. Thus, while this information is presented before the observations from the case study are described, many of the insights and understandings included in this chapter actually co-emerged as part of the data construction process. Moreover, I present the information in this chapter with the hope that it will help readers better understand the observations about the multiple-use of the EQAO Grade 9 assessment which are presented in chapter 6.

Grade 9 Mathematics Instruction and Assessment in Ontario

Ontario's Mathematics Curricula and Courses

School districts in Ontario are required to offer mathematics courses that are consistent with the curriculum documents developed by the provincial Ministry of Education (OME, 2009). These curriculum documents provide information about mathematics instruction, guidelines for classroom assessment and specific content expectations for each mathematics course (OME, 2005).¹⁹ Mathematics teachers are required to follow these curriculum documents as they plan the instructional and assessment activities they will use in their classroom. As there are two languages of instruction in Ontario, French and English, the Ministry produces separate mathematics curriculum documents in each language. However, with the exception of the language used, the content of the French and English mathematics curriculum documents is quite similar.

In Ontario, as many as three mathematics courses may be offered in either language of instruction at the Grade 9 level: Academic, Applied and Locally Developed courses. Academic courses “develop students’ knowledge and skills through the study of theory and

¹⁹ Ontario’s curriculum documents are comparable to standards in American educational jurisdictions.

abstract problems. These courses focus on the essential concepts of a subject and explore related concepts as well. They incorporate practical applications as appropriate.” (OME, 2005, p.6). Curriculum expectations for the Grade 9 Academic mathematics course are organized in four content areas or strands: number sense and algebra, linear relations, analytic geometry, and measurement and geometry. Applied courses “focus on the essential concepts of a subject, and develop students’ knowledge and skills through practical applications and concrete examples. Familiar situations are used to illustrate ideas, and students are given more opportunities to experience hands-on applications of the concepts and theories they study” (OME, 2005, p.6). Curriculum expectations for the Grade 9 Applied mathematics course are organized in three strands: number sense and algebra, linear relations, and measurement and geometry. In addition to Academic and Applied courses, in some schools students in Grade 9 may take a locally developed mathematics course. Curriculum expectations for locally developed courses are established by individual school districts and approved by the Ministry of Education (OME, 2005, p.7). Students who take locally developed mathematics courses do not complete the EQAO Grade 9 assessment and for this reason these courses are not further discussed in this study.

The EQAO Grade 9 Assessment of Mathematics is directly linked to the Grade 9 Academic and Applied mathematics courses. According to EQAO, the Grade 9 assessment is intended to “evaluate the knowledge and skills that The Ontario Curriculum expects students to have learned by the end of Grade 9” (EQAO, 2006, p.1). For this reason, EQAO assessments are developed for the Grade 9 Academic and the Grade 9 Applied mathematics courses in each language which results in four different assessments: the French-language and English-language Academic mathematics assessments and the French-language and

English-language Applied mathematics assessments. Except for the language used, the French-language and English-language assessments are considered to be parallel to one another and include similar mathematics content. However, the Academic and Applied assessments for each language are entirely separate assessments with each based on different Ministry of Education curriculum expectations.

Ontario Assessment Policies and Practices

To better understand the way teachers use the EQAO Grade 9 assessment as part of students' grades, it is helpful to be aware of a few policies, terms and common practices related to classroom assessment in Ontario. While some aspects of classroom assessment practice are regulated by policies established by the Ministry of Education and apply to teachers in all school districts across the province, other aspects of assessment practice are under the jurisdiction of each school district. For this reason assessment procedures and policies may differ across school boards. In this section, I outline some assessment policies at the Ministry and school district levels that relate to teachers' use of the EQAO Grade 9 assessment as part of students' grades.

Term Marks and Summative Grades

At the high school level, the Ministry of Education requires that students be given a report card grade which is expressed as a percent value. Teachers are required to ensure that 70% of a student's final grade is based on work completed throughout the term and 30% of the final grade is based on work completed towards the end of the term (OME, 2005). Many secondary teachers refer to the 70% portion of the grade as the term mark and the 30% portion of the grade as the summative mark. Although Ministry curriculum documents do not refer to term and summative marks, many of the participants in my study of the EQAO

Grade 9 assessment use these expressions and, for that reason, I use these terms as I present my observations.

With regard to term and summative grades, many school districts have a policy which requires that the summative mark be based on more than one task. For example, the Ottawa-Carleton School district (OCDSB) assessment policy document specifies that for high school courses the summative mark must be based on two or more tasks, one of these tasks may be a written examination but this written examination cannot count for more than 20% of the course grade (OCDSB, 2008). In mathematics departments across Ontario, it is fairly common practice for the 30% summative mark to be based on two components: a performance task and a written exam. The allocation of marks from each of these tasks varies. For example, in some schools there is a departmental policy that the course grade includes 15% for each component while in other schools the course grade might include 20% for the performance task and 10% for the written examination. The nature of the written examination also varies from school to school. In some schools, each teacher develops their own final examination and administers it to the students in their class. This is sometimes referred to as a non-common examination. In other schools, teachers in the mathematics department work collaboratively to establish a common examination which is administered to all Grade 9 Academic students and another common examination which is administered to all Grade 9 Applied students. In describing my case study observations, I use the term in-school final examination to refer to these assessments and I specify whether the examination is common or non-common.

Development, Administration and Validation of the EQAO Grade 9 Assessment

Design and Development of the Assessment

The EQAO Grade 9 Assessment of Mathematics was first administered in January 2001 (EQAO, 2009a). The design and development process for this assessment has been documented by EQAO and a substantial quantity of information about this process is available on their website. For example, there is an explanatory paper on the development, scoring and reporting of the original Grade 9 Assessment of Mathematics (EQAO, 2001), a series of framework documents for subsequent versions of the assessment (EQAO, 2006, 2009c) and a series of technical reports detailing the psychometric properties of various versions of the assessment (EQAO, 2008, 2009d). A detailed review of each of these documents is not necessary for this dissertation. However, an overview of the item types, administration process, scoring methods and approach to reporting is provided based on the information included in these EQAO documents. These details are included because they directly relate to teachers' use of the Grade 9 assessment as part of students' grades. I begin with a description of the many ways that teachers are involved with the design and development of the EQAO Grade 9 Assessment of Mathematics.

Teachers' Involvement in Item Development, Administration and Scoring

The design and development of the Grade 9 Assessment of Mathematics is under the jurisdiction of EQAO and is conducted by psychometricians, mathematics content experts, data analysts and others. However, classroom teachers are also involved in many aspects of the development process. To begin with, as was evident in my interviews with EQAO personnel, many of the content experts at EQAO were formerly mathematics teachers in Ontario classrooms. In addition, many practising mathematics teachers are involved in

developing new assessment items each year (EQAO, 2009d). Teams of teachers participate in item writing activities and review potential items both to assess their connection to the curriculum and to ensure that items do not contain material that would be inappropriate for specific groups of students. Teams of teachers are also involved in developing the item specific rubrics which are used for the official scoring of the open-response items and in selecting examples of student responses to these items.²⁰ Many teachers are also involved in the official scoring of the EQAO Grade 9 assessment. Every summer, qualified mathematics teachers are hired on a temporary basis and trained to complete the official scoring of the open-response items (EQAO, 2009d). Another way that teachers are involved with the EQAO Grade 9 assessment is that each Grade 9 mathematics teacher is responsible for the administration of the assessment in their own classroom (EQAO, 2007). Thus, mathematics teachers are involved in many aspects of the development, administration, and scoring of this assessment.

Types of Items on the EQAO Grade 9 Assessment

Evolution of item types.

The types of items included on the EQAO Grade 9 Assessment of Mathematics have changed considerably since the assessment was first developed. In its original form, the assessment included three types of items: multiple-choice, short answer open-response and extended multi-part open-response. There was also a plan to include more complex mathematical investigations as part of the assessment. These investigations were pilot tested during the first year of administration but were dropped from the assessment the following year. The extended open-response items were dropped a few years later. Currently, the

²⁰ Examples of student responses to the open-response items are identified at each of the four levels of achievement and used as anchor papers to guide the official scoring of the open-response items on the assessment.

EQAO Grade 9 assessment includes two types of items: multiple-choice and short answer open-response. When the extended open-response items were dropped, the proportion of multiple-choice items on the assessment substantially increased. These changes in item format are described to show the evolution of item types on the assessment and because comments are made by some participants in my study with regard to the extended open-response items and the investigations which were included on earlier versions of the assessment but were not a part of the assessment when my study was conducted.

Operational and field-test items.

The EQAO Grade 9 assessment includes both operational and field-test items. Operational items are questions that count toward the student's official EQAO score whereas field-test items are embedded in the assessment as part of the item development process. Field-test items may become operational items on future tests but students' responses to items that are being field-tested are not included in their official EQAO score (EQAO, 2009c). Neither the students nor the classroom teachers are aware of which items are operational and which are being field-tested. No more than 20% of the items on any version of the assessment are field-test items (EQAO, 2009c). However, in order to field-test a sufficient number of items each year, different field-test items are included in various batches of the assessment booklets. This means that students in one classroom or in one school who are taking the assessment at the same time will be completing the same operational items but may be completing different field-test items.

Released items.

Each EQAO Grade 9 assessment typically includes three field-test and 24 operational multiple-choice items, as well as one field-test and seven operational open-response items.

The items are distributed across two test booklets with 13 or 14 multiple-choice items and four open-response items included in each booklet. Every year EQAO releases a selection of the items which were used as operational items on the previous year's assessment. These items, referred to as released items, are available on the EQAO website and are intended to be used by students, parents and teachers. Teachers may decide to use these items in any part of their mathematics program. For example, they may include released items on unit tests or quizzes or use them to help students to prepare for taking the EQAO Grade 9 assessment.

To familiarize the reader with the EQAO Grade 9 assessment items, I have included four released multiple-choice items from the Grade 9 Applied Assessment as Appendix A and one released open-response item (called "Population Plans") from the Grade 9 Academic assessment as Appendix B. I refer to the Population Plans item later in this section as I describe the process which is used for the official scoring of the open-response items.

Administration of the EQAO Grade 9 Assessment

About 100,000 students enrolled in Grade 9 Academic mathematics courses and 50,000 students enrolled in Grade 9 Applied mathematics courses complete the EQAO Grade 9 assessments each year (EQAO, 2009a). Most schools in Ontario have a semestered schedule so that students complete their entire mathematics credit in either the fall semester (Sept – Jan) or the spring semester (Feb – June). A few schools are not semestered and their mathematics courses are taught from September to June. To accommodate this, an Academic and an Applied EQAO assessment is created and administered in January for students who take their mathematics course in the first semester. A second set of assessments are developed each year to be administered in June. Students who are studying in a school that is not semestered take the June EQAO assessment along with those students who completed

their mathematics course in the second semester. Students who take Grade 9 mathematics courses during the summer are not required to complete the EQAO Grade 9 Assessment of Mathematics.

Just as the item types have changed over the years, the administration process for the EQAO Grade 9 assessment has also changed. The amount of time for testing has been reduced as the number and type of items has changed. Earlier versions of the Grade 9 assessment were administered in 70-minute sessions over 3-5 days whereas the current version is administered during two 50-minute class periods. Detailed instructions are provided to principals and classroom teachers by EQAO in a test administration guideline distributed to schools every year (i.e. EQAO, 2007). At the beginning of each school year, EQAO identifies a testing window for each semester which is the time period during which schools must administer the assessment. The testing window for the assessment is about 15 days in length. For example, for the fall semester the testing window might be from January 7 to 22. Individual schools can choose any two consecutive days within this window to administer the assessment but they must ensure that the completed assessments are returned to EQAO by the end of the testing window. Teachers who choose to mark the EQAO assessment items must ensure that this process is also completed within the testing window.

Official Scoring of the EQAO Grade 9 Assessment

The ‘official’ scoring of the EQAO Grade 9 Assessment is overseen by EQAO. Detailed information about many aspects of the scoring process can be found in the most recently released EQAO technical report (2009d). Here, I provide a summary of the procedures used in the official scoring and indicate the materials EQAO makes available for classroom teachers as they mark the assessment items to use in their students’ grades. This

information relates to many of the comments about scoring which are made by teachers who participated in this case study. The official scoring procedures for the EQAO Grade 9 assessment differ for each item type so each is discussed separately.

Official scoring of multiple-choice items.

The official scoring of multiple-choice items is done electronically by scanning students' response sheets after the assessments are returned to EQAO. Marking keys for released multiple-choice items which appeared on the assessments in previous years are available on the EQAO website but EQAO does not provide classroom teachers with a marking key for the multiple-choice items on the current years' assessment (EQAO, 2009a).

Official scoring of open-response items.

The official scoring of the open-response items takes place during the summer and is conducted by teams of qualified teachers who are hired on a temporary basis and trained for this task. EQAO provides details about the qualifications and training of these individuals as well as the scoring procedures they are required to follow in the technical report (EQAO, 2009d). In this section, I summarize this scoring process and indicate the scoring information that is available to classroom teachers who choose to score the open-response items.

EQAO has a generic scoring rubric for the open-response items which indicates the qualities which students' responses must have for each of the four scoring codes or achievement levels (see Appendix C).²¹ This generic rubric stays the same from one version of the EQAO assessment to the next so that it can be applied to any open-response item on the assessment. In addition to this generic rubric, item-specific scoring rubrics and student anchor responses which provide examples of students' work at each of the four levels of

²¹ Codes of 10, 20, 30 and 40 on the open-response items equate to levels of 1, 2, 3, and 4 on the EQAO assessment.

achievement are developed for each open-response item. To illustrate these scoring resources, I have included the item-specific scoring rubric for the Population Plans item in Appendix D and an anchor response for this item in Appendix E. The anchor response included in the appendix shows an example of a student response which was rated as a Code 30 or Level 3 but anchor responses for the other three levels of achievement are also available for this item on the EQAO website. In an effort to increase the reliability of the official scoring process, teachers who are hired to score the open-response items are trained to use the generic and item-specific scoring rubrics as well as the anchor responses when they mark students' responses.

Achievement levels of official scores.

Once the official scoring process is complete, the results are combined and reported as levels on a scale from 1 to 4. The definitions for each level are shown in Table 1. The achievement levels which are used for the EQAO Grade 9 assessment scores are the same as the achievement levels teachers are required to use in their classroom assessment practice. Thus, Table 1 is taken from the EQAO Framework document (EQAO, 2009c, p.14) but nearly identical text appears in the Ministry of Education mathematics curriculum document (see OME, 2005, p.18).

Table 1: Achievement Levels for the EQAO Grade 9 Assessment of Mathematics

Level	Description
1	• Achievement falls much below the provincial standard, while still reflecting a passing grade
2	• Achievement that approaches the provincial standard
3	• Represents the provincial standard of achievement. Parents of students achieving at Level 3 can be confident that their children will be prepared for work in the next grade
4	• Achievement that surpasses the standard

Open-response scoring information available to classroom teachers.

Teachers who decide to mark the open-response items on the EQAO assessment to include in their students' grades do not have access to all of the scoring resources used for the official scoring. The generic scoring rubric is posted on the EQAO website but the item-specific scoring rubrics and anchor responses for the current year's assessment are not available to classroom teachers at the time when they are marking the EQAO assessment items. EQAO encourages teachers to use the generic rubric as they prepare students to take the EQAO assessment (EQAO, 2009c) and teachers may also decide to use this rubric when they are marking students' responses to the open-response items. A comparison of the generic rubric with the item-specific rubric (see Appendices C & D) shows that the item-specific rubric provides minimal additional information to guide the scoring process but much more detailed information is provided in the anchor responses. These anchor responses provide examples of students' responses at each of the four levels of achievement along with a detailed rationale for why the level was assigned. The item specific information provided in the anchor responses is a key part of the official scoring process but classroom teachers do not have access to this information for the open-response items on the current year's

assessment. That is, classroom teachers have access to the generic scoring rubric but only have access to the item specific scoring rubrics and anchor responses for items from the *previous years' assessments*. This means that classroom teachers can have a general sense of how EQAO assesses open-response items but they do not have any item specific details at the time when they are marking these items to include them as part of their students' grades.

Official Score Reports for the EQAO Grade 9 Assessment

Official score reports are released in the fall of the school year immediately following the year when the assessment was administered. Individual score reports are provided for students and their parents. Score reports are also produced by EQAO for each school and district, and for the entire province. Annual school, district and province-wide reports are released to the public and can be easily accessed on EQAO's website. These reports indicate the percentage of students who achieved at each of the four levels on the assessment for both the Academic and Applied courses as well as information about students' achievement broken down by gender for each course. The percentage of students who achieve at or above the provincial standard of Level 3 is often used as a way of summarising results on the EQAO Grade 9 assessment.

Recent Trends in Students' Performance on the EQAO Grade 9 Assessment

Data showing the percentage of students across the province who achieved at or above the provincial standard of Level 3 over the past five years is shown for the Grade 9 Applied and Academic assessments in Table 2 (EQAO, 2009e).

Table 2: Percentage of Ontario Students at or above Level 3 on the EQAO Grade 9 Academic and Applied Assessments of Mathematics

	School Year				
	2004-05	2005-06	2006-07	2007-08	2008-09
EQAO Grade 9 Assessment of Mathematics – Applied	27%	35%	35%	34%	38%
EQAO Grade 9 Assessment of Mathematics – Academic	68%	71%	71%	75%	77%

One trend which can be seen in Table 2 is that the percentage of students achieving at or above Level 3, for both the Applied and Academic assessments, has tended to increase across the five years shown. Another trend evident in Table 2 is the considerable difference between the results for the Academic assessment and the Applied assessment across all five years. The percentage of students scoring at or above the provincial standard of Level 3 for the Applied assessment is consistently lower than for the Academic assessment. The difference in the percentage of students scoring at or above Level 3 between the two assessments is greater than 35% across all five years, even though separate assessments are developed for each course. Concern with regard to the consistently low number of students in Applied courses who achieve at or above Level 3 is expressed by Wolfe et al. (2004) who recommend that EQAO conduct research to investigate the appropriateness of the Grade 9 assessment for Applied courses. As will be shown, several participants in my study also express their concern with regard to students' performance on the Grade 9 Applied assessment.

Stakes Associated with the EQAO Grade 9 Assessment

As noted earlier, EQAO currently oversees four assessments which all publicly funded schools are required to administer (ie. the Grades 3, 6, 9 and 10 assessments). Each

of these assessments has stakes associated with it in the sense that the assessment results for each individual school within a district are released to the public. Schools where many students are not reaching the provincial standard of Level 3 may feel pressure to improve their performance on the assessment. Beyond these accountability-related stakes which are relevant to all EQAO assessments, different additional stakes are associated with each assessment. For the grades 3 and 6 assessments there are no direct stakes for students; successful completion of the assessment is not required for students to pass these grades and scores from the EQAO assessments are not included in students' grades. For these reasons, these EQAO assessments are often described as having low stakes. In contrast, successful completion of the Ontario Secondary Schools Literacy Test (OSSLT) which is administered in Grade 10 is a requirement for high-school graduation and, for this reason, this assessment is typically considered to be a high-stakes assessment. Successful completion of the Grade 9 Assessment of Mathematics is not required for graduation or to obtain credit for Grade 9 mathematics courses. However, students' responses to the items may be included as part of students' grades. Thus, the stakes for students which are associated with the Grade 9 assessment are less than the stakes for the OSSLT but greater than for the Grades 3 and 6 assessments. In addition, the stakes associated with the Grade 9 assessment differ from school to school depending on the weight given to the assessment when it is used as part of students' grades.

Existing Evidence for Validation of the EQAO Grade 9 Assessment

A review of existing validity evidence for the EQAO Grade 9 assessment is an important aspect of this study. As noted in chapter 3, the validation process for an assessment can be considered in two stages: the development stage and the appraisal stage

(Kane, 2006). The information provided by an assessment developer is part of the development stage and tends to have a confirmationist bias as it is gathered to support the intended use of the assessment (Haertel, 1999; Kane, 2006). This information is a necessary part of the process of validation but it is not sufficient in ensuring that the interpretations made on the basis of the assessment can be warranted. Cronbach observes “it is not at all uncommon ... for the test developer to claim validity by construction, bolstering the claim with a detailed account of the construction process” (1971, p.456). For this reason, sources of evidence gathered by individuals and groups outside the assessment development process after the assessment is in use must also be included in the validation process and this is the evidence which is collected in the appraisal stage of validation.

In this section, I consider existing validity evidence for the EQAO Grade 9 assessment from both the development and the appraisal stages. I summarize the evidence from each stage to indicate the kinds of evidence which have been gathered for validating the inferences from the EQAO Grade 9 assessment but I do not provide a critical analysis of the findings from each validity-related study. Conducting a critical analysis of these findings is a significant undertaking which would be an essential part of a comprehensive validation for this assessment but is not necessary given the focus of this dissertation.

Validation Evidence Provided by EQAO

The information EQAO provides about the design and development of the Grade 9 assessment constitutes one important source of validation evidence. The framework documents provide detailed information about test design, item development, field testing, and equating (EQAO 2006, 2009c). Successive versions of these framework documents have been released by EQAO and demonstrate how the assessment has changed since it was first

administered in 2001. These documents also include an assessment blueprint which maps each question on the assessment to expectations in the mathematics curriculum. This information provides evidence for some of the inferences which are made from this assessment particularly with regard to the aspects of mathematics that the assessment claims to measure. Additional information is included in the technical report (EQAO, 2009d) which provides some evidence of the extent to which the scoring procedures for this assessment can be considered to contribute to valid inferences about students' understanding of the Grade 9 mathematics curriculum. The framework documents and technical reports along with other documentation provided on the EQAO website indicate that a considerable quantity of evidence has been collected in the development stage of validation for the EQAO Grade 9 assessment. To get a sense of the evidence which has been collected in the appraisal stage of validation for this assessment I consider studies commissioned by EQAO as well as validity-related studies which have been conducted by independent researchers.

Studies Commissioned by EQAO

In 2002, EQAO began a comprehensive review project which was intended to encompass all aspects of their assessment program. As part of this project, EQAO commissioned a number of research reports and consultation papers. Some of these studies relate to the entire EQAO assessment program while others focus on the EQAO Grade 9 Assessment of Mathematics. To date, the most wide-ranging study commissioned by EQAO was an external evaluation of the assessment processes in all four EQAO assessment programs (Wolfe et al., 2004). This external review was conducted over a period of ten months by a team of researchers at the Ontario Institute for Studies in Education at the University of Toronto. EQAO provided these researchers with access to EQAO staff,

assessment materials and documentation, as well as electronic files of assessment data (Wolfe et al., 2004, p.2). In addition to conducting their own analyses, these researchers consulted with a number of assessment experts in North America as they considered various aspects of the EQAO assessment process. The final report which emerged from this review discusses the purposes, test design, item development, scoring, scaling and reporting associated with each of the EQAO assessments (Wolfe et al., 2004). The authors identify the strengths of each assessment and point out areas where the assessments could be enhanced. They make eight overall recommendations for improving EQAO's assessment practices. Each recommendation includes a statement that applies to the entire assessment program as well as suggestions that are specific to individual assessments. One overall recommendation these authors make is that EQAO initiate and carry out an active program of validity research (Wolfe et al., 2004, p.67). Thus, a key finding of this external review of the EQAO assessment program is that more effort toward validation is needed.

Wolfe et al. (2004) make a number of observations and recommendations that refer specifically to the Grade 9 Assessment of Mathematics. Many of their observations and recommendations have serious implications for the validation of inferences made from this assessment. While an extensive discussion of each recommendation is beyond the scope of this discussion, I mention two recommendations which relate to teachers' use of the EQAO Grade 9 assessment items as part of students' grades. One recommendation is that EQAO clarify the purposes that the Grade 9 assessment is intended to serve and the other recommendation relates to changing the design of the Grade 9 assessment to increase the validity of inferences which are made. I return to each of these recommendations and discuss

them more fully in the final section of this chapter when I summarize existing research into teachers' use of the EQAO Grade 9 assessment as part of students' grades.

Other studies which were commissioned by EQAO as part of this comprehensive assessment review project focus specifically on the Grade 9 Assessment of Mathematics.

Eleven studies were located with this focus and are listed in Table 3.

Table 3: Studies of the EQAO Grade 9 Assessment of Mathematics Commissioned by EQAO

Author(s) & Year	Title of the Study
Childs, Falenchuk & Dunn (2004)	<i>Score reliability for the Grade 9 assessment</i>
Dunn, Childs & Xu (2004)	<i>Incomplete student work and scoring approaches for the Grade 9 Assessment of Mathematics</i>
Emenogu (2003)	<i>Teachers' instructional practices, students' opportunity to learn and differentiated item functioning the Grade 9 Assessment of Mathematics</i>
Emenogu, Pang & Falenchuk (2004)	<i>Incomplete assessments and the Grade 9 Assessment of Mathematics</i>
Falenchuk, Childs & Dunn (2003)	<i>Scaling of the Grade 9 assessment: Guidelines for calibration</i>
Falenchuk, Childs & Dunn (2004)	<i>An evaluation of the alignment approach for the Grade 9 Assessment of Mathematics</i>
Herbert, Dunn & Luthra (2004)	<i>The validation of psychometric procedures: Teachers perceptions and opinions of the Grade 9 Assessment of Mathematics</i>
Klinger, Falenchuk & Pang (2004)	<i>Simultaneous versus separate calibration of the Grade 9 Assessment of Mathematics</i>
Suurtamm & Moisan (2004)	<i>Examining the multiple versions of the Grade 9 Assessment of Mathematics</i>
Suurtamm, Moisan & Luthra (2004a)	<i>Does the mathematics being measured by the Grade 9 Assessment of Mathematics represent the "important" mathematics to be learned in the Grade 9 curriculum?</i>
Suurtamm, Moisan & Luthra (2004b)	<i>How well does the Grade 9 Assessment of Mathematics reflect the Ontario curriculum?</i>

As noted earlier, a critical analysis of each of these studies is not included in this dissertation. Rather, the titles of these studies are provided as an indication of the kinds of validity evidence EQAO has collected for the Grade 9 Assessment of Mathematics. A review of these titles suggests that the focus is primarily on the psychometric characteristics (e.g., Childs, Falenchuk & Dunn, 2004; Dunn, Childs & Xu, 2004; Klinger, Falenchuk & Pang, 2004 etc.) and content of the assessment (e.g. Suurtamm, Moisan & Luthra 2004a, 2004b). While a detailed review of these psychometric characteristics is not necessary for this summary of the validity evidence for the EQAO Grade 9 assessment, in general these studies provide evidence that the assessment is a reliable measure of individual achievement and assesses a reasonable sample of the Grade 9 mathematics curriculum. Further, as noted below, EQAO has responded to some of the specific concerns raised in these studies by making changes to the assessment or by providing additional explanatory text concerning the limitations of the assessment.

The studies which focus on the content of the EQAO Grade 9 assessment merit a little further discussion because their findings relate more directly to teachers' use of the EQAO assessment items as part of students' grades. In these studies, Suurtamm, Moisan & Luthra explore the connection between items on the EQAO Grade 9 assessment and the Ontario mathematics curriculum. The authors attempt to match items on the assessment with the individual curriculum expectations. Their findings suggest that matching items from the assessment to the curriculum is problematic in that some items can be matched to more than one curriculum expectation. Further, the authors conclude that while many of the EQAO assessment items match the curriculum expectations reasonably well, this alignment does not guarantee that the important mathematics concepts in the curriculum are being assessed by

the items on the EQAO assessment. As will be seen in chapter 6, the teachers who participated in this study also express concerns with regard to the content of the EQAO Grade 9 assessment items.

I located two additional validity-related studies of the EQAO Grade 9 assessment which were funded by EQAO but were not part of the assessment review project described above (Ross & Gray, 2005; Suurtamm, Lawson & Koch, 2008). A brief summary of each study is provided as these studies relate to teachers' use of the assessment as part of students' grades.

Ross and Gray (2005) compared scores from the EQAO Grade 9 assessment with report card grades using the 2002 (n=2450) and 2003 (n=2404) scores for one school district. Students' official EQAO results were compared with their nearest preceding report card grade. Given this design, Ross and Gray were treating students' report card grades as a predictor of their EQAO scores. No mention is made of teachers' use of the assessment as part of students' grades. However, the study appears to have used mid-semester report card grades for the Grade 9 students. Presumably, these mid-semester grades were collected prior to the administration of the EQAO Grade 9 assessment and do not include grades that were based in part on teachers' scoring of items on the EQAO assessment. Ross and Gray report a significant difference between report card grades and EQAO scores for both Academic and Applied mathematics courses with the mean grade on the report card being significantly lower than the mean score on the EQAO. The agreement between report cards and EQAO scores was higher for Academic courses than for Applied courses and Ross and Gray interpret this finding as suggesting that the EQAO assessment may be better aligned with classroom assessment practices for the Academic course than it is for the Applied course.

This finding is interesting in light of the persistent difference between the number of students achieving a Level 3 on the EQAO Academic and Applied assessments described earlier in this chapter (see Table 2). In a general sense, this study provides some criterion-related evidence for validating inferences made from the EQAO Grade 9 assessment and suggests that there may be some concerns particularly for the EQAO Grade 9 Applied assessment.

Suurtamm, Lawson, and Koch (2008) use document analysis, information provided by EQAO and their own mapping of items on the assessment to the curriculum expectations to consider the coherence of the EQAO Grade 9 assessment with the curriculum. In this study the authors demonstrate that the EQAO Grade 9 assessment overlooks the investigative nature of the Ontario mathematics curriculum. It is interesting to note that since this research was conducted EQAO has added an explicit statement to the framework document indicating that those curriculum expectations which require mathematical investigations are not evaluated on the assessment and are better assessed by classroom teachers (EQAO, 2009c).

I have summarized the studies commissioned by EQAO that relate to the validation of Grade 9 assessment in this section as they constitute part of the appraisal stage of validation. As noted, it is important to consider studies conducted by independent researchers as well and I turn to this group of studies in the next section.

Studies of the EQAO Grade 9 Assessment Conducted by Independent Researchers

Although I searched for studies which were done by independent researchers that could contribute to the validation of inferences made from the EQAO Grade 9 assessment, this search produced very few results. I was able to locate two doctoral dissertations (Kitto,

2006; Lock 2001) which provide some insights for the validation of this assessment. Each of these dissertations relates directly to teachers' use of the assessment as part of students' grades and a description of both studies is included later in this chapter where I discuss existing research on the use of the EQAO Grade 9 assessment by teachers as part of students' grades. Beyond these two dissertations, I was not able to locate any published articles which provide evidence that could contribute to the validation of the EQAO Grade 9 assessment. This was somewhat surprising as I was able to locate a number of studies which relate to the validation of other EQAO assessments. For example, I was able to locate a number of recently published studies that would be relevant to the process of validation for the EQAO Grade 10 literacy assessment (e.g. Luce-Kapler, 2005; Ricci, 2004).

A Summary of Existing Validity Evidence for the EQAO Grade 9 Assessment

My search for validity evidence for the EQAO Grade 9 assessment suggests that most of the evidence which is currently available has been provided by EQAO in the development stage of validation or commissioned by EQAO as an initial part of the appraisal stage of validation. The studies located suggest that EQAO has focused their validation efforts principally on the psychometric properties of the assessment and on the alignment of the assessment with the curriculum. As noted earlier, Wolfe et al. (2004) call for a more extensive validity program for this assessment. Similarly, Crundwell (2005) and Volante (2007) lament the lack of validity evidence for the EQAO assessments and note that while these assessments may have adequate evidence of content validity there is a serious lack of evidence regarding other aspects of validity such as the consequences of the assessment. A number of researchers have noted that the validation process for many large-scale assessments tends to focus on the relationship between the content of items and the subject

area and/or on the selection of items with adequate psychometric properties (e.g, Kane, 2008; Sicolý, 2002; Sireci, 2007). These researchers maintain that this emphasis on content and psychometrics is not an adequate approach to validation. This observation certainly seems to pertain to the validation of the EQAO Grade 9 assessment and I add my voice to the call for more validity-related research for this assessment. Further, I suggest that there is a particular need for research conducted by researchers who are independent of EQAO.

In this section I have summarized the validity evidence which is currently available for the EQAO Grade 9 assessment. I now turn to a discussion of the research which has been conducted related specifically to the use of the EQAO Grade 9 assessment by teachers as part of students' grades. I begin by describing teachers' use of large-scale assessments as part of students' grades across Canada and then present some studies which discuss this practice with reference to the EQAO Grade 9 assessment.

Research on Using the EQAO Assessment in Students' Grades

Teachers' use of large-scale assessments as part of students' grades is not uncommon across Canada. Two studies which demonstrate the frequency of this practice in Canada are summarized in this section along with four studies which specifically consider teachers' use of the EQAO Grade 9 assessment as part of students' grades.

Evidence of the frequent use of large-scale assessments as part of students' grades in Canada is found in information gathered in the Pan-Canadian Assessment Program (PCAP). PCAP assesses the achievement of 13-year-old students in reading, mathematics and science in each province in Canada (Council of Ministers of Education Canada [CMEC], 2009). As part of this assessment program, PCAP gathers background information by administering questionnaires to students, teachers and school administrators. To get a sense of the

assessment practices that may influence student's achievement on the PCAP, one item on the questionnaire asks teachers about their use of province-wide assessments as part of students' grades. Results from this item are summarized in Table 4 and indicate that teachers in every province or territory include the results from their province-wide assessments as part of their students' grades (CMEC, 2009). However, the percentage of teachers who report this practice varies considerably across provinces. The practice is most common in Saskatchewan where half of the teachers indicate that they include province-wide assessment results in their students' grades and is least common in Prince Edward Island where only 5% of teachers report using province-wide assessments as part of their students' grades. In Ontario, 27% of teachers who completed the PCAP questionnaire indicate that they use province-wide assessments as part of their students' grades. The PCAP questionnaire was completed by teachers of students in Grades 7, 8 and 9 because the assessment is age-based (i.e. 13-year olds) rather than grade-based (CMEC, 2009, p.14). Accordingly, the percentage reported for Ontario includes teachers who use the EQAO Grade 9 assessment.²² The information from this PCAP questionnaire demonstrates that teachers' use of large-scale provincial assessments such as the EQAO Grade 9 assessment as part of students' grades is a fairly common practice across Canada. However, no information about how this practice takes place is provided. For example, it is not clear if teachers in other provinces are marking the assessment prior to returning them for official scoring as is done with the EQAO Grade 9 assessment or if they receive official scores from the province-wide assessments in time to include them in their students' grades.

²² In fact, it must include these individuals as teachers are not permitted to mark the EQAO Grade 6 assessment to include it as part of students' grades and the official score reports are not returned to schools until after report card grades have been issued.

Table 4: Percentage of Teachers who Report Using Province-wide Assessments as Part of Students' Grades by Province or Territory

Province/Territory	% of Teachers
Saskatchewan	50%
Alberta	48%
Yukon	47%
Newfoundland	45%
New Brunswick	41%
Quebec	35%
Manitoba	32%
Ontario	27%
British Columbia	25%
Nova Scotia	24%
Prince Edward Island	5%

Another reference to the use of large-scale assessments as part of students' grades is made by Klinger et al. (2008) who document large-scale assessment practices in each province in Canada. With regard to teachers' use of large-scale assessments as part of students' grades, Klinger et al. (2008) indicate that in Alberta, teachers are encouraged but not required to mark their students' large-scale assessment responses and to use the mark as part of their students' grades. They indicate that this approach to encouraging teachers' use of the assessment items applies to all of the province-wide assessments administered in Alberta and is not limited to the mathematics assessment. In Saskatchewan, teachers are also encouraged to score student work on provincial assessments in reading, mathematics and writing and to share feedback with students before they return the assessments for central scoring. However, no mention of including the assessment results as part of students' grades is made. In Prince Edward Island, teachers are required to include their province-wide Grade

9 assessment of mathematics as 10% of student's grades.²³ In the Yukon scores on their province-wide Grade 9 assessments of both mathematics and language arts are counted as 25% of students' course grades. With regard to the EQAO Grade 9 assessment in Ontario, these authors observe "teachers commonly include a portion of the provincial test in calculating students' final grades, to a maximum of 10%" (p.17). The source of this information is not indicated and no reference for the value of 10% is provided. My research indicates that no maximum value is indicated in any of the EQAO documents or in the assessment guidelines provided by the Ministry of Education. Nonetheless, Klinger et al. (2008) provide evidence that the use of large-scale assessments as part of students' grades takes place in many Canadian educational jurisdictions.

In addition to these studies which provide information about teachers' use of large-scale assessments as part of students' grades across Canada, I located four studies which specifically discuss teachers' use of the EQAO Grade 9 Assessment of Mathematics. Each study is described in this section.

The use of the EQAO Grade 9 Assessment as part of students' grades is considered by Lock (2001) in her dissertation examining the influence of this assessment on the classroom practices of Grade 9 mathematics teachers. Each of the four teachers that Lock interviewed in her study indicated that they use items from the Grade 9 assessment as part of their students' grades. These teachers reported marking different kinds of items from the assessment with some including only multiple-choice items and others using a variety of the item types that were included on the assessment that year. In addition, the EQAO assessment

²³ The % of teachers indicated for Prince Edward Island in Table 4 seems quite low given that teachers are required to include their province-wide Grade 9 assessment of mathematics as 10% of students' grades. However, the PCAP survey was conducted in the school year prior to the year that the policy requiring teachers to include the assessment in grades was introduced.

contributed differing amounts toward students' final grades; 10% for some teachers and 15% for others (teachers were located at different schools and some were from different school boards). Lock does not mention whether board or departmental policy influenced the teachers' decisions to use the assessment items as part of their students' grades but she does indicate that two of the teachers decided to use EQAO items as part of students' grades because they thought it would ensure that students took the assessment seriously. Lock provides further evidence of teachers' concerns with students taking the assessment seriously in the responses teachers gave to a questionnaire she developed. On this questionnaire 37.6% of teachers agreed with the statement "I sense that my students did not take the EQAO assessment seriously" (p. 225). Lock's research was conducted at the time of the first administration of the EQAO Grade 9 assessment and provides evidence that the use of this assessment as part of students' grades was occurring in some schools from the beginning of the assessment program.

Additional information about the use of the EQAO assessment as part of students' grades is provided by Kitto (2006) in his dissertation examining teachers' perceptions of the validity of the EQAO Grade 9 assessment. As part of his research, Kitto interviewed teachers who were teaching Grade 9 mathematics courses to gather information about their views and experiences with the EQAO Grade 9 assessment. Kitto reports that seven of the ten teachers he interviewed had decided to use the EQAO assessment as part of their students' grades. As observed by Lock (2001), the amount that the assessment contributed to students' grades varied among these teachers. Kitto reports that the most common reason teachers gave for using the assessment as part of students' grades was that their mathematics department required them to do so. The second most common reason for using the

assessment as part of students' grades reported by these teachers was to encourage students to take the assessment more seriously. Most of the teachers Kitto interviewed indicated that they marked only the multiple-choice items to include in students' grades. Kitto also notes that the teachers in his study felt that the negative consequences of the EQAO assessment were particularly strong for students taking the Grade 9 Applied mathematics course. Further, Kitto's interview findings suggest that teachers who are more familiar with how the EQAO assessments are generated and marked are more sympathetic to the assessment. He states "it appears that having more knowledge and understanding of the process leads to more acceptance" (p.253).

In addition to these teacher interviews, Kitto also conducted a survey. One question on his survey was "What is the contribution of the EQAO assessment to your final report card mark?" (p.344). A total of 63 teachers responded to this question. The responses Kitto received are summarized in Table 5.

Table 5: Teachers' Responses to *What is the contribution of the EQAO assessment to your final report card mark?*

Response Option	% of Teachers
0%	16%
<10%	48%
10-20%	29%
21-30%	8%
>30%	0%

According to this questionnaire item, 85% of teachers who responded to this questionnaire use the EQAO Grade 9 assessment as part of their students' grades and approximately 50% of teachers indicate that the assessment contributes less than 10% toward that grade.

In the external review Wolfe et al. (2004) were commissioned by EQAO to conduct (described earlier in this chapter), there is no mention of the fact that teachers are permitted to mark the Grade 9 assessment and use their teacher-derived scores as part of students' grades. However, while these authors do not explicitly refer to this practice, the recommendations for improvements to the EQAO Grade 9 assessment which Wolfe et al. make include changes in the design of the assessment which relate directly to teachers' use of the assessment as part of students' grades. In particular, Wolfe et al. recommend changing the design of the Grade 9 assessment so that some of the EQAO items would remain in the school to be marked by teachers while other items would be returned to EQAO for official scoring. According to the authors, this two-part design would serve the purpose of measuring individual student achievement in mathematics as well as providing information at the school, board, and provincial levels about trends in mathematics achievement. The authors suggest that those items which are marked by teachers might replace the in-school mathematics exam portion of the summative grade and that this would save teachers time because the items and marking guides would be provided by EQAO rather than developed in individual schools (Wolfe et al., 2004). If EQAO adopted this design, the items that teachers would mark and use for students' grades would not be the same items as those that are officially scored by EQAO. Again, Wolfe et al. make this design recommendation and discuss using EQAO items as a replacement for in-school exams without any explicit reference to the current way teachers select, mark and use the EQAO Grade 9 assessment items as part of their students' grades.

One other study I located that considers teachers' use of the EQAO Grade 9 assessment as part of students' grades is an internal study which was conducted by EQAO

but not published. I located this study as a result of my email inquiries with a director at EQAO (personal communication, October 15, 2008) who indicated that this was the only research that EQAO has conducted with regard to teachers' use of the assessment as part of students' grades.²⁴ This study attempted to determine if informing students that part of the assessment would be used for their grades had an effect on students' performance on the assessment (Pang, 2004). EQAO collected data for this study by including a question on the class tracking sheet which is completed by teachers as they administer the assessment. Grade 9 teachers from the January testing session of the 2003-04 school year were asked to indicate if they had informed students that the assessment would be used as part of their grades. Separate data were collected for Grade 9 Academic and Applied courses. The report is quite brief and does not provide a comprehensive description of the results but does indicate that 88% of teachers for the Grade 9 English-language Applied mathematics assessment (n=12,341) responded 'yes' indicating that students had been informed that the assessment would be used as part of their grades. Data at this level of specificity are not indicated for the Grade 9 English-language Academic mathematics courses. However, as noted in the report, the way the question was stated included two implicit questions (i.e. Is the assessment used as part of student's grades? and Are students informed that the assessment will be used as part of their grades?) but teachers were asked to provide only one response. Thus, it is not clear if a 'no' response indicated was that the assessment was not used or that teachers had not told the students that it would be used. A 'yes' response is somewhat clearer as it suggests that students were told that the assessment would be used as part of their grades and implies that the assessment was being used for this purpose. Of course, it is possible that

²⁴ The director more recently informed me that EQAO is currently working with a research team to develop questionnaire items for teachers and students related to this practice. However, the research is in the pilot test stage and no results are available.

teachers informed students that the assessment would be used as part of their grades but subsequently decided not to use the assessment for that purpose. Despite the problem with the design of this study, Pang (2004) concludes that being told that the assessment would be included in students' grades did influence the performance of students in Applied courses in that those who were told that the assessment would be used attempted to respond to more items and scored higher on their overall performance than those who either were not told that the assessment was being used or where the assessment was not used as part of students' grades.

CHAPTER 5

METHODOLOGY, RESEARCH DESIGN AND DATA CONSTRUCTION

In this chapter I describe my methodology, present the research questions, summarize the research design and describe the methods I used to conduct my in-depth case study of the multiple-use of the EQAO Grade 9 Assessment of Mathematics.

Methodology of the Inquiry

The methodology of this inquiry is based on a sociocultural view of assessment, ideas concerning the re-conceptualization of validity theory and the interpretivist epistemological stance that I adopt. In previous chapters, I indicated some aspects of sociocultural studies that influence my thinking and I discussed the ideas about re-conceptualizing validity theory which I have drawn on for this inquiry. In this section, I begin by summarizing my understanding of interpretivist inquiry and then I illustrate how I combined this view of research with a sociocultural view of assessment and with ideas about re-conceptualizing validity theory to establish the parameters for this inquiry.

My Understanding of Interpretivist Inquiry

Interpretivism is a term that has been used to encompass a broad range of theoretical frameworks such as critical theory, poststructuralism, feminist standpoint theory, pragmatism, hermeneutics and postmodernism (Howe, 2001; Lincoln & Guba, 2000; Moss, 2005). While many disparate views are found among these frameworks, some areas of consensus can also be identified (Erickson, 1986; Howe, 2001, Lincoln & Guba, 2000). I draw on these areas of consensus to describe some aspects of interpretivism that underpin my methodology. I also discuss some ways of enhancing the quality of interpretivist research that I use in this study.

View of Knowledge in Interpretivism

For interpretivists, knowledge is actively constructed and co-created as individuals seek to understand or make meaning from their experiences (Erickson, 1986; Howe 2001; Taylor, 1979). The active construction of knowledge is continually enabled and constrained by the various human activities which are taking place (Erickson, 1986; Rabinow & Sullivan, 1979). In interpretivist inquiry, researchers consider aspects of the complex web of social, political and economic activities surrounding a phenomenon to be an integral part of what is being studied (Howe, 2001; Rabinow & Sullivan, 1979). In educational research, these activities take place at many levels beginning with an individual student and expanding to the classroom, the school, the district and beyond. Further, because knowledge is actively constructed, interpretivist researchers acknowledge that multiple perspectives exist (Lincoln, 1995; Rabinow & Sullivan, 1979). Various individuals and groups may have different points of view regarding an issue and these diverse perspectives need to be considered in the research process. However, in any one inquiry, both the range of activities and the number of points of view that can be explored is limited. Accordingly, researchers identify the perspectives that are included in the scope of their inquiry at the same time as they acknowledge the implications of those not considered (Lincoln & Guba, 2000, Moss et al., 2006). Howe characterizes the interpretivist view of knowledge concisely when he states that knowledge “must be seen as actively constructed - culturally and historically grounded, as laden with moral and political values, and as serving certain interests and purposes” (2001, p.202).

The Purpose of Research

Interpretivist inquiry does not seek universal principles or laws to predict or explain human behavior (Rabinow & Sullivan, 1979, Taylor, 1979). Rather, the purpose of interpretivist inquiry is to make coherent meaning of human action; to find coherence between the actions of an individual or group and the meaning of those actions for that individual or group (Taylor, 1979). However, as Taylor explains, the process of making coherent meaning need not result in a logical explanation of an action or event. The meaning of an action for an individual or group may be illogical and full of confusion and contradictions. In interpretive inquiry, the process of making coherent meaning includes the adequate depiction of the confusion and contradictions of individuals and groups.

Given this view, in interpretivist inquiry the research process often begins with describing the concrete details of an event including an account of what took place, a description of what the event means to those who are involved as well as a description of the context that shaped and was shaped by the event (Erickson, 1986; Howe, 2001; Moss et al., 2006). In this way, interpretivist research can contribute to an enhanced understanding of the events that take place in an educational setting and can reveal interconnections between theory and practice. The understandings which emerge from interpretivist inquiry can also enable and promote social justice, discourse and diversity and many interpretivists indicate that they engage in educational research with these goals in mind (e.g. Greene, 1994; Howe, 2001; Lincoln, 1995).

These ideas have influenced the way I think about the purpose of research. In my research I am attempting to gain a better understanding of a phenomenon, in this case the

multiple-use of large-scale assessments, to share that understanding with others and, where possible, to use that enriched understanding to promote more equitable educational practice.

Methodology and Method in Interpretivist Research

According to many interpretivist researchers, *methodology* and *method* are related but distinct concepts (Howe, 2001; Lincoln & Guba, 2000). Howe (2001, 2004) maintains that in interpretivist inquiry, the methodology of an inquiry is established by considering the inferences that are being made from the data. These inferences are not dictated by the research methods that are used but emerge from the theoretical perspectives guiding the research. For instance, Howe (2004) contends that experimental and quasi-experimental methods can be used without adopting a positivist stance. In a similar vein, a number of other researchers describe studies where qualitative methods are used in a positivist manner and quantitative methods are used in an interpretivist manner (e.g. Ercikan & Roth, 2006; Howe, 2004; Mason, 2006; Moran-Ellis et al., 2006). This distinction between methods and methodology means that a wide range of research methods may be used in interpretivist research. In addition, many researchers use multiple sources of evidence to get at the various perspectives which are valued in interpretivist inquiry. Methods such as case studies, interviews, focus groups and participant observation are particularly valuable in interpretivist research because they enable researchers to develop an in-depth understanding of participants' views about a particular phenomenon as well as an appreciation of the complex activities which may be influencing participants (Erickson, 1986; Howe, 2001; Moss et al., 2006). For these reasons, case studies and interviews are important research methods in this inquiry.

Ensuring Quality in Interpretivist Research

Interpretivism has been described as *postepistemic* and *antifoundational* because it is an approach which rejects the adoption of fixed standards or definitive frames of reference by which truth can be universally known (Greene, 1994; Lather 1993, 2001; Lincoln, 1995; Lincoln & Guba, 2000). However, this rejection of fixed standards should not be construed as suggesting that there is no need for or means of ensuring quality in interpretivist research. As Moss points out, acknowledging the existence of multiple interpretations in no way implies that these different interpretations cannot be comparatively evaluated (1998, p.58). Numerous criteria for evaluating the quality of interpretive research have been suggested (Lincoln & Guba, 2000). The approaches to ensuring and evaluating quality in interpretivist inquiry which have most influenced my thinking are those expressed by Erickson (1986), Lather (1993; 2001), Lincoln (1995), Mishler (1990), Moss (1998, 2005), Schwandt (1996) and Wolcott (1990). Some of these authors use the term validity when they discuss the evaluation of interpretivist inquiry while others suggest a variety of terms such as transparency, quality and trustworthiness. Regardless of the terminology, in this inquiry I have used some approaches suggested by these authors to enhance the quality of this research. One recommendation these authors make is that interpretivist researchers should attempt to be aware of the lens they bring to the research process; that is, to be aware of their own interpretive activity. I have attempted to do this throughout this dissertation by describing the frameworks I draw on and by chronicling the changes in my perspective as the project unfolded. Another means of ensuring quality in interpretivist inquiry suggested by these methodologists is that researchers provide a sufficiently detailed account of the processes they have used so that the reader can judge the strength of evidence supporting any

assertions or claims the researcher makes (Erickson, 1986; Lather, 2001; Moss, 2005). While judging how much detail to report remains a challenge, this recommendation has guided each phase of my inquiry and in writing this dissertation I have attempted to provide a sufficiently detailed account of the methods I used to enable readers to evaluate the evidence I present.

Combining an Interpretivist Epistemology, a Sociocultural View of Assessment and Notions of the Re-conceptualization of Validity Theory

The methodology of this inquiry is based on my pairing of an interpretivist epistemological stance with a sociocultural view of assessment and with ideas about re-conceptualizing validity theory. In this section I describe three aspects of my methodology that illustrate how these perspectives have influenced this project: approaching the problem of multiple-use from multiple perspectives, developing a research design that facilitates contextualization, and conducting an in-depth case study of a local practice.

Approaching the Problem of Multiple-use from Multiple Perspectives

Reading about interpretivist approaches to research and sociocultural views of assessment increased my awareness of the value of considering multiple perspectives when attempting to understand the implications of the multiple-use of large-scale assessments for the process of validation. I concur with Cobb (1994) who notes that rather than attempting “to subjugate research to a single, overarching theoretical scheme that is posited a priori”, researchers can enrich inquiry by investigating a problem from different perspectives (p. 19). Cobb states “in doing so, we would give up the quest for an acontextual, one-size-fits-all perspective. Instead, we would acknowledge that we, like teachers, cast around for ways of making sense of things as we address the situated problems of our practice” (p.19). A similar view is expressed by Brookhart (2004) who argues that drawing on several theoretical

traditions in a study can enrich the research process even if methodological tensions sometimes result.

The idea of approaching the practice of multiple-use from multiple perspectives influenced this inquiry on two levels. On an overall level, I explore how validation might proceed where multiple-use is known to occur by drawing on two different theoretical perspectives. I begin with how this practice might be viewed within a current measurement-based conception of validation but I also consider how this practice may be viewed from another perspective. That is, I consider how ideas from sociocultural studies can contribute to the process of validation given multiple-use practices. Therefore, on the broadest level of this inquiry, I am thinking of the problem of multiple-use from multiple perspectives. This approach is also consistent with the value of drawing on other disciplines to expand current conceptions of validity theory, as discussed in chapter 3.

The value of considering multiple perspectives influences the methodology of this inquiry on another level and that is that I am interested in exploring the perspectives of various individuals and groups with regard to the multiple-use of the EQAO Grade 9 assessment. In this inquiry, I consider the perspectives of EQAO assessment developers as well as those that are expressed by classroom teachers, mathematics department heads, and principals from different schools who are using the assessment as part of their students' grades. In future inquiries, I would be interested in getting an understanding of the views of classroom teachers who have chosen not to use the assessment as part of their students' grades, as well as the views of students taking the EQAO Grade 9 assessment, their parents, administrators in various school districts and policymakers at the Ministry of Education.

A Research Design that Facilitates Contextualization

Given the interpretivist and sociocultural influences on my thinking, as I approached this inquiry I wanted to use research methods that enable the process of contextualization. As described in chapter 1, I use the term contextualization to indicate a view of context as something that is relational and emergent. In this inquiry then, activities which take place as part of the practice of multiple-use both shape and are shaped by the context where they take place. Accordingly, rather than focusing on statistical evidence and looking for correlations between students' official EQAO scores and whether or not the assessment is included in their grades, I designed a study that uses school-level case studies and interviews with EQAO personnel to gain an understanding of the multiple-use of the EQAO Grade 9 Assessment of Mathematics. Conducting school-level case studies is a research method that facilitates the contextualization of the practice of multiple-use in that as a researcher I can gather information about how the practice of multiple-use shapes and is shaped by the context where it takes place including some understanding of the activities of individual teachers, mathematics department heads, and principals. Similarly, understanding how this practice is viewed by EQAO requires document analysis as well as interviews to appreciate how and why this practice has evolved. I also wanted to get a sense of how the practices described by teachers in the school-level case studies relate to and are influenced by policies and practices in school districts and across the province. Thus, these are some levels where contextualization takes place in this inquiry.

Studying Local Practice to Inform Theory

Erickson (1986) states that interpretivist researchers study a situation “to uncover the different layers of universality and particularity that are confronted in the specific case at

hand – what is broadly universal, what generalizes to other similar situations, what is unique to the given instance” (Erickson, 1986, p.130). Similarly, Moss states that “from the perspective of interpretive social science, the usefulness of a case does not depend on its typicality or even on the success of the practices it represents. Rather, it depends on the richness of the detail which allows the reader to draw instructive comparisons with their own contexts of work” (2003, p.13). These statements suggest the valuable insights which can come from studying local practice and they are consistent with the views of validity theorists who see locally situated studies as an effective means of expanding on existing conceptual frameworks for validity theory (as described in chapter 3). Case studies are an appropriate means of studying local practice because they provide an opportunity to gather detailed information about a practice (Stake 1995, 2000, 2008; Stevenson, 2004; Yin, 2003, 2006). Thus, my investigation of the multiple-use of the EQAO Grade 9 assessment is intended to be an in-depth case study of a local practice that will help to better understand that practice as well as to build on current conceptual frameworks for validation.

In describing my methodology I have summarized how I joined my interpretivist epistemological stance and my sociocultural view of assessment with current notions of re-conceptualizing validity theory to frame my study of the multiple-use of large-scale assessments. For me, these theoretical frameworks suggest the value of approaching the question from multiple perspectives, of using a research design that enables contextualization of this practice and of using a case study approach. I now describe the research methods I used within this methodology by discussing the research questions and design and detailing the process through which the data was constructed.

Research Questions and Design

Some aspects of an interpretivist methodology are established a priori while others evolve as an inquiry proceeds (Erickson, 1986; Ercikan & Roth, 2006). Aspects such as the research questions and the selection of analytic techniques often evolve (Davis & Sumara, 2006; Erickson, 1986). In this section I provide an account of the research questions that guided this inquiry and a discussion of how the research design addresses each question.

Emerging Research Questions

Erickson (1986) makes a case for articulating clear research questions when conducting interpretivist research. He states “framing research questions explicitly and seeking relevant data deliberately enable and empower intuition, rather than stifle it” (p.140). Erickson also notes that research questions often change as the research progresses. Similarly, Davis and Sumara (2006) describe research questions as “moving targets” (p.149). My experience constructing research questions throughout this inquiry corroborates these views. Given that the purpose of this inquiry is to understand the implications of the multiple-use of large-scale assessments for the process of validation, I developed six questions to guide the research:

1. What is the pattern of uses associated with the EQAO Grade 9 Assessment of Mathematics?
2. How does teachers’ use of the EQAO Grade 9 Assessment of Mathematics as part of students’ grades take place?
3. How is teachers’ use of the EQAO Grade 9 Assessment of Mathematics as part of students’ grades viewed by EQAO?
4. Is there evidence to suggest that there are interactions between teachers’ use of the assessment as part of students’ grades and EQAO’s use of the assessment for accountability purposes?
5. How can Kane’s argument-based model of validation be applied to the multiple-use of the EQAO Grade 9 Assessment of Mathematics?

6. In what ways can the concepts of boundary objects and boundary encounters contribute to the process of validation given the multiple-use of the EQAO Grade 9 Assessment of Mathematics?

These questions were not established at the outset but developed throughout the inquiry process. Earlier versions of these questions were used to design the study and identify the data sources. As the research progressed the wording of the questions changed and some questions were dropped and others added in an effort to refine the focus of the inquiry. For example, it was only during the analysis of the data that it became clear that I needed to look for specific examples of practices that would provide the evidence to support or refute a claim that there are interactions between the two uses of the EQAO Grade 9 assessment. Thus, the fourth research question regarding interactions emerged during the analysis process.

Addressing the Research Questions

In this section I describe how the research design addresses each question. There are five sources of data in this study: a province-wide questionnaire, school-level case studies in three schools, analysis of EQAO documents, interviews with EQAO personnel and my research journal. Details regarding each data source and the methods of data construction for each source are provided in the next section.

1. What is the pattern of uses associated with the EQAO Grade 9 Assessment of Mathematics?

To address this research question, I gathered information from EQAO documents and searched the literature for published articles related to the EQAO Grade 9 assessment. Through this approach, I became aware of the uses of the EQAO Grade 9 assessment that are suggested by EQAO as well as some of the ways the assessment is used by other individuals and groups. Drawing information from these sources together, I created the pattern of uses

presented in chapter 6. Within this pattern of uses, I focus on two specific uses: the use of the assessment results by EQAO for accountability purposes and the use of the assessment items by teachers as part of students' grades.

2. How does teachers' use of the EQAO Grade 9 Assessment of Mathematics as part of students' grades take place?

To address this research question, I needed to gather details about how the practice of using the EQAO Grade 9 assessment as part of students' grades takes place. This question could be addressed at many levels and I decided to focus on two aspects. First, I wanted to get a sense of how common teachers' use of the EQAO Grade 9 assessment is across the province. I felt that understanding this was important because teachers' practices in one school might be affected by practices taking place in the schools and districts around them. Fortunately, at the time when I was designing my study, I was also a member of a research team that was conducting a province-wide questionnaire of Grades 7-10 mathematics teachers regarding various aspects of their classroom practice. I was able to include a few items on this questionnaire to gather information about teachers' use of the EQAO Grade 9 assessment. Having the opportunity to gather this data was one reason that this aspect was included in the research design and is a further illustration of the way my research questions emerged as the project unfolded. Details concerning the province-wide questionnaire are provided later in this chapter.

I also wanted to better understand how teachers use the EQAO Grade 9 assessment in their own school and district. To gather this information, I decided to conduct school-level case studies in three schools where the assessment is being used as part of students' grades, making sure that each school was located in a different district. My decision to consider three schools in different districts was based on my sense that teachers' use of the EQAO

Grade 9 assessment as part of students' grades might be influenced by a wide variety of factors. Given the many differences across schools and districts, I felt that examining the practice in only one school would reveal a very limited number of aspects of teachers' use of this assessment. Looking at practices in three schools gives a better sense of the variety of practices that are taking place and of the range of factors which may influence those practices. Because there are three schools in the design, some comparison between the schools is inevitable. However, my focus is on identifying the salient factors which emerge in each school and district rather than on comparing or evaluating practices across these settings.

At each school I interviewed several Grade 9 mathematics teachers, the mathematics department head and the principal. In interviewing a number of individuals in each school, I was not looking for consistency in the practices reported within a school, I was trying to get an understanding of the range of practices and opinions these educators have regarding teachers' use of the EQAO Grade 9 assessment as part of students' grades. Individual teachers within one school may use different practices, and have different views. In addition, teachers' perspectives may differ from those of the mathematics department head or the principal in their school. Drawing on the information from all of these individuals enables me to develop a more coherent account of how teachers use the EQAO Grade 9 assessment as part of students' grades. A description of the procedures I used to make observations at each school is included later in this chapter.

3. How is teachers' use of the EQAO Grade 9 Assessment of Mathematics as part of students' grades viewed by EQAO?

To gain an appreciation of how EQAO regards teachers' use of the assessment as part of students' grades I began with an extensive analysis of EQAO documents. I expected

that the document analysis would reveal adequate information about EQAO's current policies regarding teachers' use of the EQAO Grade 9 assessment as part of students' grades. However, I anticipated that information about EQAO's reasons for allowing this practice to take place as well as the evolution of the policies surrounding this use might not be well documented. In addition, I expected that the document analysis would not provide many insights into the challenges or tensions which emerge from teachers' use of the assessment as part of students' grades from the perspective of EQAO. For these reasons, I included a series of interviews with EQAO personnel in my research design. A description of the procedures I used to gather information about EQAO's perspective is included later in this chapter.

4. Is there evidence to suggest that there are interactions between teachers' use of the assessment as part of students' grades and EQAO's use of the assessment for accountability purposes?

The fourth research question is focused on the interactions which might occur between teachers' use of the assessment items as part of students' grades and EQAO's use of the assessment results for accountability purposes. As noted earlier, I define an interaction as the situation which is created when a practice associated with one use has an impact on the validity of interpretations which can be made for the other use. By understanding the procedures and perspectives of teachers as well as the policy and perspectives of EQAO with regard to teachers' use of the assessment as part of students' grades, interactions between the two uses might become apparent. The approach I used during the analysis of the data to identify the interactions is described later in this chapter.

5. How can Kane's argument-based model of validation be applied to the multiple-use of the EQAO Grade 9 Assessment of Mathematics?

The fifth research question which frames my study is to consider how Kane's argument-based model of validation can be applied to the multiple-use of the EQAO Grade 9 assessment. To address this research question, I attempt to construct an interpretive argument that accounts for the use of the assessment results by EQAO for accountability purposes as well as the use of the assessment items by teachers as part of students' grades. I draw on the various data sources to identify the assumptions and inferences which would be inherent in these two uses and to get a sense of the warrants and backing that might be needed to support those assumptions. In particular, I consider the extent to which Kane's model addresses the interactions between the two uses.

6. In what ways can the concepts of boundary objects and boundary encounters contribute to the process of validation given the multiple-use of the EQAO Grade 9 Assessment of Mathematics?

The sixth research question considers how the concepts of boundary objects and boundary encounters might contribute to the validation of the EQAO Grade 9 assessment. To address this question I draw on the document analysis, school observations, and interviews with EQAO personnel to provide evidence of how the EQAO Grade 9 assessment functions as a boundary object. From these data sources, I also identify activities involving teachers and EQAO personnel that are consistent with the concept of a boundary encounter. I demonstrate how considering the EQAO assessment as a boundary object and identifying these boundary encounters enables a better understanding of the interactions between the two uses and thereby contributes to the process of validation for this assessment.

Summary of the Research Design

This research design enables me to gather detailed information about the multiple-use of the EQAO Grade 9 assessment. It enables me to understand the practices teachers use and to contextualize their practices. The design also enables me to consider the perspectives of teachers, mathematics department heads, school principals, and EQAO personnel with regard to this practice. In focusing on two uses of the assessment (EQAO's use of the assessment for accountability purposes and teachers' use of the assessment items as part of students' grades), this design makes it possible to look for evidence of interactions. In addition, I get a sense of how Kane's argument-based approach to validation might be applied to this instance of multiple-use and of how the process of validation might be enriched by considering the concepts of boundary objects and boundary encounters. The design of this case study is consistent with an interpretivist approach to research, with the sociocultural view of assessment that I adopt and with current notions of the re-conceptualization of validity theory in that I conduct a locally situated study to gather detailed information about the practice of multiple-use, consider a range of perspectives regarding this practice and include a discussion of the contextualization of this practice.

The Process of Data Construction

I concur with the view that data is not simply collected but is deliberately constructed. A number of researchers have observed that the process of data construction is active, decision driven, and value-laden (Ercikan & Roth, 2006; Erickson 1986, 2004). As Erickson (1986) states "fieldnotes, videotapes, and site documents are not data. Even interview transcripts are not data. All these are documentary materials from which data must be constructed through some formal means of analysis" (p.149). According to this view, the

construction of data takes place through the analysis process and the data that is constructed provides evidence of any claims which are made by the researcher (Erickson 1986, 2004). Following this approach, in this section I describe the strategies I used to construct data from each of the five data sources in this study. I begin with an account of my research journal as it overlaps with the other data sources in the study.

My Research Journal

Erickson (1986) maintains that keeping a journal can enhance the plausibility of a research project and is essential for recording events and documenting how a research project changes as data sources are collected and analyzed. I followed Erickson's advice and began making entries in my research journal shortly after my thesis proposal was accepted. I used the journal to record the details of my research activities as well as my thoughts and concerns throughout the inquiry. For example, I made brief notes in my journal while I was conducting interviews and then made more detailed entries immediately after each school visit or EQAO interview was completed. These journal entries provide documentation of aspects of the school visits and EQAO interviews which were not captured on the audio recordings such as the conditions under which the interview took place, comments made by the participant before or after the audio recording was made, and events taking place at the time of the interview which might have had an impact on the participant's comments.

To illustrate the value of my research journal as part of the process of data construction, I have included three excerpts from the journal on the topic of interviewing. These excerpts document some of the ways that I adjusted the interview process as the research proceeded as well as some of my reflections on the challenges I encountered in

conducting a single interview with each research participant. The first excerpt was made immediately after I interviewed three teachers at one of the case study schools.

Three interviews in a row is challenging and this will influence the way I do the EQAO interviews where I will have to do three or four interviews in one afternoon . . . you get confused about which topics have been addressed in which interview . . . for next week I will prepare a ‘tally sheet’ for each interview and use it discreetly to ensure all topics are covered with all participants as much as possible . . . it was interesting to see the value of adjusting the interview protocol according to different participants. The way you ask questions, the way you interrupt or don’t depends a great deal on the dynamic and on the characteristics and personality of the participant. In the third interview today the participant was calm and thoughtful and more areas were approached because he seemed to want to engage in thinking about those different areas whereas the principal had so much he wanted to communicate that there was no time to consider a dialogue or even to ask very many questions. The transcripts will be fascinating but will not be able to make that dynamic especially clear. (Journal entry, April 30, 2008)

The second excerpt was made immediately after I conducted four EQAO interviews in which I used the ‘tally sheet’ approach mentioned in the first excerpt.

The ‘checklist’ approach worked well . . . I told them ahead of time that I was using it to keep myself on track – not to evaluate their responses. (Journal entry, May 7, 2008)

The third excerpt was made after the final case study interviews had been conducted but before all of the EQAO interviews had been completed.

Overall, I feel the limitation of one interview as a technique. Participants seem to feel rushed – and they often are – so they answer quite concisely. Also, they don’t seem to realize that researchers are really interested in the details of their experiences. Just as they are warming up and feeling comfortable the whole thing is winding up . . . there is also a sizable variation across participants – some take their time and are more reflective but most seem to feel rushed by their schedules, by the timetable, by the tape recorder etc. . . . balancing my needs as a researcher for lots of detail, long interviews, a quiet spot and a chance to revisit their ideas later with their realities as teachers is not easy. (Journal entry, May 29, 2008)

These entries indicate some of my reflections on the interview process at a fairly early stage of the inquiry and they connect to several entries later in the journal where I continue to reflect on interviews as a means of constructing data.

In addition to using the research journal to record observations which could not be gathered on the audio recordings, I also continued to make journal entries throughout the analysis of the other data sources and to document the way my ideas about the multiple-use of the EQAO Grade 9 assessment emerged as the research progressed. Thus, I constructed data from the entries in my research journal by making journal entries, and by reading and re-reading earlier journal entries as I conducted the analysis of the other data sources in the study. Reviewing my research journal was also very beneficial as I wrote this account of my inquiry.

Province-wide Questionnaire

In this section I describe the province-wide questionnaire which was one of the data sources in my case study. I provide details concerning the development and distribution of the questionnaire, identify the items which relate to teachers use of the EQAO Grade 9 assessment and summarize the process that I used to analyse teachers' responses to these items.

The Curriculum Implementation in Intermediate Mathematics (CIIM) Research Project

In the period from early 2006 to the fall of 2008, I was a member of a research team that conducted a large-scale study of curriculum implementation and teacher practices for Grades 7-10 mathematics in Ontario, Canada. The study was called the Curriculum Implementation in Intermediate Mathematics (CIIM) project and was conducted by a team of researchers at the University of Ottawa (Suurtamm & Graves, 2007). The project included

case studies in intermediate mathematics classrooms, focus group interviews with mathematics educators, and a web-based questionnaire which Grades 7-10 mathematics teachers across the province were invited to complete. As a member of this research team, I contributed to various aspects of the project by conducting literature searches, gathering and analysing observations from focus group interviews and case studies, and working collaboratively on the design, administration and analysis of the province-wide questionnaire. Participation in the CIIM project was a valuable experience which provided an opportunity for me to further develop my research skills. My participation in this project also contributed to my dissertation in more direct ways. First, I was able to include items about teachers' use of the EQAO Grade 9 assessment on the province-wide CIIM questionnaire. The development, administration and analysis of this questionnaire are described in the next section of this chapter. Another way my involvement with the CIIM project contributed to my dissertation was in helping me locate schools willing to participate in my research. More specific details about this are also included in this chapter. The third benefit of the CIIM research for this inquiry was that I became more familiar with the mathematics curriculum and with some current classroom practices in Grade 9 mathematics in Ontario. This familiarity was an asset as I drafted interview guides, conversed with interview participants and conducted the data analyses for this study and it is another aspect of the process of contextualization that took place in this inquiry.

Questionnaire Development and Distribution.

The CIIM questionnaire was designed to address the research questions of the CIIM project (Suurtamm & Graves, 2007). In developing the 45 items on this questionnaire, the research team examined other questionnaires for mathematics teachers including those from

the Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA) and a questionnaire that EQAO administers to gather information on teachers' instructional practices and attitudes regarding mathematics teaching. A paper version of the CIIM questionnaire was pre-tested by distributing it to a group of practising teachers. This pre-test led to minor revisions to some of the items and helped determine the amount of time participants would need to complete the questionnaire. A web-design team then transformed the paper version into a web-based questionnaire.

The web-based questionnaire was distributed to intermediate mathematics teachers across the province. The distribution process began with sending letters to the directors of every English-language school district in the province inviting them to participate in the research. School districts who accepted our invitation were asked to indicate the number of schools in their board where intermediate mathematics was taught. With that information, we prepared packages for each school consisting of a covering letter for the school principal and letters to be distributed to each intermediate mathematics teacher in that school. The letters to the mathematics teachers described the project and invited them to participate by responding to the online questionnaire. Each letter included information about how to access the questionnaire on the internet as well as an individual access code. The access codes were intended to ensure that the questionnaire was only completed by teachers who were teaching intermediate mathematics at the time the questionnaire was administered. All questionnaire responses were anonymous but teachers were asked to indicate their school district at the end of the questionnaire. The questionnaire was available online from May to mid-July 2006. A total of 42 of the 60 English-language school boards in Ontario participated in the questionnaire and 1,096 responses from mathematics teachers in Grades 7-10 were received.

CIIM Questionnaire Items Related to the EQAO Grade 9 Assessment

The CIIM questionnaire included 43 selected-response items and 2 open-ended items. The items address a range of topics such as teaching experience, background in mathematics, classroom practices, and professional development. For some of the items on the questionnaire, teachers were asked to identify one mathematics class that they were teaching during the school year and to use that class as the basis for their answers. Among the 1096 Grades 7-10 mathematics teachers who responded to the CIIM questionnaire, 272 indicated that they were describing their experiences teaching Grade 9 Academic (n=150) and Grade 9 Applied (n=122) mathematics classes.

Three of the selected-response items on the CIIM questionnaire asked teachers about different aspects of their use of the EQAO mathematics assessment. One of these items directly relates to teachers' use of the EQAO assessment as part of their students' grades and teachers' responses to this item are included in this dissertation. The wording of the item was "In this class to what extent do you use EQAO assessment results in mathematics to form part of student's final mark". Participants were asked to respond by selecting 'Not at all', 'A little', 'Some' or 'A lot'. In addition to this selected-response item, an open-ended item on the CIIM questionnaire asked teachers to describe a beneficial professional development experience they had had with regard to mathematics teaching. Although this item was not focused on the EQAO Grade 9 assessment, it generated some responses related to teachers' involvement with EQAO activities and for this reason, this item became an additional data source for my inquiry. The wording of this item was "Please describe a professional development experience that has positively influenced the way you teach mathematics".

Analysis of Questionnaire Responses

One of my contributions to the CIIM project was to conduct the analysis of the questionnaire data. Because the questionnaire was web-based, responses from participants were automatically stored in a data file. The data file was sorted by questionnaire numbers which were automatically assigned as each participant entered their access code and completed the questionnaire. Each participant is represented as a row in the data matrix and each variable is represented as a column. Because many of the items had several parts, the 45 items on the questionnaire generated a total of 284 variables. This data file was imported directly into SPSS for analysis. In the analysis of the selected-response items, descriptive statistics including frequency counts, percentages and cross-tabulations were used to summarize teachers' responses. The use of descriptive statistics is appropriate because almost all of the selected-response questionnaire items generated categorical responses rather than interval level data. Responses to the selected-response item that relates to teachers' use of the EQAO Grade 9 assessment are presented in chapter 6.

Responses to the open-ended questionnaire items were automatically compiled by the web-based questionnaire software as a text file which specified each participant's questionnaire number and provided the full text of their responses. A content analysis was conducted for each open-ended item. In this analysis, responses were sorted according to a variety of topics that were of interest in the CIIM project. One of the topics included in this analysis was large-scale assessment. Responses to the open-ended items which mentioned large-scale assessment were copied into a separate file. Within that file, I flagged any responses that mentioned the EQAO Grade 9 assessment. Through this process, I located four responses to one open-ended item in which teachers identified their involvement with

the EQAO assessment as a professional development experience that had positively influenced the way they teach mathematics. These four responses provide additional evidence of the way that teachers' involvement in EQAO activities has influenced their use of the EQAO assessment as part of students' grades and are discussed in greater detail in chapter 7.

School-level Case Studies

The school-level case studies I conducted in three schools where the EQAO Grade 9 assessment is used as part of students' grades are the third data source in this study. In this section I describe how I located the participating schools, the process I used to gather contextual information about these schools, the procedures I used to identify and interview participants and the process I used to analyse and construct data from these observations.

Identifying Schools for the Study

The main criterion for selecting schools for this study was that the EQAO Grade 9 assessment was being used as part of students' grades by at least some teachers in the school. The three schools who participated in this study were identified as a result of my involvement in the CIIM project. Two of the schools, Scottsdale Collegiate and Taylor Brook High School were case study schools in the CIIM project.²⁵ During the case study phase of the CIIM project, the topic of the EQAO Grade 9 assessment came up several times as we conducted observations in various schools. Several CIIM participants at Scottsdale Collegiate and Taylor Brook High School expressed interest in further discussing these ideas when they learned of my dissertation research and both schools subsequently agreed to participate in my research. Birch Park Secondary School was not a case study location in the CIIM project. However, access to Birch Park was also obtained as a result of my

²⁵ Throughout this dissertation all individual participants and schools are identified using pseudonyms.

involvement with the CIIM project in that another researcher on the CIIM team who was employed by the school district in which Birch Park is located expressed interest in my research and facilitated my access to the school by describing my research to the director of the school district and to a principal at one secondary school in the district. Both the school district and the principal of the case study school subsequently gave their consent to participate in my research.

Gathering Information about the Schools

I gathered information about each school from a number of sources. One source of information was each school's website. From these websites I gathered information about the number of students at the school, school improvement plans which describe areas of concern or focus for the school and school-wide assessment policies. Another source of information was each of the school district websites. From these websites I gathered information about the number of high schools in the district, specific initiatives in the district, as well as board-wide assessment policies and procedures. I also reviewed the EQAO Grade 9 assessment score reports for each school and district for the previous five years. As will be shown in chapter 6, using this information, helped me to contextualize the information participants provided about teachers' use of the assessment as part of their students' grades.

Identifying Interview Participants in each School

In each school I conducted interviews with the principal, the mathematics department head and several Grade 9 mathematics teachers. After consent to conduct the research was obtained from each school district and principal, participants were recruited at each school. To begin this process an email was sent to each principal describing the research and inviting

them to participate in an interview. Another email was sent to each mathematics department head inviting them to participate in an interview and asking them to forward my invitation to participate in the research to all of the teachers in their department teaching Grade 9 Academic or Grade 9 Applied mathematics courses that year. Individuals were asked to contact me if they were interested in participating in the research and interviews were scheduled based on each participant's availability. The number of teachers interviewed at each school varied on the basis of teachers who were interested in participating in the research or who were available to complete the interview. A summary of the interviews I conducted in the three schools is provided in Table 6.

Table 6: Overview of Interviews in School-level Case Studies

School Pseudonym	Interviews Conducted
Birch Park Secondary School	• Principal • Mathematics Department Head • Individual interviews with three Grade 9 mathematics teachers
Scottsdale Collegiate Institute	• Principal • Mathematics Department Head • Individual interview with one Grade 9 mathematics teacher • Comments about the use of the EQAO Grade 9 assessment made by one Grade 9 mathematics teacher in a CIIM focus group interview
Taylor Brook High School	• Principal • Mathematics Department Head • Individual interviews with three Grade 9 mathematics teachers

Ethical Considerations for School-level Case Studies

Prior to commencing my research in schools, I followed the University of Ottawa ethics procedure to obtain approval to conduct this research and to get informed consent from each participant. Erickson (1986), among others, maintains that research participants

need to be as informed as possible of the purposes of the research and to be protected from risks such as embarrassment and administrative sanction which might result from their participation. Erickson also notes that many research participants assume that the researcher's purposes are in some way evaluative and I have observed this tendency on a number of occasions when interviewing participants for other research projects. In this inquiry, I was concerned that participants might have the impression that my purpose was to evaluate the practices they use in marking and including the EQAO Grade 9 assessment items in their students' grades. In particular, I was concerned that references to *validity* in the description of my research might be misinterpreted by teachers as an intention to evaluate their practice. As I wanted participants to feel comfortable in describing what they do as they use the assessment and not to think that I was evaluating their practices, I did not use the word validity in the description of the project I included on the consent form. I also tried to reduce the feeling that participants might have of being evaluated during the interviews by avoiding making evaluative comments when participants were describing their practices even if those practices were not consistent with the policies of their school, district, EQAO or the Ministry of Education.

Interview Protocols

Interviews at each of the three schools were conducted using the interview protocols I developed prior to beginning the school visits. These protocols were used as a guide and interesting topics which emerged during each interview were also pursued. In addition, ideas that emerged in the interviews at one school influenced the interviews I conducted in the other schools as well as the interviews I conducted with EQAO personnel. In this way, my interview questions varied somewhat as I became aware of new practices and additional

relevant issues. I began conducting interviews in the first school (Birch Park Secondary School) but had not completed all the interviews in this school when I began conducting the interviews in the second (Scottsdale Collegiate) and third (Taylor Brook High School) schools. All interviews were audio-recorded and transcribed and follow-up emails were sent to some participants to clarify comments they had made or to ask about a detail that been overlooked in the interview.

One protocol was developed and used for the interviews with the mathematics teachers and department heads (see Appendix F). I developed the questions in this protocol and then refined them on the basis of a pilot interview I conducted with a high school mathematics teacher and department head. In these interviews, I began by seeking some information about each participant's teaching experience and then posed an open-ended question asking each participant to describe, in as much detail as they could, the process they used to include the EQAO Grade 9 assessment as part of students' grades. I then referred to the more specific questions in the protocol and asked some of these questions depending on the details each participant had included in their response to the initial prompt. Through this process of asking a general question followed by more specific questions, I was able to get a sense of many aspects of teachers' use of the EQAO Grade 9 assessment as part of students' grades including how the assessment items are chosen and scored, the amount the assessment contributes to students' grades and the reasons each participant had for using the assessment items for this purpose. The other advantage of beginning with a general question was that participants could mention issues that I was not aware of rather than answering the specific questions I had thought of in advance of the interview. In this way, I became aware of a number of issues related to teachers' use of the EQAO Grade 9 assessment that I was

not familiar with when I developed the interview protocol. My approach to interviewing follows many of the suggestions made by Seidman (2006) with respect to the value of asking open-ended questions and the importance of further exploring issues which emerge during the interview.

The interviews with the mathematics department heads were based on the same protocol that I used with teachers but I encouraged department heads to respond to each question both in their capacity as department head and with respect to their own experiences as teachers using the EQAO Grade 9 assessment as part of their students' grades. At the end of each interview, I asked if the participant had any other comments about using the EQAO assessment as part of students' grades. This question was intended to give participants an opportunity to voice any opinions which they had not yet expressed and to emphasize or clarify ideas that they had mentioned earlier in the interview. In closing, I encouraged each participant to contact me by email if they thought of things they would like to include after the interview was complete.

A separate protocol was developed and used to interview the school principals (see Appendix G). These interviews followed a similar approach to the one I used with teachers and department heads. I began by seeking some information about each participant's experience as a principal, then posed an open-ended question, asking each principal to describe how the assessment was used as part of students' grades in their school, and following up with specific questions from the protocol depending on what the principal had included in their initial response. The topics I asked principals about include: when the use of the EQAO Grade 9 assessment as part of students' grades first started in their school, the benefits and difficulties which they associate with this use of the assessment, their awareness

of school, board or Ministry of Education policies related to this practice, as well their sense of how other schools in their school district use the EQAO Grade 9 assessment as part of students' grades. I ended the interview with each principal as I had done with teachers and department heads by asking if they had any other comments and by encouraging them to contact me if they thought of any additional comments they would like to make.

An Overview of the Visits to each School

While the same approach was used for each school in the case study, the number and sequence of visits to each school varied based on the availability of teachers, department heads and principals. I visited Birch Park Secondary School on three occasions. The first visit took place on April 14, 2008 and during this visit I interviewed the principal in the school's main office. The second visit took place on April 24, 2008 when I interviewed the mathematics department head in the mathematics department office during his preparation period. The third visit took place on May 29, 2008 when I conducted three separate interviews with teachers who had been teaching Grade 9 mathematics courses during the winter semester. These interviews were conducted in the mathematics department office during the lunch break. The interviews at Birch Park Secondary School varied in length from 7 to 33 minutes and were conducted in the month immediately preceding the administration of the EQAO Grade 9 assessment for that semester.

I visited Scottsdale Collegiate on two occasions. The first visit took place on April 18, 2008 when I interviewed the principal. As this interview also included some content related to the CIIM project, it was conducted jointly by me and one of the CIIM principal researchers. The interview took place in a small meeting room in the school's main office. My second visit took place on June 24, 2008 when I conducted individual interviews with

one Grade 9 mathematics teacher and with the mathematics department head. Each of these interviews took place in a mathematics classroom amidst stacks of returned textbooks as teachers prepared for the end of the school year. As the interviews on this second visit did not relate to the CIIM project, I conducted them by myself. The interviews at Scottsdale Collegiate varied in length from 14 to 25 minutes and were conducted approximately two weeks after the administration of the EQAO Grade 9 assessment for that semester had taken place. This meant that teachers had just finished marking the EQAO Grade 9 assessment and including it as part of their students' grades. Additional information from another Grade 9 mathematics teacher at Scottsdale Collegiate was obtained from a focus group interview conducted as part of the CIIM project a few months prior to my visits to this school.

I visited Taylor Brook High School over a period of four days at the end of May 2008. As this school was a case study site for the CIIM project, I conducted my interviews at the same time as the CIIM case study was being conducted. The interview with the mathematics department head, who was teaching two classes of Grade 9 Applied mathematics, took place on the first day of the CIIM case study. This interview was conducted by me and one of the CIIM principal researchers and took place at the end of the school day in the participant's classroom. On the second day, the interview with the principal was conducted. As this interview also included questions related to the CIIM project, it was conducted by me and one of the CIIM principal researchers and took place in the principal's office. On the final day of the CIIM case study I conducted three separate interviews with teachers who were teaching Grade 9 mathematics courses at Taylor Brook that semester. These interviews were conducted during the teachers' preparation period and took place in the mathematics department office. As these interviews did not include any

questions related to the CIIM project, I conducted them by myself. The interviews at Taylor Brook High School varied in length from 8 to 96 minutes and were conducted about a week before the beginning of the EQAO testing period for that semester.

Analysis of School-level Case Studies

In my analysis of the school-level case study observations, I began by identifying some background information for each school such as the size of the school and district, the geographic location, EQAO results etc. As noted, my observations from the schools included information from the school and district websites, EQAO Grade 9 assessment results for each school and district for the previous five years, notes made in my research journal during my school visits as well as transcripts from the audio-recorded interviews. I reviewed each of these data sources and used the information in them to write descriptions of each school.

I then analysed the interview transcripts in order to develop a detailed description of how teachers use the assessment as part of their students' grades. The process I used was an iterative one which evolved somewhat as the analysis proceeded. To begin, I read through each transcript a number of times. As I read through the transcripts, I created a number of observation tables with each table focused on a different aspect of teachers' use of the EQAO Grade 9 assessment as part of their students' grades. For example, I created a table for the way participants select the EQAO assessment items they will score, a table for the procedures they follow to score the items, a table for the meaning participants ascribe to the EQAO assessment results, and a table to record the tensions that participants associated with their use of the EQAO Grade 9 assessment as part of students' grades. While some of the topics for these observation tables were clear to me as I began the analysis process, others became evident as I read the transcripts and re-considered my research questions. In this

way, the topics for the analysis were guided by the research questions but also emerged from the data. Within each observation table, I made a separate entry for each comment related to the topic of that table. For each entry, I identified the participant, wrote a brief summary of the comment and/or cut-and-pasted the text from the interview transcript into the table and indicated the page number where the comment could be found. Thus, each table includes my observations about one aspect of teachers' use of the EQAO Grade 9 assessment across the three school locations.

This analysis process resulted in ten observation tables detailing various aspects of teachers' use of the EQAO Grade 9 assessment. In an effort to report these findings in a reasonably concise manner, for this dissertation I have combined my observations from some of these tables and I have omitted those tables that are not directly related to my research questions. Accordingly, the following six topics from my analysis are reported in this dissertation:

1. *Weighting* (the amount that teacher-derived EQAO assessment scores contribute to a students' overall grade for the course)
2. *Selection of items to be scored* (which EQAO assessment items teachers and/or department heads score and how the items are chosen)
3. *Scoring procedures* (how teachers and/or department heads score the items they select)
4. *Meaning ascribed to the official and teacher-derived EQAO scores* (as indicated by teachers, department heads and/or principals)
5. *Rationale and benefits of using the assessment as part of students' grades* (as indicated by teachers, department heads and/or principals)
6. *Tensions emerging from the use of the assessment as part of students' grades* (as indicated by teachers, department heads and/or principals)

The first three topics focus on the procedures teachers use to include the EQAO assessment items in their students' grades while the last three topics focus more on participants' views of this use of the assessment. As I analysed the interview transcripts,

some teachers and department heads talked about their experiences in the context of their current school and contrasted those experiences with experiences they had had in using the EQAO Grade 9 assessment for grades at other schools in the same school district where they had previously taught. In constructing the observation tables, I decided to include the experiences of these teachers from their previous schools as their comments often include variations in practice which I had not observed elsewhere and provide evidence of the variety of practices that are used within each of the school districts in the study.

Once I had constructed the tables for these six topics, I reviewed the entries in them and wrote a summary of each topic across the three schools. For some of the aspects, before I began to write the summary I organized the entries within the observation table into themes. For example, in the table focused on tensions that emerge from teachers' use of the EQAO assessment as part of students' grades, I organized the entries according to six themes (i.e. tensions related to providing feedback to students and parents, tensions related to time constraints and test security etc.). While the initial draft of the summary for each topic was based on my analysis of the interview transcripts, I also reviewed my research journal, the information that I had gathered from the EQAO score reports and from the school and district websites and added any information that would contribute to a coherent description of how teachers use the assessment as part of students' grades in these schools. I then read each transcript again and compared it with the summaries to ensure that I had adequately reflected the range of practices and views expressed by the participants. Throughout this process, I was mindful of Erickson's (1986) advice to avoid placing emphasis on comments or events that recur frequently in data sources at the expense of less common but equally important comments or events. In constructing the data for this study, my interest was more

on the range of practices and views that were expressed by each participant than in the frequency that a particular practice or view occurred and for this reason I did not make frequency counts of the entries in each table. The summaries which resulted from this analysis are presented in chapter 6.

Two additional stages in my analysis of the observations from the schools took place once these summaries had been written. First, I read through the analysis tables in order to identify evidence of interactions between teachers' use of the assessment as part of students' grades and EQAO's use of the assessment for accountability purposes. More detail regarding how this part of the analysis was conducted is provided in the section of chapter 6 where I describe the interactions that I discovered. The final stage in my analysis of the observations from the schools was to create a table where participants' comments that might be considered evidence of boundary objects and encounters were listed. More detail regarding how this analysis was conducted is provided in chapter 7 where I discuss how I applied the ideas of boundary objects and encounters to the multiple-use of the EQAO Grade 9 assessment.

EQAO Document Analysis

My analysis of EQAO documents took place over a period of 18 months that began before I conducted the school-level case studies and continued for several months after the interviews were complete. I conducted a close reading of all documents currently provided on the EQAO website which pertain to the EQAO Grade 9 assessment of mathematics. These documents include: successive versions of the Grade 9 assessment framework (EQAO 2006, 2009c), technical reports (EQAO 2001; 2008, 2009d), research papers commissioned by EQAO (as listed in chapter 4), administration guidelines for principals and teachers

(EQAO 2007), annual assessment reports for various schools, districts and the province, released items, scoring rubrics and anchor responses for the Grade 9 Academic and Applied assessments, media releases, and a number of other documents.

As I reviewed each document I made notes of any information related to teachers' use of the EQAO Grade 9 Assessment of Mathematics as part of students' grades. The information gathered from these documents was an important part of establishing the pattern of uses for this assessment (described in chapter 6) and for understanding the current policies EQAO has regarding teachers' use of the EQAO Grade 9 assessment as part of students' grades. This EQAO document analysis also guided the development of the interview protocols for the interviews I conducted at the schools and with EQAO personnel. In addition, I used the information from the EQAO document analysis along with the information from my interviews with EQAO personnel to write a summary of the perspective of EQAO regarding teachers' use of the Grade 9 assessment as part of students' grades.

Interviews with EQAO Personnel

The last data source in this case study that remains to be described are the interviews that I conducted with EQAO personnel. In this section I summarize the process I used to conduct the interviews and analyze the interview transcripts.

Identifying Interview participants at EQAO

To recruit EQAO personnel to participate in this research, I contacted a research officer at EQAO and asked her to distribute an email to individuals working on the Grade 9 assessment inviting them to participate in this research. Four individuals who are currently involved with the development and administration of the EQAO Grade 9 assessment

indicated that they were willing to participate in my study (Marie, Mitchell, Melanie and Dana). These four individuals have varying amounts of experience with the Grade 9 assessment. Both Mitchell and Marie have a considerable amount of experience as they worked on the assessment when it was originally being developed and throughout the years as changes were made to the assessment. Dana also has a lot of experience with the Grade 9 assessment as she began working at EQAO shortly after the first year the assessment was administered. One participant, Melanie, is fairly new to EQAO and had only three years experience working on the Grade 9 assessment.

After I had conducted the interviews with these four individuals, I had a number of remaining unanswered questions so I contacted two other individuals who had been mentioned as valuable sources of information by some of the first four participants. One of these participants, Susan, worked on the earliest versions of the EQAO Grade 9 Assessment of Mathematics and has continued to work for EQAO in a number of other capacities since that time. The other individual, Carol, is a consultant who has been hired by EQAO to work on a number of projects involving the EQAO Grade 9 assessment over a period of several years. In the end, I conducted a total of six interviews with individuals who had been involved in the design, development, and administration of this assessment. I also contacted the director of assessment and reporting at EQAO by email to clarify a few points that remained unclear after discussing this use of the assessment with these six EQAO personnel.

Interview Protocol

The interviews I conducted with EQAO personnel were guided by the questions shown in Appendix H. I began each interview by asking the participant to describe their position at EQAO and their experience with the EQAO Grade 9 assessment. I then asked an

open-ended question in which I indicated that I was interested in learning about EQAO's view of teachers' use of the Grade 9 assessment as part of students' grades including any policies that EQAO had established and the reasons EQAO had for permitting teachers to use the assessment in this way. I asked each participant to provide as much detail as they could about these topics. As shown in Appendix H, I followed this open-ended question with a series of specific questions depending on the response each participant had given to the initial prompt. I asked the open-ended question at the beginning with the hope that participants would feel free to comment on aspects of this practice that they felt were important and so that the interview would not be limited to those issues I was aware of before the interviews began. As with the interviews I conducted in the schools, I ended each interview by asking if the participant had additional comments and indicating that they should feel free to contact me if they thought of other comments they would like to make. The interview protocol was used as a guide and themes or topics that emerged from participants' comments were explored as each interview progressed. In addition, insights gained from earlier interviews sometimes influenced the questions which were posed in subsequent interviews.

Conducting Interviews with EQAO Personnel

To conduct the interviews with EQAO personnel, I travelled on two occasions to EQAO'S main office in Toronto, Ontario. During my first visit, in May 2008, I conducted the first four interviews with EQAO personnel. These interviews were conducted in a conference room at EQAO's offices where I met with each participant separately. The second visit took place in October 2008. During this visit, I conducted a single interview in which both participants took part. These two participants had worked together on various

aspects of the EQAO Grade 9 assessment over a period of several years and felt that they could provide more comprehensive responses if they were interviewed at the same time. The interviews with the six EQAO personnel were between 14 and 69 minutes in duration and were audio-recorded and transcribed. Follow-up emails were sent to some participants to clarify some details from the interviews. As with the interviews in the schools, I made brief notes in my research journal during each interview and more extensive notes immediately after each interview was finished.

Analysis of EQAO Interviews

In my analysis of the EQAO interviews, I followed a similar approach to the one I used for the interviews in the schools. I read each transcript a number of times and created tables of observations related to three topics: EQAO's policy and rationale for permitting teachers' to use the Grade 9 assessment as part of students' grades, the tensions identified by EQAO personnel related to teachers' use of the assessment, and the views expressed by EQAO personnel with regard to the comparability of EQAO results given the variation in teachers' use of the assessment for students' grades across schools and districts. In each table entry I summarized the participant's comment and/or cut-and-pasted the text from the transcript directly into the table. I noted the participant's pseudonym and the page in the transcript where the comment could be found. I used the information in these tables to write a description of EQAO's perspective on these three aspects of teachers' use of the Grade 9 assessment. As I wrote these descriptions, I reviewed the information recorded in my research journal, gathered from my analysis of EQAO documents and contained in the email communications I had with the EQAO director and added any relevant points. Once I had written a draft of each description, I read each transcript once again to ensure that I had not

missed any details and that I had reflected the range of opinions expressed by the participants. The descriptions which resulted from this analysis are included in chapter 6.

There were two additional stages in my analysis of the EQAO interviews. One of these was that I read through the tables to identify evidence of interactions between teachers' use of the assessment as part of students' grades and EQAO's use of the assessment for accountability purposes. More detail regarding how this part of the analysis was conducted is provided in the section of chapter 6 where I describe the interactions that I discovered. The final stage in my analysis of the EQAO interviews was to search the transcripts for comments which suggest that the assessment may function as a boundary object or encounter. More detail regarding this aspect of the analysis is provided in chapter 7 where I discuss the use of the concepts of boundary objects and encounters in validating the multiple-use of the EQAO Grade 9 Assessment of Mathematics.

I also want to describe an aspect of my analysis of the EQAO interviews that I had not anticipated in my research design. During my interviews with the six EQAO personnel, I discovered that five of them had previously been mathematics teachers and four had used the Grade 9 assessment as part of students' grades in their own classrooms prior to working at EQAO. During the interviews, these four individuals provided their opinions as EQAO personnel and they also described some of the approaches they had taken as teachers to include the assessment in students' grades. Some of the procedures described by these individuals were different than those that I had learned about through my school-level case studies. Accordingly, as part of the analysis of the EQAO interviews I read the transcripts of those participants who had been teachers and added their comments to the tables that I had created for weighting, item selection and scoring in the analysis of the interview transcripts

from the schools. I also included the practices described by these EQAO former teachers in my summaries in chapter 6 whenever the practice was a variation that I had not observed in the three schools. I include these EQAO former teachers' experiences in my summaries to demonstrate the wide range of practices that take place across Ontario as teachers use the EQAO Grade 9 assessment as part of students' grades.

CHAPTER 6

OBSERVATIONS OF THE MULTIPLE-USE OF THE EQAO GRADE 9 ASSESSMENT

In this chapter I present my observations of the multiple-use of the EQAO Grade 9 Assessment of Mathematics which address the first four research questions (the fifth and sixth research questions are addressed in chapter 7). The chapter is divided into four main sections, one for each research question. As some of the sections are quite long, a brief summary of the findings for each research question is included at the end of each section and a discussion of the findings across the four research questions is included at the end of the chapter.

Research Question 1 – The Pattern of Uses

What is the pattern of uses associated with the EQAO Grade 9 Assessment of Mathematics?

To address the first research question, I draw on information from my analysis of EQAO documents as well as the interviews I conducted with EQAO personnel. Using these data sources, I explain the challenges I encountered in identifying the intended use of the EQAO Grade 9 assessment and then identify the multiple-uses I found for this assessment. I also discuss how the multiple-use of the EQAO Grade 9 assessment relates to the practice of multiple-scoring (which was described in chapter 2) and describe EQAO's current policy regarding teachers' use of the EQAO Grade 9 assessment.

Identifying the Intended Use of the EQAO Grade 9 Assessment

In chapter 2 I demonstrated that large-scale assessments are often used in complex ways. I argued that understanding the pattern of uses associated with an assessment is an essential part of the process of validation and suggested that one way to establish the pattern of uses is to begin by identifying the intended use of an assessment and then look for

evidence of other uses. Accordingly, to establish the pattern of uses associated with the EQAO Grade 9 assessment I began by attempting to determine the intended use for this assessment. As noted earlier, EQAO was established with a mandate to develop and administer province-wide tests, to evaluate the accountability of school districts to the public and to evaluate the quality and effectiveness of education in Ontario (Government of Ontario, 1996, s.3). With specific reference to the Grade 9 assessment, EQAO states that the assessment “evaluates the knowledge and skills that The Ontario Curriculum expects students to have learned by the end of Grade 9” (EQAO, 2009c, p.5). EQAO identifies the purpose of this assessment as “to assess the level at which students are meeting the provincial curriculum expectations in mathematics up to the end of Grade 9” (EQAO, 2009c, p.6) and to report these results for individual students, schools, school districts and the province. Thus, both the EQAO mandate and the explicitly stated purpose for the Grade 9 assessment suggest that the intended use of this assessment is for accountability purposes.

And yet, as I continued to analyse the EQAO documents and the interviews I conducted with EQAO personnel, I found considerable ambiguity regarding the intended use of this assessment. One source of ambiguity originates in the EQAO framework document itself where, as quoted in the previous paragraph, the purpose of the Grade 9 assessment is stated (EQAO, 2009c). In this framework document, in addition to stating the purpose for the assessment, EQAO provides a list of “benefits” of the assessment (EQAO, 2009c, p.6).²⁶ The benefits which are listed are: providing data to assist schools and boards in improvement planning and target setting, supporting the successful implementation of the curriculum, improving understanding of assessment practices and levels of achievement among

²⁶ Both the 2006 and 2009 versions of the framework document include similar statements regarding the purpose and benefits of the Grade 9 assessment.

educators, and improving understanding of assessment practices among the public (EQAO, 2009c, p.6). The framework document does not explain how these benefits differ from the purpose of the assessment and arguably, some or all of these benefits could be considered intended uses of this assessment. If these benefits are essentially additional uses then evidence supporting them would need to be gathered as part of the process of validation.

It is interesting to note that the benefits which are claimed by EQAO for the Grade 9 assessment are similar to many of the purposes which Stobart (2009) identified as being associated with the national curriculum assessments that are used for accountability purposes in England (as described in chapter 2). Stobart indicates that there are 14 purposes for the national curriculum assessments and he identifies six of the “key purposes” in his article: (i) to assess the attainment of individual pupils in English and mathematics at the end of specific grades; (ii) to assess the performance of schools, local authorities and government by aggregating the results; (iii) to help teachers assess their pupils’ attainment and complement teachers’ own assessments; (iv) to set the standard of attainment in each subject; (v) to raise standards by setting national targets based on levels achieved by pupils, schools and local authorities; and (vi) to monitor changes in standards over time (Stobart, 2009, p. 167). While the language that is used with reference to the national curriculum assessments differs somewhat from that used for the EQAO Grade 9 assessment, many of the purposes for these two assessments are very similar. Specifically, scores from both assessments are used to assess individual students, aggregated to assess the performance of schools and districts, used to improve the assessment practices of teachers and used to set targets for improvement planning. The similarity between the benefits indicated by EQAO and the multiple purposes of the national curriculum assessments provides support for

considering many of the benefits of the EQAO Grade 9 assessment to be additional uses that should be considered as part of the process of validation for this assessment. This similarity between these two assessments also makes me wonder if many other large-scale assessments claim a similar range of uses and/or benefits for their assessments.

In further considering the benefits listed in the EQAO framework document, I suggest that some benefits may be characterized as intended consequences of the assessment while others constitute additional intended uses for the assessment (see chapter 3 for a discussion of intended and unintended consequences of large-scale assessments). For example, providing support for the implementation of the curriculum and improving the understanding of assessment practices by the public may be appropriately viewed as intended consequences of the assessment in the sense that these benefits are actually consequences which result from the use of the Grade 9 assessment for other uses. I consider these benefits to be *intended* rather than *unintended* consequences because they are explicitly identified by EQAO in the framework document rather than emerging as an unanticipated result of the use of the assessment. However, I would argue that benefits such as providing data to assist with improvement planning and target setting are more appropriately identified as multiple-uses of the assessment because the assessment results are specifically used for these purposes (i.e. to set targets and track improvement). Thus, while the framework document explicitly states a purpose for this assessment, the list of benefits in the framework creates confusion with regard to the intended use(s) of the EQAO Grade 9 assessment and distinguishing the intended use(s) of the assessment from intended consequences and additional multiple-uses is not a straightforward matter.

Further ambiguity regarding the intended use of the EQAO Grade 9 assessment is found in the administration manual which is distributed to teachers and principals each year (EQAO, 2007). It is this document which indicates that teachers may select some or all of the items on the assessment, mark students' responses to these items and include the results as part of their students' course grade. Surprisingly, while teachers' use of the assessment as part of students' grades is included in the administration manual, it is not included in the list of benefits in the framework document nor is it mentioned anywhere else in the framework document. From reading the administration manual and the framework documents, it is not clear how teachers' use of the assessment as part of students' grades is distinguished by EQAO from the explicitly stated purpose or from the identified benefits of the assessment. I contend that teachers' use of the assessment as part of students' grades is an additional multiple-use of the assessment, and one which, as will be shown, has been permitted since the assessment was first administered. Thus, in some senses, teachers' use of the assessment as part of students' grades is difficult to clearly distinguish from the intended use of the assessment.

Further evidence supporting the contention that some of the benefits in the framework document can be considered to be part of the intended use for this assessment is found in the interviews I conducted with EQAO personnel. One participant, Susan, who was involved with the EQAO Grade 9 assessment in the early stages of the development process indicated that improving teachers' classroom assessment practices and encouraging teachers to implement the mathematics curriculum more fully were important purposes for this assessment from the outset. Susan offered two examples to illustrate how these purposes were reflected in the assessment items. She noted that the development of items on the

Grade 9 assessment which refer to graphing calculators and motion detectors was intended to increase teachers' use of technology in their classrooms. In addition, the inclusion of items on the assessment which draw on more than one strand was intended to encourage teachers to integrate mathematics strands as they develop their own classroom assessment activities.

Susan states:

Geometry was at the end of the curriculum and was at the end of the textbook and was at the end of whatever else so teachers would say "well, I haven't taught that to the kids yet" and our [EQAO's] comment was always "you don't start at the beginning and work through the curriculum. You don't have to start at the beginning of the textbook or the curriculum" . . . so one of the things we had done in our early questions was a number of questions were cross-strand questions so that we could demonstrate how you could be assessing measurement and whatever else. (Carol & Susan interview, p. 20).

Susan's comment suggests that more than one intended purpose has been associated with the EQAO Grade 9 assessment from the outset.

The ambiguity regarding the intended use(s) of the EQAO Grade 9 assessment which I found in my analysis of EQAO documents and in my interviews with EQAO personnel has also been discussed in several articles. Earl and Torrance (2000) argue that each assessment program at EQAO had dual purposes from the outset; accountability and school improvement and Wolfe et al. (2004) demonstrate that the purposes of the EQAO assessment have expanded since the assessment was developed. In addition to accountability and school improvement which Wolfe et al. identify as the two original purposes for the assessment program, these authors point out that the EQAO Grade 9 assessment has been used for several other purposes including individual student reporting, modeling assessment practices for teachers and encouraging teachers to implement the mathematics curriculum. Klinger et al. (2008) make similar observations about the expanding purposes of the EQAO assessments and argue that changes in the design of these assessments are needed to support

these additional purposes. Thus, my case study observations demonstrate that identifying the intended use of the EQAO Grade 9 assessment and distinguishing it from intended consequences and multiple-uses of the assessment is a challenge and these findings are consistent with existing literature concerning the purpose of this assessment.

In spite of this ambiguity, in my description of the pattern of uses for this assessment I decided to identify accountability as the intended use of the EQAO Grade 9 assessment and then treat the other uses as additional multiple-uses. I made this decision for two reasons. First, as cited above, the principal focus of the EQAO assessment program is for accountability and the explicitly identified purpose of the Grade 9 assessment is to monitor student achievement of the curriculum from year-to-year. The second reason I identify accountability as the intended purpose is that many of the changes that have been made to the Grade 9 assessment since it was first administered in 2001 support the use of the assessment for accountability purposes rather than the use of the assessment for any of the other additional purposes. For example, on the individual reports provided to students, EQAO previously reported Grade 9 mathematics scores by strand. However, at the present time students receive one overall score only. As indicated in Wolfe et al. (2004), more detailed individual reporting supports the use of the assessment as an indication of individual students' understanding of mathematics but providing such detailed information requires more items than can be answered in a reasonable amount of time. Further, as explained by one of the EQAO personnel I interviewed, EQAO's decision to report only an overall score provides less detailed information for individual students but provides an adequate level of detail for accountability purposes. Thus, the change in scores that are reported to individual

students supports the use of the assessment for accountability at the expense of the use of the assessment for individual student reporting.

Similarly, while investigations and extended multi-part problems arguably provide more information about individual students' mathematics learning, these item types are no longer included on the EQAO Grade 9 assessment and, as noted in chapter 4, the proportion of multiple-choice items has increased. Several comments made by two of the EQAO personnel I interviewed indicate that the changes in item type were made to increase the reliability and utility of the assessment for accountability purposes rather than to support other purposes associated with the assessment. The following interchange between Carol and Susan took place as Susan was explaining the history of the EQAO Grade 9 assessment and is an example of the comments made by EQAO personnel that express the view that changes to the assessment have focused principally on the use of the assessment for accountability purposes.

Susan: At that time I was working with some people from the field who were innovative math leaders in the province and that's where the investigations came from.

Carol: Susan is still mourning the passing of the investigations.

Susan: They [the investigations] were difficult to develop ... they were expensive for scoring and they demanded technology for every kid and you know, on and on and on. So, we field tested a number of them the first year, as you may remember, and even though we never used them we posted them because we thought they were excellent classroom assessment stuff.

(Carol & Susan interview, p.5-6)

Susan's comments in this interchange and elsewhere in the interview reveal that she sees the value of the investigations and extended multi-part problems which were included in the first few years of the assessment for assessing students' mathematical

understanding but recognizes that these item types are not tenable in the use of the assessment for accountability purposes.

Identifying Multiple-uses of the EQAO Grade 9 Assessment

Having provided an argument for identifying accountability as the intended use for this assessment, I consider the other benefits or uses claimed by EQAO as well as the use by teachers as part of students' grades to be additional multiple-uses for this assessment. In addition to these multiple-uses, as mentioned in chapter 2, the Fraser Institute uses the results from the EQAO Grade 9 assessment in their calculations for ranking high schools in Ontario (Cowley & Easton, 2008b). Thus, a number of multiple-uses of the EQAO Grade 9 Assessment of Mathematics are taking place with some suggested by EQAO and others involving individuals or groups outside of the assessment development agency. A visual representation of five multiple-uses of the EQAO Grade 9 assessment is shown in Figure 7 using the same conventions which were used in the figures in chapter 2.

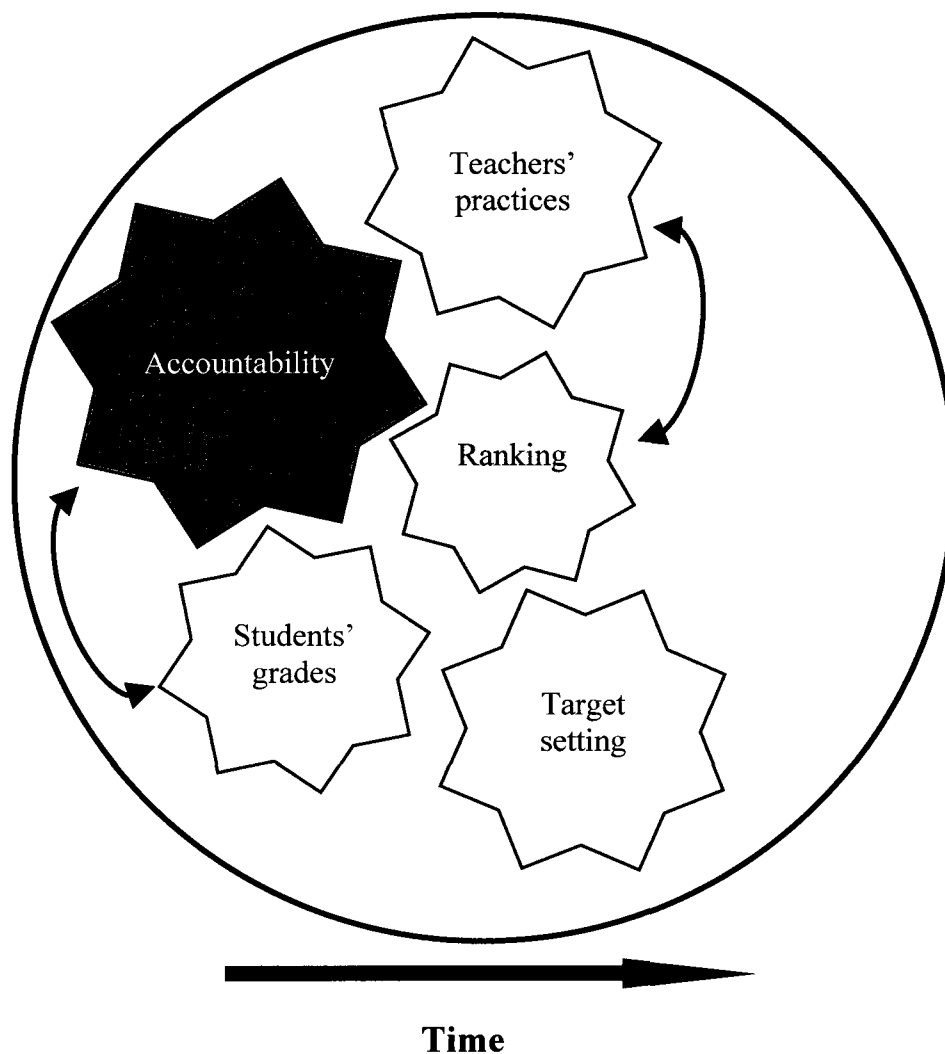


Figure 7. A visual representation of five multiple-uses of the EQAO Grade 9 Assessment of Mathematics.

Each of the uses for the EQAO Grade 9 assessment shown in Figure 7 relate to a single administration of the assessment and this single administration is represented as the large outer circle in the diagram. The intended use is shown as the largest gear inside the circle and is identified as accountability. Four additional uses are shown as smaller gears: ranking (use by the Fraser Institute for ranking schools), students' grades (use by teachers to contribute to students' grades), teachers' practices (benefit claimed by EQAO that the

assessment will improve teachers' assessment practices), and target setting (benefit claimed by EQAO that the assessment provides data for target setting). Because each of these uses involves different groups or individuals, each use has been shaded in a different manner. In order to keep the diagram legible, I have shown a total of five uses for this assessment but others could be added. For example, only two of the benefits that EQAO has listed in the framework document are included in Figure 7 (EQAO, 2009c). Each multiple-use in this diagram is represented as a gear rather than a circle to indicate that the uses cannot be assumed to be independent of one another. Arrows connecting some of the uses are also included to emphasize the potential for interactions among the uses. The relative position of each multiple-use and the interactions that are shown between the uses in this figure are arbitrary. Without conducting empirical studies, it is not clear if and to what extent these multiple-uses might interact with one another. For example, the use of the assessment for ranking might interact more directly with the use for accountability or with the use related to target setting than it does with the use by teachers as part of students' grades.

Focusing on Teachers' Use as Part of Students' Grades

In the remainder of this case study, I focus on two of the multiple-uses that are identified in this diagram; the use by EQAO for accountability purposes and the use by teachers as part of students' grades. I focus on two uses rather than the whole pattern in an effort to better understand the way that these uses may interact. I have chosen to focus on the use for accountability as it is the ostensible intended use for this assessment and to focus on the use by teachers as part of students' grades as it was this practice that inspired this inquiry. In addition, the potential for interactions between teachers' use of the assessment as part of students' grades and the use for accountability is considerable. As suggested in

Figure 7, many complex interactions may take place between the other uses of the Grade 9 assessment. These interactions are not explored in this inquiry but could be the focus of future research.

Multiple scoring of the EQAO Grade 9 assessment.

The multiple-use of the EQAO Grade 9 assessment includes the practice of multiple scoring, as described in chapter 2. In particular, in using the assessment as part of students' grades, teachers do not use the official assessment results provided by EQAO, rather, they select and mark some or all of the items on the assessment. This means that students' responses to at least some of the items on the EQAO Grade 9 assessment are being scored twice. The first scoring is conducted by teachers and takes place before the assessments are returned to EQAO. I refer to these scores as the teacher-derived scores. The second scoring is conducted by EQAO and results in what I refer to as the official scores. The scoring procedures that are developed and carried out by individual classroom teachers are according to these teachers' own criteria whereas the official EQAO scoring uses a scoring rubric and anchor responses developed by EQAO and carried out by a group of trained teachers (as described in chapter 4). Thus, while the multiple-use of the EQAO Grade 9 assessment includes the practice of multiple scoring, this instance of multiple scoring differs from the multiple scoring practices which are described in chapter 2 in that in those instances of multiple scoring each scoring procedure was overseen by the assessment developers whereas for the EQAO Grade 9 assessment different scoring criteria are used for the teacher-derived scores than are used for the official scores and the assessment developers have authority over only one of the scoring events. I return to the significance of this characteristic towards the end of this chapter when evidence of the interactions between the two uses is presented.

EQAO'S current policy regarding teachers' use of the assessment for grades.

In my document analysis, I located EQAO's current policy regarding teachers' use of the Grade 9 assessment as part of students' grades. The administration manual for the assessment indicates that teachers may select some or all of the items on the assessment, mark students' responses to these items and include the results as part of their students' course grade (EQAO, 2007). According to this manual, the decision to use the assessment in this manner is at the discretion of individual teachers, principals and school boards. EQAO specifies that in their use of the assessment items as part of students' grades teachers must conform to the following conditions:

- All or some components of the assessment may be marked on completion of the assessment and prior to the return of materials to EQAO.
- Marks may be used for the course only according to Ministry of Education guidelines.
- The marking methodology is a school or board decision.
- Marks **must not** be made on student work.
- Student work **must not** be copied.
- Use the package of materials marked 'teacher' for in-school marking. This package must be returned to EQAO along with all other assessment materials.
- Any materials generated during marking that include references to the content of the questions must also be returned to EQAO.

(EQAO, 2007, p.12, emphasis in original)

I include the details of the EQAO policy because they relate to many of the observations reported later in this chapter. The explicit statement that the marking methodology is a school or district decision is of particular importance for this study. As will be shown, the fact that EQAO does not provide item-specific scoring guides for the open-response items or marking keys for the multiple-choice items has a considerable impact on the practices that classroom teachers use as they mark students' responses to the items. It is also worth noting

that, given the EQAO policy and current Ministry of Education assessment guidelines, the use of EQAO assessment items as part of students' grades could be as part of a student's 70% term mark or as part of their 30% summative mark.

An additional factor regarding teachers' use of the assessment in grades that I want to highlight is that teachers in other grades are not permitted to mark items from the other EQAO assessments or to include results from these assessments as part of their students' grades. That is, EQAO does not permit teachers to mark items from the Grade 3 and 6 assessments of reading, writing and mathematics or from the Grade 10 literacy assessment. I return to this point later in this chapter in my discussion of observations from the interviews with EQAO personnel.

In this section of this chapter I have addressed the first research question, *What is the pattern of uses associated with the EQAO Grade 9 Assessment of Mathematics?* I presented the pattern of uses for the EQAO Grade 9 assessment based on my analysis of EQAO documents and the interviews I conducted. The pattern of uses clearly indicates that multiple-uses of this assessment are taking place and that a form of multiple scoring also occurs. In this section I also summarized EQAO's current policies regarding the use of the assessment by teachers as part of their students' grades. I now present my observations of how teachers' use of the EQAO Grade 9 assessment as part of students' grades takes place.

Research Question 2 – Describing Teachers' Use

How does teachers' use of the EQAO Grade 9 Assessment of Mathematics as part of students' grades take place?

To address this question I draw on data from the province-wide questionnaire as well as information I collected in the school-level case studies. I begin by considering the

prevalence of this practice across the province and then I provide a description of the procedures and perspectives of teachers with respect to including the EQAO Grade 9 assessment in their students' grades.

Prevalence of Teachers' Use of the Assessment across Ontario

Teachers' responses to one selected-response item on the CIIM questionnaire provide an indication of the prevalence of teachers' use of the EQAO Grade 9 assessment across the province. In this item teachers were asked "In this class to what extent do you use EQAO assessment results in mathematics to form part of students' final marks" and could respond by selecting 'Not at all', 'A little', 'Some' or 'A lot'. Responses to this item are shown in Table 7.

Table 7: Teachers' Reported use of the EQAO Grade 9 Assessment of Mathematics as Part of Students' Grades

	Not at all	A little	Some	A lot
Academic (n=150)	8%	37%	39%	16%
Applied (n=122)	10%	35%	40%	16%

Table 7 indicates that the majority of teachers describing their experiences with Grade 9 Academic and Applied mathematics courses report using the assessment as some part of their students' grades. Specifically, 92% of teachers describing their experiences with Grade 9 Academic and 90% of teachers describing their experiences with Grade 9 Applied mathematics courses indicate that they use the assessment as some part of their students' grades. Given the similarity in the pattern of results for the Academic and Applied courses, these results can be summarized by stating that 91% of Grade 9 teachers who participated in this questionnaire include the EQAO assessment as part of students' grades. The findings from the CIIM questionnaire are similar to those observed by Kitto (2006) and Pang (2004),

as described in chapter 4. In Kitto's study, 85% of teachers indicated that they use the EQAO Grade 9 assessment as part of their students' grades and in the internal study conducted by EQAO, Pang found that 88% of teachers indicated that they had told Grade 9 Applied students that the assessment would be used as part of their grades.

An additional analysis was conducted with the responses to this item to get a sense of how widespread the use of the assessment as part of students' grades is across school districts in Ontario. The final item on the questionnaire asked teachers to indicate their school district so that I was able to compute a cross-tabulation of responses to the EQAO assessment item by district. As noted, the CIIM questionnaire included responses from 42 of the 60 English-language school districts in Ontario. Of the 42 responding districts, answers to this item were received from Grade 9 teachers in 39 different districts. In all but one of these 39 districts (i.e. 97% of districts that responded) at least some teachers indicated that they use the assessment as part of their students' grades and in 27 of these 39 districts (i.e. 69% of responding districts) every Grade 9 teacher who completed the item indicated that they use the assessment as part of their students' grades. However, in 12 of the 39 districts (i.e. 31% of responding districts) at least one teacher indicated that they do not use the assessment as part of their students' grades while other teachers in the same district indicated that they do use the assessment as part of their students' grades. This pattern of responses suggests that in these 12 districts individual teachers or schools decide whether or not they will use the EQAO Grade 9 assessment as part of students' grades rather than there being a district-wide policy.

While this questionnaire item provides a sense of the extent to which the EQAO assessment is used by teachers as part of students' grades across Ontario, these observations

do not provide any details about how that use takes place or about teachers' views regarding this practice. Information about these aspects of teachers' use of the EQAO Grade 9 assessment is presented in the next section by drawing on the three school-level case studies.

A Description of Teachers' Use of the Grade 9 Assessment as Part of Students' Grades

The data from the school-level case studies are presented according to the analysis process described in chapter 5. I begin by summarizing the data for each school which is part of the contextualization of the practices surrounding teachers' use of the assessment in each school. I include information about the size and geographic location of the school and district, school and district results on the EQAO Grade 9 assessment for the five years prior to the study, as well as information that I was able to gather about the approach each district has adopted regarding teachers' use of the EQAO Grade 9 assessment as part of students' grades. In reporting this information I have taken steps to protect the identity of individuals, schools and districts. For example, I have provided approximate values for school and district sizes, used general terms to describe geographic locations and created pseudonyms for schools and participants.

After presenting some information about each school, I describe my observations for the six aspects of teachers' use of the assessment which are identified in chapter 5. Three of these aspects relate to the procedures participants use as they mark the EQAO assessment items (i.e. weighting, selection of items, and scoring procedures) and three of these aspects relate to participants' views regarding this use of the assessment (i.e. meaning ascribed to the scores, rationale and benefits of using the assessment as part of grades and tensions emerging from this use of the assessment). As will be shown, summarizing these aspects of

teachers' use of the EQAO Grade 9 assessment as part of students' grades provides a detailed account of this practice.

Contextualization of Participating Schools

Birch Park Secondary School's geographic location and size.

Birch Park Secondary School is located in a small, predominantly rural school district which has fewer than 5 high schools. The school has approximately 1100 students in Grades 7-12 and is situated in a middle- to lower-income residential neighbourhood in a city with a population of about 15,000 people. In addition to students from the neighbourhood where the school is located, a large number of students attending Birch Park Secondary School are bussed in from rural areas surrounding the city.

Interviews conducted at Birch Park Secondary School.

The interviews I conducted at Birch Park Secondary School are listed in Table 8. Each participant's pseudonym, position and teaching assignment at the time of the interview are indicated along with a summary of the information they provided about their teaching experience.

Table 8: Participants at Birch Park Secondary School

Pseudonym	Position	Math Teaching Assignment	Teaching Experience
Allen	Principal	None	11 years as principal with 7 years at Birch Park and 4 years at another high school in the same board - several years teaching experience prior to becoming principal
Francis	Mathematics Department Head & Teacher	One class Grade 9 Applied math; 2 classes Grade 12 University math	9 years teaching math including 3 years as department head at Birch Park

Pseudonym	Position	Math Teaching Assignment	Teaching Experience
			• first time teaching Grade 9 Applied math, has taught Grade 9 Academic math in previous years • no indication of experience with EQAO item writing or scoring
William	Teacher	• One class Grade 9 Academic math; one class Grade 12 Workplace math	• 4 years teaching mostly in economics and history • first year teaching math
Margot	Teacher	• One class Grade 9 Applied math; 2 other math courses	• 5 years teaching math
Shannon	Teacher	• One class Grade 9 Academic math	• approximately 5 years teaching Grades 7 & 8 • several years teaching high school, mostly science & physical education • taught Grade 10 Applied math in previous year • first time teaching Grade 9 math

Characteristics and concerns of Birch Park Secondary School.

In his interview, the principal, Allen, indicated that a relatively large proportion of the students attending Birch Park come from families with low socioeconomic status (Allen interview, p. 5). Allen also commented that a major focus in the school is to support students who are not succeeding in their courses. He stated “it’s how to meet the challenges of those students who are not succeeding . . . there are some students who are falling between the cracks and there are students who need more support and how do we, within the capacity that we have, how do we build that in?” (Allen interview, p.5). Allen indicated that while

students in the Grade 9 Academic mathematics courses at Birch Park were doing well, students in the Grade 9 Applied mathematics courses were struggling and in need of greater support (Allen interview, p.4). He also noted that the teachers at Birch Park are relatively new to teaching with approximately 30% having less than five years teaching experience. He expressed the view that many of these teachers are still developing their teaching skills and are more focused on issues such as time management and acquiring useful teaching resources than on refining their approach to assessment (Allen interview, p.9).

Background of Grade 9 mathematics teachers at Birch Park.

At Birch Park Secondary School I interviewed the mathematics department head as well as three teachers in the department. All four of these individuals were teaching Grade 9 mathematics courses during the semester when the interviews took place. Two of the teachers I interviewed, William and Shannon, indicated that their subject specialty was not mathematics, that they were teaching Grade 9 mathematics for the first time and that this was their first time preparing to use the EQAO assessment as part of their students' grades (see Table 8). The other two participants indicated that their subject specialty was mathematics and that they had been teaching Grade 9 mathematics and using the assessment as part of students' grades for several years. The two experienced mathematics teachers were teaching Grade 9 Applied mathematics while the two less experienced mathematics teachers were teaching Grade 9 Academic courses in the semester when the interviews were conducted.

EQAO Grade 9 assessment results for Birch Park.

Results on the EQAO Grade 9 Assessment of Mathematics for the school, district (board), and province are shown in Figures 8 and 9 for the five years prior to the year of this study. Each graph indicates the percentage of students who achieved at or above the

provincial standard of Level 3 on the assessment. Figure 8 shows the results for students who completed the *Grade 9 Applied* assessment and Figure 9 shows the results for students who completed the *Grade 9 Academic* assessment. The results in these graphs were taken from the EQAO school report for Birch Park Secondary School and from the report for the school district where Birch Park is located.

One trend which can be seen in these graphs is that the number of students achieving Level 3 is considerably lower for the *Grade 9 Applied* courses than it is for the *Grade 9 Academic* courses for the school, district and province. Another observation which can be made from these graphs is that the EQAO *Grade 9* assessment scores for the school district where Birch Park is located are consistently higher than the provincial results for both the *Academic* and *Applied* assessments. In terms of results over time, Birch Park's scores on the *Applied* and *Academic* assessments seem to vary widely from year to year but the extent of variation is somewhat less for the *Academic* assessment than it is for the *Applied* assessment.

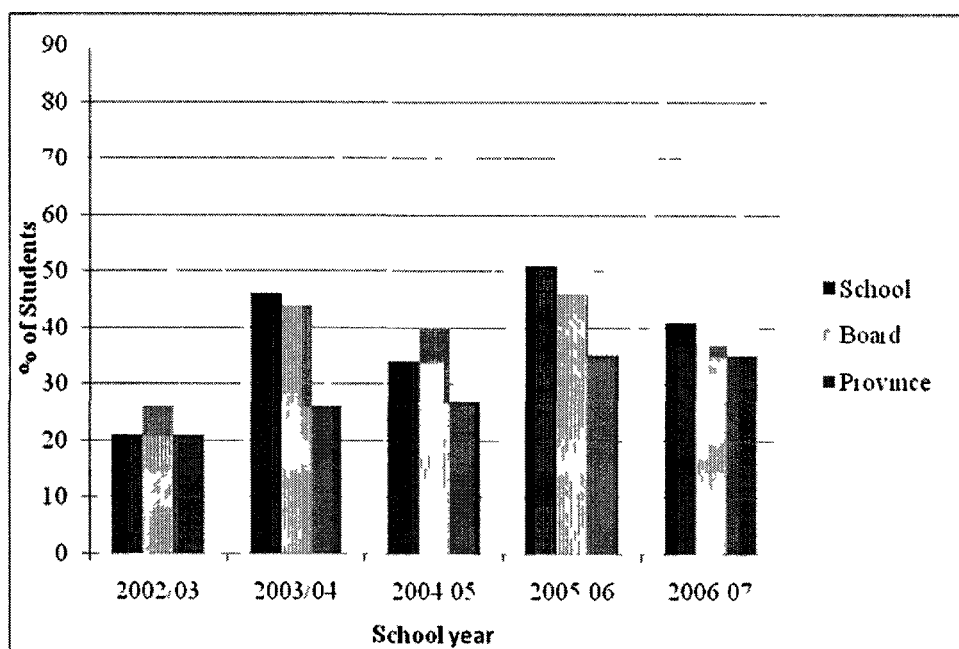


Figure 8. Percentage of students at or above Level 3 on the EQAO Grade 9 Applied Assessment of Mathematics by school, board and province for Birch Park.²⁷

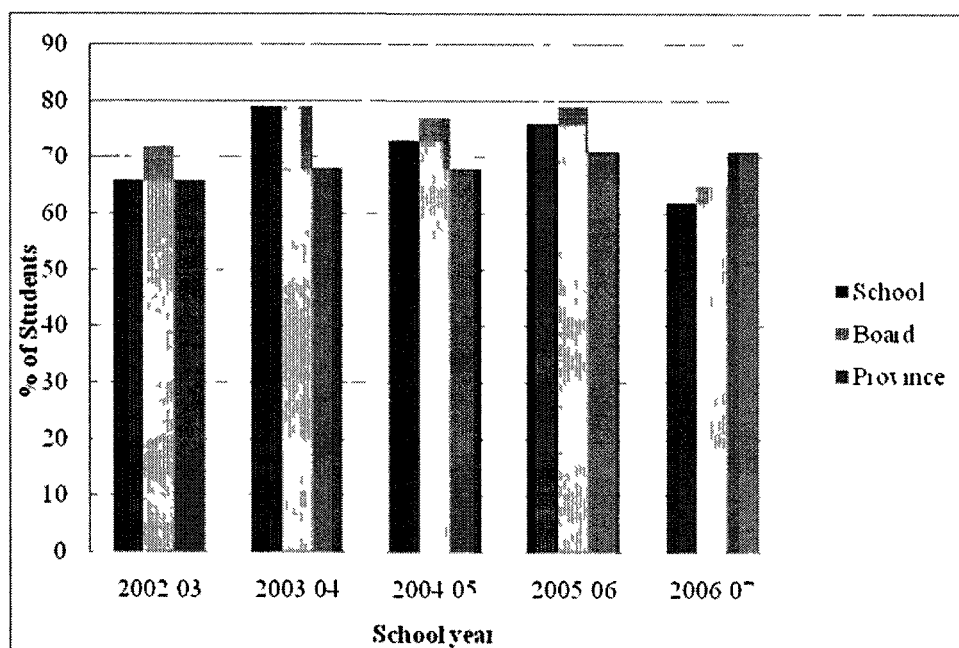


Figure 9. Percentage of students at or above Level 3 on the EQAO Grade 9 Academic Assessment of Mathematics by school, board and province for Birch Park.

²⁷ EQAO advises that changes in results from 2004/05 to 2005/06 must be interpreted in the context of significant revisions made to the Grade 9 Applied mathematics curriculum in 2005

School district approach to using the EQAO assessment for grades.

I was unable to find any mention of the use of the EQAO Grade 9 assessment as part of students' grades on the website for the school district where Birch Park is located. In fact, I was unable to locate any policies or procedures on the district's website related to student assessment or evaluation. However, I did get some sense of the approach this district has taken toward this use of the EQAO assessment from the comments made by the principal. Allen has been a high school principal in this district for the past 11 years which includes the entire period that the EQAO Grade 9 assessment has been administered. In his interview Allen indicated that when the EQAO Grade 9 assessment was first administered in 2001 teachers across the district were distrustful of the assessment and tended to see it as serving the purposes of the Ministry of Education. He states:

I would say in the early years that they [teachers] did not use it because there was still that, I guess, mistrust of EQAO and some angst about teachers saying this is a government imposed program and not necessarily a teacher-friendly government imposed program. So, there was a little bit of resistance. They certainly did the program but they didn't necessarily use it [as part of their students' grades]. (Allen interview, p. 2).

However, Allen indicated that after a few years as teachers at Birch Park became more familiar with the EQAO Grade 9 assessment they began to use it as part of their students' grades. Allen indicated that teachers at Birch Park have been using the assessment as part of students' grades for about five years (Allen interview, p.3). However, he also noted that the district does not have a policy concerning the use of the EQAO Grade 9 assessment as part of students' grades and each high school decides whether or not they would like to use the assessment for this purpose (Allen interview, p.5).

Scottsdale Collegiate's geographic location and size.

Scottsdale Collegiate is located in a mid-sized, predominantly urban and suburban school district which has approximately 20 high schools. Scottsdale Collegiate has approximately 1000 students in Grades 9-12 and is situated in a suburban, middle-income neighbourhood. According to the principal, Michael, the majority of students at Scottsdale Collegiate reside in the neighbourhood. He indicated that the level of parental involvement at the school is quite high and described the parents in the school community as fairly well educated (Michael, interview, p. 7). Michael also indicated that the majority of students at Scottsdale come from a single elementary school which is located nearby.

Interviews conducted at Scottsdale Collegiate.

The interviews I conducted at Scottsdale Collegiate are listed in Table 9. Each participant's pseudonym, position and teaching assignment at the time of the interview are indicated in this table along with a summary of the information they provided about their teaching experience.

Table 9: Participants at Scottsdale Collegiate Institute

Pseudonym	Position	Math Teaching Assignment	Teaching Experience
Michael	Principal	• None	• 10 years experience as principal including years as principal at Scottsdale • 20 years teaching experience prior to becoming a principal mostly in business, accounting and law

Pseudonym	Position	Math Teaching Assignment	Teaching Experience
Mark	Mathematics Department Head & Teacher	Grade 12 math courses	- approximately 25 years experience teaching math including 9 years experience as department head - several years experience with Grade 9 EQAO item writing and reviewing
Whitney	Teacher	One class Grade 9 Academic math; one class Grade 9 Applied math; one class Grade 10 Applied math	3 years teaching experience - math and science teaching qualifications - second year teaching at Scottsdale
Leanne	Teacher	One class Grade 9 Academic math	2 years teaching at Scottsdale - several years teaching math and science at previous schools

Characteristics and concerns of Scottsdale Collegiate.

In his interview the principal, Michael, indicated that assessment is an area which both Scottsdale Collegiate and the school district have been focusing on and he described some professional development initiatives that had recently been held in the school and the district on this topic (Michael interview, p.3). Michael noted that a new assessment and evaluation policy was due to be released by the school district and that Scottsdale Collegiate had already begun to implement some aspects of this policy. In addition, both Michael and the mathematics department head, Mark, emphasized that Scottsdale Collegiate has a strong focus on supporting new teachers and on encouraging collaboration and consistency among the teachers in the department (Michael interview, p. 3; Mark interview, p.2). Michael noted

that one challenge which he faces in achieving a high degree of consistency among teachers at Scottsdale Collegiate is the turnover in teaching staff which occurs as experienced teachers seek opportunities for leadership roles in other schools and new teachers are hired to fill their positions (Michael interview, p. 6).

Background of Grade 9 mathematics teachers at Scottsdale Collegiate.

At Scottsdale Collegiate I gathered information about teachers' use of the EQAO Grade 9 assessment from the mathematics department head, Mark, as well as two Grade 9 mathematics teachers, Whitney and Leanne. During the semester when I conducted the interviews, Mark was teaching Grade 12 mathematics but was not teaching any Grade 9 mathematics courses. However, he indicated that he has taught both Academic and Applied mathematics at the Grade 9 level many times in the past and that he has used the EQAO Grade 9 assessment as part of students' grades in previous years. He also indicated that he has had experience with writing and reviewing items for the EQAO Grade 9 assessment. Both Whitney and Leanne have their mathematics teaching qualifications and both were relatively new to Scottsdale Collegiate at the time of the study. However, Leanne had several years of teaching experience prior to coming to Scottsdale, whereas Whitney had just finished her third year of teaching at the time of the interview. There were two other teachers teaching Grade 9 mathematics courses at Scottsdale Collegiate who were not available to participate in this study.

EQAO Grade 9 assessment results for Scottsdale Collegiate.

Results on the EQAO Grade 9 Assessment of Mathematics for the school, district (board) and province are shown in Figures 10 and 11 for the five school years prior to the year of my study. Each graph indicates the percentage of students who achieved at or above

the provincial standard of Level 3 on the assessment. Figure 10 shows the results for students who completed the Grade 9 Applied assessment and Figure 11 shows the results for students who completed the Grade 9 Academic assessment. The results in these graphs were taken from the EQAO school report for Scottsdale Collegiate and from the report for the school district where Scottsdale is located. As noted for Birch Park Secondary School, one trend which can be seen in these graphs is that the number of students achieving Level 3 is considerably lower for the Grade 9 Applied courses than it is for the Grade 9 Academic courses for the school, district and province. These graphs also show that the results at Scottsdale Collegiate in both Academic and Applied courses are consistently higher than the results for the school district and the province. Given that the results for this school district are also frequently higher than the provincial results for both the Academic and Applied assessments, these results suggest that the performance of students at Scottsdale Collegiate on the EQAO Grade 9 assessments are among the highest of schools across the province.

In terms of trends in EQAO assessment results over time, Figure 10 shows that while the percentage of students at Scottsdale Collegiate achieving at or above the provincial standard on the Grade 9 Applied assessment is quite high compared to the district and province, it has been steadily decreasing each year. That is, students in the Grade 9 Applied courses over the past five school years are scoring less well on the EQAO Grade 9 assessment relative to students in the district and province. In contrast, as shown in Figure 11, the percentage of students achieving Level 3 on the EQAO Grade 9 Academic assessment has been relatively stable or shown a modest gain over these five school years.

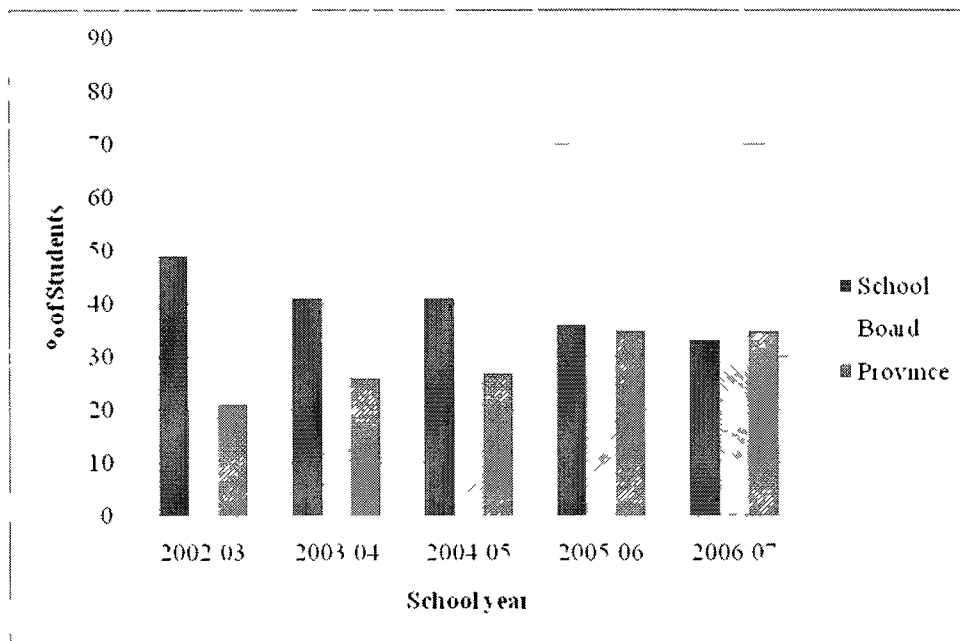


Figure 10. Percentage of students at or above Level 3 on the EQAO Grade 9 Applied Assessment of Mathematics by school, board and province for Scottsdale Collegiate.²⁸

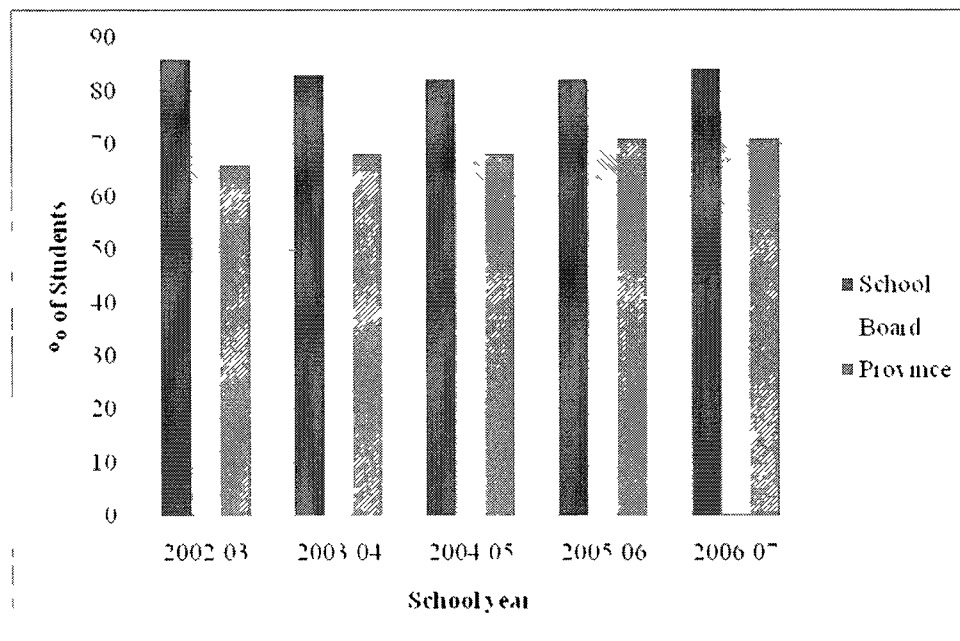


Figure 11. Percentage of students at or above Level 3 on the EQAO Grade 9 Academic Assessment of Mathematics by school, board and province for Scottsdale Collegiate.

²⁸ EQAO advises that changes in results from 2004/05 to 2005/06 must be interpreted in the context of significant revisions made to the Grade 9 Applied mathematics curriculum in 2005

School district approach to using the EQAO assessment for grades.

The website for the school district where Scottsdale Collegiate is located includes a number of policy documents related to various aspects of assessment and evaluation. While teachers' use of the EQAO Grade 9 assessment as part of students' grades is not explicitly mentioned, this district does have a policy related to the assessment tasks which teachers include in the summative portion of their students' grades. The policy states that high school teachers must ensure that the 30% summative portion of students' grades is based on at least two different assessment tasks. The policy also specifies that one of these tasks may be a written examination but the written examination cannot count for more than 20% of a students' final grade.

I gathered some information about the approach this school district has taken toward the use of the EQAO Grade 9 assessment as part of students' grades in my interview with the mathematics department head. Mark has about 25 years experience in this school district and has been a mathematics department head for the entire time that the EQAO Grade 9 assessment has been administered. He indicated that when the assessment was first administered in 2001, the teachers in the district were not inclined to use it as part of students' grades but they gradually began to do so. He states:

I think at the beginning there was no intent for it to, well I guess there was an uncertainty about what EQAO was going to look like and there was trepidation about how the kids were going to perform on EQAO and so no one, there was no requirement by anyone that it should be included in the marks and I think at the beginning the intent certainly was that it not be included in the marks but then that evolved to a more open-ended thing. (Mark interview, p. 3)

Mark indicated that most schools in this district are currently using the assessment as part of students' grades but that there is not a consistent pattern. He also said that he knew some schools within this district with a large proportion of English as a Second Language (ESL)

learners that do not use the assessment as part of their students' grades. This pattern of a gradual increase in the use of the assessment as part of students' grades over the years since the assessment was first introduced is similar to the one described by the principal at Birch Park Secondary School.

Taylor Brook High School's geographic location and size.

Taylor Brook High School is located in a large, predominantly urban and suburban school district that has approximately 30 high schools. Taylor Brook has about 1300 students in Grades 9 to 12 and is situated in a suburban, middle-income neighbourhood. While some students live in the neighbourhood in which the school is situated, many are bussed in from surrounding neighbourhoods. According to the mathematics department head, students at Taylor Brook come from five elementary schools located in the neighbourhoods adjacent to the high school (Diane interview, p.10).

Interviews conducted at Taylor Brook High School.

The interviews I conducted at Taylor Brook High School are listed in Table 10. Each participant's pseudonym, position and teaching assignment at the time of the interview are indicated in this table along with a summary of the information they provided about their teaching experience.

Table 10: Interviews Conducted at Taylor Brook High School

Pseudonym	Position	Math Teaching Assignment	Teaching Experience
Brian	Principal	None	- More than 20 years teaching - approximately 10 years as a principal including 4 years as principal at Taylor Brook

Pseudonym	Position	Math Teaching Assignment	Teaching Experience
Diane	Mathematics Department Head & Teacher	Two classes Grade 9 Applied math; one class Grade 12 Data Management	17 years teaching experience in math - approximately 6 years as department head - several years experience with Grade 9 EQAO item writing and reviewing
Adrian	Teacher	- One class of Grade 9 Academic math this semester - One class of Grade 9 Academic math last semester	- two years teaching experience - all teaching experience in math and at Taylor Brook
Yvette	Teacher	- One class Grade 9 Academic math this semester - Two classes Grade 9 Academic math last semester	- third year teaching - first year at Taylor Brook - teaches physics mostly with some math
Owen	Teacher	- One class Academic Grade 9 math this semester; two classes Grade 12 Calculus this semester - two classes Grade 12 Functions and one class Grade 9 Academic last semester	15 years teaching experience - all teaching experience at Taylor Brook - teaches mostly math and a little science

Characteristics and concerns of Taylor Brook High School.

In his interview, the principal, Brian, noted that while the school has historically been primarily an academic school, in recent years an increasing number of students who are at-risk in terms of their social, academic and economic situations have begun to arrive at Taylor Brook (Brian interview, p. 2). There has also been an increase in the proportion of ESL

students at Taylor Brook who require additional support. Brian indicated that a major focus in the school is on motivating and supporting students in Applied level courses including mathematics (Brian interview, p.3). He described a number of strategies related to developing students' timetables, the use of non-common in-school examinations, and the careful tracking of students who are at-risk which are being used at Taylor Brook in an effort to support these students.

Background of Grade 9 mathematics teachers at Taylor Brook.

At Taylor Brook High School I interviewed the mathematics department head, Diane, who was teaching two classes of Grade 9 Applied mathematics during the semester when the study was conducted. I also interviewed three teachers who were teaching Grade 9 Academic mathematics courses that semester. Diane has 17 years of experience in teaching math including six years as a mathematics department head. She has also had experience in item writing, reviewing items and selecting anchor responses for open-response items on the EQAO Grade 9 assessment. Two of the other teachers I interviewed, Adrian and Yvette are qualified mathematics teachers but were relatively new to teaching with two and three years experience respectively. The other teacher I interviewed, Owen, is an experienced mathematics teacher who has taught Grade 9 Academic and Applied courses for several years as well as mathematics courses at each of the other grade levels. During the year when I did this study, there were 10 classes of Grade 9 Academic mathematics and 3 classes of Grade 9 Applied mathematics taught at Taylor Brook (Diane interview, p. 3).

EQAO Grade 9 assessment results for Taylor Brook.

Results on the EQAO Grade 9 Assessment of Mathematics for the school, district (board) and province are shown in Figures 12 and 13 for the five school years prior to the of

the study. Each graph indicates the percentage of students who achieved at or above the provincial standard of Level 3 on the assessment. Figure 12 shows the results for students who completed the Grade 9 Applied assessment and Figure 13 shows the results for students who completed the Grade 9 Academic assessment. These results were taken from the EQAO school report for Taylor Brook High School and from the report for the district where the school is located. The graphs show that the percentage of students at Taylor Brook High School at or above Level 3 has been consistently much higher for students taking the Grade 9 Academic assessment than for students taking the Grade 9 Applied assessment. While this observation was also made for the other schools in this study, the difference between results for the Academic and Applied assessments is greater at Taylor Brook than the difference that was observed for either Birch Park or Scottsdale Collegiate. Figure 13 also shows that results for students taking the Grade 9 Academic assessment are comparable or somewhat higher than the results for the district and for the province. However, as shown in Figure 12, results for students taking the Grade 9 Applied assessment are consistently considerably lower at Taylor Brook than those of the district and province.

With regard to Taylor Brook's Grade 9 EQAO scores, both the principal and the mathematics department head commented in their interviews that while the scores for the Academic courses have been reasonably good in recent years, the scores for the Applied courses are among the lowest in the district. In his interview the principal referred to the Applied results as a "shattering statistic" (Brian interview, p.2) and both Brian and Diane indicated that improving EQAO scores for Applied students is a major focus for this mathematics department (Brian interview, p.3; Diane interview, p.4).

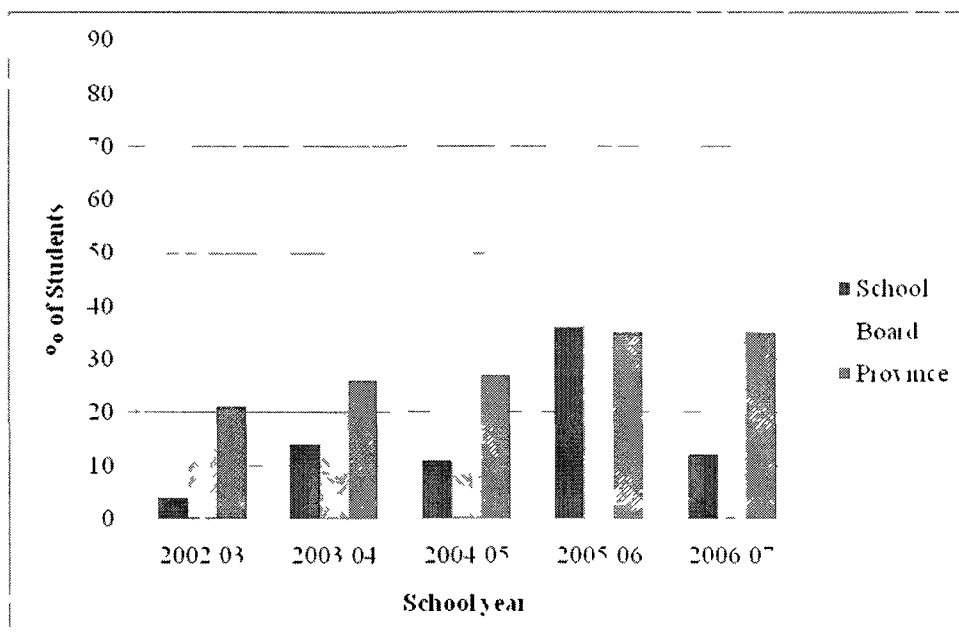


Figure 12. Percentage of students at or above Level 3 on the EQAO Grade 9 Applied Assessment of Mathematics by school, board and province for Taylor Brook ²⁹

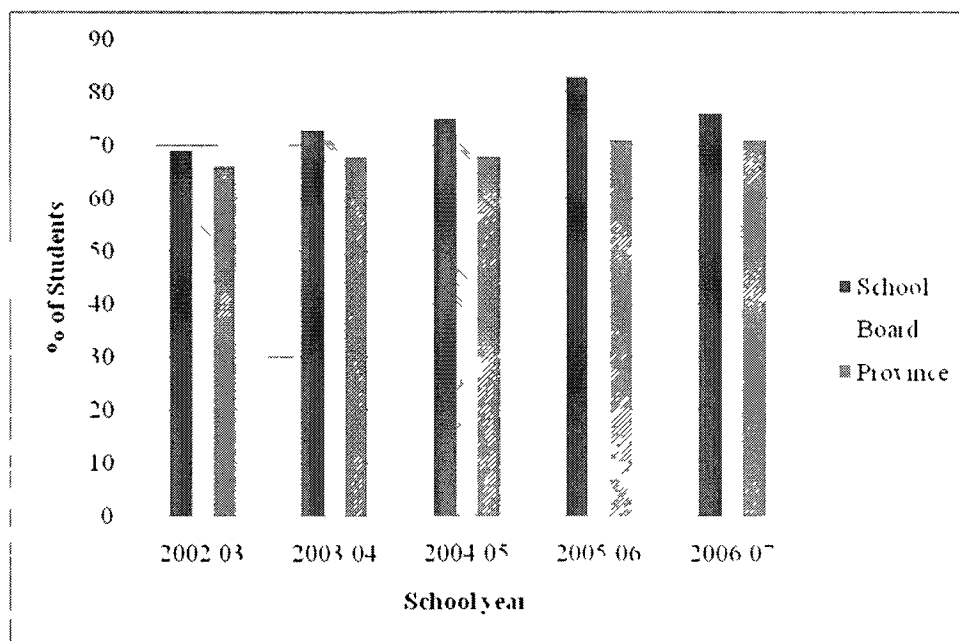


Figure 13. Percentage of students at or above Level 3 on the EQAO Grade 9 Academic Assessment of Mathematics by school, board and province for Taylor Brook

²⁹ EQAO advises that changes in results from 2004/05 to 2005/06 must be interpreted in the context of significant revisions made to the Grade 9 Applied mathematics curriculum in 2005

School district approach to using the EQAO assessment for grades.

With regard to policies for the school district where Taylor Brook High School is located, I was unable to locate any policies on the district website that refer to teachers' use of the EQAO Grade 9 assessment as part of students' grades. While I did find assessment and evaluation policies that identify the purpose and qualities that assessments should have in this school board, I was unable to locate specific details such as the number of assessment tasks that must be used in summative grades. Given the size of this school board and comments made by participants in the interviews I conducted, I expect that these policies exist but are not posted on the board's website.

However, I was able to get a clear sense of the approach this district has taken toward the use of the EQAO Grade 9 assessment as part of students' grades from my interview with the mathematics department head at Taylor Brook who has 17 years experience teaching mathematics in this district. In her interview, Diane indicated that for the first few years when the EQAO Grade 9 assessment was administered this district had a policy which prohibited teachers from using the Grade 9 assessment as part of students' grades. According to Diane, the district did not consider using the EQAO Grade 9 assessment as part of students' grades to be a "valid" assessment practice (Diane interview, p.5). However, a few years before my study took place the school board reversed their position and began to require that teachers include the assessment as 5% of students' grades (Diane interview, p.5). Diane indicated that this decision was made by the district in an attempt to improve EQAO Grade 9 assessment results for schools in the district. She indicated that the use of the assessment as part of students' grades has been discussed at district-wide mathematics department head meetings and that most schools have decided to mark only the multiple-

choice items on the assessment (Diane interview, p.5). Diane also noted that according to district policy, individual schools can decide which items they would like to mark and use as the required 5% component so that some schools may choose to use the open-response items in addition to the multiple-choice items.

An Overview of the Participating Schools

To provide a sense of the range of characteristics of the schools who participated in this study, I offer a brief overview of the information just detailed. One school is located in a small predominantly rural school district, one is located in a mid-sized, urban/suburban district and one is located in a large urban/suburban district. The schools vary in size from approximately 1000 to 1300 students. Each school has distinct student characteristics and concerns they are focused on addressing.

The mathematics department heads and teachers I interviewed at these schools had a range of teaching experience from 2 years to over 25 years. A few teachers were teaching Grade 9 mathematics courses and using the EQAO Grade 9 assessment as part of students' grades for the first time but most had had several years of experience with both these practices. In addition, two of the mathematics department heads had experience with developing and reviewing items on the EQAO Grade 9 assessment. Apart from these two department heads, none of the teachers who were interviewed in this study had participated in the official summer scoring of the EQAO Grade 9 assessments or had been involved in writing or reviewing assessment items for EQAO.

In terms of performance on the EQAO Grade 9 assessment, at one school, Scottsdale Collegiate, students in both the Academic and Applied mathematics courses score consistently well above the school district and provincial averages. However, at the other

two schools performance on the EQAO Grade 9 assessment is comparable to or somewhat lower than that of their districts and the province. In addition, at Taylor Brook High School, scores on the EQAO Applied mathematics assessment have been a serious concern as they have been considerably below those of the district and the province for several years.

With regard to policies at the school district level, evidence from the interviews and websites suggests that the small rural school district has few policies regarding assessment and evaluation while the two larger school boards have established more policies regarding teachers' assessment practices. One of the district schools boards in my case study (the largest board) has a policy requiring that teachers include the EQAO Grade 9 assessment as part of students' grades. In the other two school boards no policy regarding this practice has been established. My observations also suggest that some schools in the two districts with no district-wide policy use the assessment as part of students' grades while other schools in the same district have decided not to take this approach.

I now turn to a description of the procedures teachers at these three schools report using as they mark the EQAO Grade 9 assessment and include these results in their students' grades.

Description of Procedures Teachers Use to Include the Assessment in Students' Grades

I present my observations of the procedures teachers use according to three aspects of this practice: the weighting of the assessment in the final grade, the selection of EQAO Grade 9 assessment items to be scored, and the scoring procedures used by mathematics teachers. For each aspect, I include observations from the three schools as well as those practices that were reported by EQAO personnel in their former capacity as mathematics teachers.

Weighting.

In using the term weighting, I am referring to the amount that teacher-derived scores on the EQAO Grade 9 assessment contribute toward each student's final grade for mathematics. As noted earlier, neither EQAO nor the Ministry of Education has established policies regarding how much the assessment can contribute to students' grades. As will be shown, the information gathered in this study provides evidence that the weighting given to teacher-derived EQAO scores can vary within individual classrooms, across classrooms within a school, across schools in a district and from district to district in the province.

At Birch Park Secondary School each teacher makes their own decisions about the weighting of teacher-derived EQAO scores in their students' grades. The mathematics department head and two of the teachers at Birch Park indicated that they mark some of the items on the EQAO assessment and include the grade as part of their students' 70% term mark. These three teachers indicated that the EQAO assessment is weighted as the equivalent of a small assignment or quiz when it is incorporated into students' term marks. However, another teacher at Birch Park indicated that he uses the EQAO assessment items as part of students' 30% summative grade. Thus, it is evident that there is variation in the amount of weighting and in whether the EQAO assessment is counted toward the term or summative part of students' grades at Birch Park Secondary School.

The weighting given to the EQAO assessment at Birch Park also varies from student to student within some classes and can change based on the quality of the items on the assessment. For instance, William indicated that he would include scores on the EQAO Grade 9 assessment items as part of students' summative grade for most students in his class but that he would not include the assessment items as part of the grades for two students who

have difficulty demonstrating their mathematical understanding on tests (William interview, p.2). Another example of the varied practices in weighting the EQAO assessment that were reported at Birch Park is the approach used by Margot. Margot indicated that in previous years she has planned to use the EQAO assessment items as the equivalent of a quiz to include in students' term marks and has informed her students that the assessment would be counted in their grades before they completed the assessment. However, when she marked the EQAO assessment items she felt that the items were not "a fair assessment" of her students' ability in mathematics and subsequently decided not to include them as part of students' grades (Margot interview, p.2). Margot indicated that she is prepared to change her weighting of the EQAO assessment as part of her students' grades on the basis of her impression of the quality of the items included on the current version of the assessment. Thus, she informs her students that the EQAO assessment will contribute to their grades and then makes a final decision as to whether she will actually include it in their grades after she has marked the assessments.

Among the three schools that participated in this study, the teachers at Birch Park placed the least weight on the EQAO Grade 9 assessment in their students' grades. In fact, this is the only school I observed where some teachers include the assessment items as part of students' 70% term mark rather than as part of the 30% summative portion of their grade. It is also the only school of those that I observed where individual teachers within one mathematics department weighted their teacher-derived EQAO results in different ways from class to class. This means that students at Birch Park who take Grade 9 mathematics courses during the same school year experience different weightings related to the EQAO Grade 9 assessment depending on the approach taken by their classroom teacher. It is

interesting to note that Grade 9 mathematics teachers at Birch Park also have a non-common performance task and a non-common in-school exam as part of their students' summative grade. Both the teachers and the department head indicated that they feel it is important for teachers to use their professional judgment to make decisions that are appropriate for the specific students in their own classrooms. For example, with regard to the use of the EQAO assessment as part of students' grades Shannon states:

So, [name of department head] said I completely have control of how I do it and that's one of the good things. He's the type of department head that gives me my professional judgment, my ability to use my professional judgment and I know that if he says consistently that this is what we do and if I have an issue with that I can sit down with him and say these are my issues, this is what I want to do. (Shannon interview, p. 5)

The approach to weighting the EQAO assessment in students' grades is quite different at Scottsdale Collegiate where the department head and all of the teachers include the EQAO assessment as 10% of the summative part of students' grades. At Scottsdale the summative grade for all Grade 9 mathematics courses is based on a common in-school final exam (10%), a common performance task (10%) and the teacher-derived mark from the EQAO Grade 9 assessment (10%). The mathematics department head indicates that consistency in assessment practices from class to class is important at Scottsdale Collegiate. He states "it's got to be a common approach with all the classes...It wouldn't be acceptable in any of our courses to have a different approach, teacher by teacher" (Mark interview, p.2). In addition to commenting on her experiences at Scottsdale, one of the teachers I interviewed at this school noted that at her previous school within the same district the EQAO assessment was used as part of the summative grade in lieu of a final in-school exam. She could not recall whether the EQAO assessment counted for 15% or 20% of students' overall grade (Whitney interview, p.1). In contrast to this practice, the mathematics department head

at Scottsdale indicated that the EQAO assessment was not being used as part of students' grades at all in some schools within the same district (Mark interview, p.4). Thus, my observations from Scottsdale Collegiate indicate a high degree of consistency in the weighting of the EQAO Grade 9 assessment within this school but considerable variation in the weighting given to the assessment across schools in the district where the school is situated. Among the three schools in my study, Scottsdale Collegiate placed the most weight on the EQAO Grade 9 assessment in their use of it as part of students' grades.

At Taylor Brook High School the mathematics department head and all of the teachers who were interviewed use the teacher-derived scores from the EQAO assessment items as 5% of the summative portion of students' grades. The remainder of the summative grade is derived from a non-common in-school exam valued at 20% and a non-common performance assessment valued at 5%. The use of the assessment as 5% of students' summative grade is consistent across the school district where Taylor Brook is located but was a relatively new practice which was established two years prior to the year of this study. Several teachers who were interviewed indicated that prior to using the EQAO assessment as part of the summative grade, the performance assessment was valued at 10% of students' overall grade. Thus, the decision to include the EQAO assessment as part of students' grades reduced the weighting on the performance task by 5%. The weighting given to the non-common in-school exam has stayed at 20% throughout this period. Among the three schools included in my study, Taylor Brook was the only school where the weighting decision was based on district policy.

I learned of another interesting approach to weighting the EQAO Grade 9 assessment from one of the EQAO personnel that I interviewed who had formerly been a Grade 9

mathematics teacher. I include Mitchell's approach in this summary as it differs from the approaches I learned about in the school-level case studies. As part of their summative grade, students in Mitchell's class completed both an in-school examination and the EQAO assessment. Mitchell marked both of these assessments and then included in each student's grade whichever was higher, the teacher-derived EQAO score or the score on the in-school examination. Mitchell explained that he used this approach because he wanted the EQAO assessment to count as part of students' grades so that students would be motivated as they complete the assessment. However, because teachers are not permitted to keep copies of students' responses to the EQAO items there is no way to provide evidence of a student's performance to students or to their parents. In Mitchell's experience, parents sometimes express concern at not being able to see their child's EQAO responses when those responses are being counted in the student's grade. Thus, any parents who were not satisfied with including the EQAO assessment in their students' grades were given the option of having the in-school examination count instead (Mitchell interview, p. 15). Given this approach, in Mitchell's classroom the weighting given to the EQAO assessment items as part of students' grades would vary from student to student within the same class.

The data from this study indicate that the weighting teachers give to the teacher-derived scores on the EQAO Grade 9 assessment as part of students' grades varies considerably. In one school each teacher makes their own decision about weighting, in another school the Grade 9 mathematics teachers come to a consensus to decide on the weighting that will be used and in another school the weighting is decided by the school district. In addition, the weighting varies from some schools not using the assessment as part of students' grades, minimal weighting at Birch Park where the teacher-derived scores are

counted as a small part of the students' term or summative grades, and moderate weighting at Taylor Brook with the assessment counting as 5% of students' overall grade. Greater weighting of the EQAO assessment takes place at Scottsdale Collegiate with 10% of students' overall grade coming from the teacher-derived marking process. The greatest weighting described by participants in this study was the use of the EQAO assessment in lieu of an in-school exam as 15-20% of students' overall grades. Thus, this study provides evidence that different weightings are sometimes used for students within one classroom, from classroom to classroom within one school and from school to school within one school district.

Selection of items to be scored.

The school-level case studies provide evidence that the selection of items on the EQAO Grade 9 assessment that teachers score and include as part of students' grades also varies from school to school. At Birch Park Secondary School each teacher independently selects the items they will mark and include in students' grades. Each teacher at Birch Park indicated that they use a selection of both the multiple-choice and the open-response items. Three of the teachers at Birch Park indicate that, if they have sufficient time, they mark all of the items on the EQAO Grade 9 assessment. However, if they have more than one class of Grade 9 mathematics in a semester or particularly large classes then they select some multiple-choice items and some open-response items to mark and include in the grade. Most of the teachers indicated that they choose items that are representative of what was done in their specific mathematics course. For example, Shannon states "What I plan on doing is taking the EQAO test after they're done writing it and selecting the questions that I feel best represent the course" (Shannon interview, p.1).

At Scottsdale Collegiate the teachers get together to decide which items they will score. According to the teachers I interviewed, they typically mark most of the multiple-choice items and then select a few of the open-response items to score. These teachers indicated that they do not score any EQAO assessment items that are not consistent with what was covered in their classes or that have wording that students might not understand. The department head also indicated that in coming to a consensus on which items to score, the teachers try to make sure that there is not too much overlap between the EQAO assessment items and the items which are on the common in-school exam and the common performance task which are the other two components of students' summative mark in all Grade 9 mathematics courses at Scottsdale (Mark interview, p.2). Whenever possible, face-to-face meetings of Grade 9 mathematics teachers are held in order to make these decisions and at other times the decisions are made through email discussions (Whitney, p.3). Thus, at Scottsdale, both multiple-choice and open-response items are scored, all teachers use the same items from the assessment and the selection of the items is done collaboratively.

At Taylor Brook High School only the multiple-choice items are scored. In addition to making the marking process faster, one teacher, Owen, indicated that marking only the multiple-choice items ensures that each teacher in the department marks the EQAO assessment in a consistent way. Owen expressed some concern that if teachers marked the open-response items in addition to the multiple-choice items, different scoring procedures might occur from one class to another. He states:

I guess I wouldn't mind using them [the open-response items] myself but I guess there would be a certain level of discomfort with realizing that I might be marking them a little differently than other teachers. So I like the fact that the multiple-choice is pretty cut and dried and that you are marking exactly the same way as everybody else. (Owen interview, p.6)

Owen's concern about the need for standardized teacher-derived marking of the EQAO assessment items from one class to another within his school is puzzling given that this school has a non-common in-school exam as well as non-common performance tasks as the other parts of the summative grade for their Grade 9 mathematics courses. This means that there is considerable variation from class to class in these aspects of the summative assessment practice at Taylor Brook and yet Owen seems to want to keep the approach to the EQAO assessment the same for all teachers. When I asked him why he was concerned about using a standardized approach to scoring the open-response items on the assessment given that the other summative tasks were non-common, he stated that it probably did not make a lot of sense but indicated that he still felt it was important for items on the EQAO assessment to be scored in a consistent manner.

The department head at Taylor Brook indicated that class size was a factor that was considered when selecting which items to score. She explained that because Academic classes are quite a bit larger (25-30 students) than Applied classes (12-15 students), marking the open-response items on the EQAO assessment would be more problematic for the Academic courses than for the Applied given the time frame in which they have to return EQAO materials (Diane interview, p. 5). She also indicated that the school wants the same approach toward marking the EQAO assessment items to be used for both the Academic and Applied courses. Thus, class size was a factor influencing the decision to mark only the multiple-choice items in this school.

One of the EQAO personnel I interviewed, Marie, who had formerly been a classroom teacher described how she selected EQAO Grade 9 assessment items to score and include in her students' grades. Marie indicated that she had taken a number of different

approaches in the years that she had used the assessment. One semester she had not had time to give her students a unit test for the geometry unit they had completed shortly before the EQAO assessment took place so she told the students that she would select all of the multiple-choice and open-response items on the EQAO assessment related to geometry, mark them and include the grade on those items as the equivalent of a unit test mark for that part of the course (Marie interview, p.3). Accordingly, for that semester, she marked only the geometry items on the EQAO Grade 9 assessment. Another approach Marie sometimes used was to consult the students in her class to decide which EQAO items to mark. After the students had written the assessment, Marie described how she had a discussion with the class and asked them to decide as a group which of the open-response items she should mark and include in their grades (Marie interview, p.6). Once they had come to a consensus, Marie developed a marking key for the item the students had identified and that was the only open-response item she scored on the assessment for that semester.

The observations from this study suggest that there is considerable variation in the selection of EQAO assessment items to be scored and included in students' grades. Some teachers score only multiple-choice items while others choose a selection of multiple-choice and open-response items to score. In some schools the selection of items is done collaboratively while in other schools individual teachers decide on the items they will score. Many factors influence the selection of items to be scored including the amount of time available for marking the EQAO assessment items, a concern with consistency in scoring approaches, the alignment of the EQAO items with the approach which has been taken in the mathematics classroom, as well as the mathematics content that has been assessed on unit tests or will be assessed on other summative assessments tasks.

Scoring procedures.

As noted earlier, EQAO does not provide teachers with a marking key for the multiple-choice items or with any item-specific scoring guides for the open-response items. Teachers who decide to include the EQAO assessment items as part of students' grades must develop their own scoring procedures. At each of the schools in this study, teachers described the procedures they developed to score the EQAO Grade 9 assessment items. In this section I offer a summary of those procedures.

At Birch Park Secondary School where teachers mark both the multiple-choice and the open-response items, each teacher develops their own scoring key and little collaboration among teachers takes place. For multiple-choice items, each teacher indicated that they work out their own solutions and then score their students' responses as correct or incorrect. For the open-response items the process is more complex. Each teacher indicated that they develop a scoring key for these items based on the way that they would mark similar kinds of items on the unit tests they develop and use during the term. In marking the open-response items on the EQAO assessment several teachers indicated that they consider the mathematical processes students use as well as the way students communicate their understanding of mathematics concepts in addition to the correct response. Each teacher described how they write out sample responses to each item in the booklet which EQAO provides for this purpose and then decide how many marks should be allocated to each item. Students' responses to the open-response items are then scored with reference to these sample responses without making any marks on the students' EQAO booklets or photocopying their responses. The booklet which contains the marking key the teachers

devised is returned to EQAO along with the students' completed booklets, as required by EQAO, so that no evidence of the marking procedure which was used remains at the school.

At Scottsdale Collegiate scoring of the EQAO items is done in a collaborative manner. Each teacher I interviewed indicated that they work together to establish a marking key for the multiple-choice and open-response items that they have decided to mark. One or two teachers collaborate to complete an initial scoring guide which includes solutions for each item as well as an indication of the number of marks to be allocated to each open-response item (Whitney interview, p.2). This initial marking key is then circulated to the other Grade 9 mathematics teachers in the department and revised on the basis of their input. For example, Whitney describes this process as follows:

As Grade 9 teachers we get together and decide which questions do we like, which questions are we going to not mark or look at because we don't like them . . . so I did it this year and I said "okay this how much I think each question should be worth" and then we kind of talk about it. "Oh, I don't like this question." "Okay, so we won't do it." "Oh, I think this question deserves more marks" and then we'll just kind of tweak it. (Whitney interview, p.2)

As a result of this process, all teachers at Scottsdale mark the same items on the assessment and use the same scoring guide when they do their marking.

At Taylor Brook only the multiple-choice items are scored and included as part of students' grades. Each of the teachers I interviewed indicated that they get together to ensure that they all have the same answers for the multiple-choice items before they begin scoring. Some teachers create a template with correct answers on an overhead transparency that can be placed over students' bubble sheet response sheets enabling the teacher to quickly score the assessment without making any marks on students' EQAO booklets. Two reasons were given by the teachers at Taylor Brook for working collaboratively to develop the marking key. First, the teachers indicated that they want to ensure that they all agree on the correct

response option. In addition, one teacher indicated that from time to time the teachers have discovered errors in the marking key they created for the multiple-choice items because of the different field-test items which can be included in each student's test booklet (see the description of field test items in chapter 4). Adrian explained "sometimes you will notice that a student who you'd expect to do particularly well may have suddenly not done well and that gives you a reason to go back and check and see whether or not there was a difference in the questions" (Adrian interview, p.3). By working collaboratively, teachers can flag if there are different field-test items in some of the booklets and reduce the risk of making errors as they mark the multiple-choice items.

These school-level case studies provide evidence that teachers have developed a wide variety of approaches for scoring the EQAO Grade 9 assessment items. While scoring the multiple-choice items was viewed as fairly straightforward by most of the participants some concern was expressed about teachers' ability to reliably discern the different embedded field-test items that may appear on the assessments administered in their department. In scoring the open-response items, teachers develop sample responses and decide how many marks they will allocate to each item. In one school that scored open-response items this process was done collaboratively while in the other school that scored open-response items this process was done by individual teachers. A number of factors seem to influence the scoring procedures which are used including concerns about consistency of scoring between teachers, consistency of scoring open-response items on the EQAO assessment in the same manner as other open-response items students have completed on classroom assessments, and the development of efficient ways of scoring that do not make marks on students' EQAO booklets.

Having summarized the range of procedures teachers use to include the EQAO Grade 9 assessment as part of students' grades, I now turn to a description of teachers' views of this use of the assessment.

Description of Teachers' Views of the Use of the Assessment in Students' Grades

To describe teachers' views of the use of the EQAO Grade 9 assessment as part of students' grades, I draw on the observations from the school-level case studies. Three aspects of the analysis I conducted are described: the meaning which participants ascribe to EQAO scores, the rationale and benefits of using the assessment as part of students' grades, and the tensions which were identified as emerging from teachers' use of the assessment as part of students' grades.

Meaning teachers ascribe to EQAO scores.

During the analysis of the interview transcripts from the school-level case studies, I created a table to keep track of comments made by participants with regard to the meanings they ascribe to both the official and the teacher-derived EQAO scores. As I indicated in chapter 3, Kane's argument-based approach begins with a consideration of the interpretations or meanings which are ascribed to the assessment scores and, for this reason, I felt I needed to get a sense of these meanings from the participants in the study. As I conducted this part of my analysis I encountered a number of difficulties. In many cases it was not possible to ascertain if a participant's comments about the meaning of a score referred to the official score or to their teacher-derived score. In other cases, the participants describe the meaning they ascribe to scores from earlier versions of the EQAO Grade 9 assessment but the meaning which they would ascribe to the current version of the assessment would presumably be different. For example, the principal at Birch Park

Secondary School indicates that the EQAO Grade 9 assessment provides students with “a very nicely detailed report” which identifies the strengths and weaknesses in their math knowledge (Allen interview, p.1). In this description, Allen is referring to the individual student reports which used to be produced by EQAO and included separate scores for each mathematics strand. As noted, current EQAO individual student reports provide one overall mathematics score.

Another challenge I encountered in describing the meaning teachers ascribe to the EQAO Grade 9 assessment is that many participants seem to ascribe a different meaning to the Academic assessment than they do to the Applied assessment. For example, one department head comments “in the Applied I find that because we’re doing a lot of activity-based, a lot of manipulatives, a lot of hands-on stuff that when it comes to EQAO there is a disconnect between how we are teaching the kids and how they actually are assessed on that. So, I see that disconnect but in the Academic I don’t see as much of a disconnect” (Francis interview, p. 2). This comment, along with similar comments made by several other participants, provide evidence that teachers may not ascribe the same meaning to the Academic and Applied assessments.

Further, as I looked across the comments participants made with regard to the meaning of the scores, I found little commonality. That is, there were almost as many meanings ascribed to the scores as there were participants in the study. Focusing on the comments made by participants in one school shows the range of views expressed. For example, at Birch Park Secondary School both Allen and Francis describe the EQAO Grade 9 assessment as a “snapshot” which provides additional information about how well students are doing in their mathematics courses (Allen interview, p.6, Francis interview, p.2).

However, both also caution that the assessment may be a more accurate snapshot for the Academic than the Applied mathematics courses and Allen notes that the scores can be significantly impacted by various factors such as the time of day when students complete the assessment. In the same school, William sees the EQAO Grade 9 assessment as a “a good barometer” and a “good overall overview” of what the Ministry of Education says students should know at the end of their Grade 9 mathematics course (William interview, p. 5) while Margot and Shannon both see the assessment as a measure of students’ ability to apply their mathematics knowledge and skill to new situations (Margot interview, p. 2; Shannon interview, p.3). In addition to these examples of the varied meanings participants ascribe to the EQAO Grade 9 assessment, as I describe the benefits and tensions teachers associate with the use of the EQAO Grade 9 assessment in the next section, several additional aspects of the meanings ascribed to the EQAO scores are revealed.

In an effort to clarify my findings, I repeated this part of my analysis. However, after repeating the analysis, I concluded that getting an understanding of the meaning that teachers ascribe to the EQAO scores would require a series of interview questions focused specifically on this topic and that this aspect of teachers’ use of the EQAO Grade 9 assessment could easily become the focus of another complete study. In chapter 7, I return to a discussion of how these observations about the various meanings of the EQAO Grade 9 assessment informed my response to the research question regarding the use of Kane’s model (see research question #5).

Rationale and benefits of using the assessment as part of students’ grades.

To gain an understanding of why participants at these schools decided to include the EQAO Grade 9 assessment items as part of their students’ grades, I considered both the

rationale that teachers indicated for their decision as well as the benefits they identified as coming from their use of the assessment as part of students' grades. In all three schools participants' comments indicate that there had been resistance to the use of the assessment as part of students' grades in the first few years after the assessment was introduced. As noted earlier in this section, the principal at Birch Park Secondary School indicated that teachers initially viewed the assessment as politically motivated and serving the needs of the government (Allen interview, p.2) and the mathematics department head at Scottsdale Collegiate indicated that the EQAO Grade 9 assessment was not used by teachers as part of students' grades during the first few years that the assessment was administered (Mark interview, p.3). The situation was even more pronounced at Taylor Brook High School where the use of the assessment as part of students' grades was initially seen as an invalid practice and was not permitted by the school district. Teachers at Taylor Brook changed their practice with regard to use of the EQAO assessment as part of students' grades when the district policy changed requiring them to include teacher-derived scores on the EQAO assessment items as 5% of students' grades. According to the participants in my study, in the other two districts, as teachers became more accustomed to the EQAO assessment items and procedures their initial reluctance to using the assessment faded and teachers began to include the assessment as part of their students' grades.

In all three schools the primary reason given for using the assessment as part of students' grades is to ensure that students take the assessment more seriously. For instance, a teacher at Taylor Brook comments:

When they know that it's for marks, they tend to take it more seriously and I think it guarantees a better overall indication of the students' performance because it could be that a student who is achieving well just decides not to take it seriously and doesn't get, you know, a true result on the EQAO test. (Adrian interview, p.2)

Numerous other teachers, department heads, and principals who participated in this study made similar comments. In all three schools, participants indicated that using the assessment as part of students' grades was necessary because no other stakes are associated with this assessment. That is, successful completion of the EQAO Grade 9 assessment is not a requirement for graduation or for obtaining credit for Grade 9 mathematics courses. These participants indicated that the EQAO Grade 9 Assessment of Mathematics contrasts with the EQAO Ontario Secondary Schools Literacy Test (OSSLT) in that successful completion of the OSSLT is a graduation requirement. Two examples of comments about the difference between the EQAO Grade 9 assessment and the OSSLT are included to illustrate this perspective. The first comment is from the department head at Taylor Brook and the second comment is from a teacher at Birch Park.

We are using it to motivate kids to try their best . . . the problem is that it doesn't count. Students work really, really hard on the Grade 10 literacy test [OSSLT] because it's a pass/fail on their record . . . for math it doesn't mean anything to them and so they don't care. (Diane interview, p.16)

The biggest problem with EQAO is just to take the test seriously. I've had to administer the literacy test [OSSLT] before and I found the students really take that seriously because it means graduation. There is a real important reason to pass that test. What I notice is my students' attitude, like this is my first year doing it, but their attitude is "Well, it's not worth any marks, we really don't care about it". (Shannon interview, p.4)

A teacher at Taylor Brook High School suggested that including the assessment as part of students' grades may be most effective in motivating students to take the assessment seriously for those students who are in danger of failing the course. He stated "I guess if I was to think of that student that was the 50 or 60% student, that didn't care about it before, maybe they care a little bit more now" (Owen interview, p.4).

Some differences were found with regard to the rationale for using the assessment as part of students' grades in the Academic and Applied courses. A number of teachers indicated that motivating the students to take the assessment more seriously is less important for students in the Academic courses who already take the assessment seriously than it is for students in the Applied courses. At the same time, several teachers expressed the view that many students in the Applied courses do not take the assessment seriously even when it is included as part of their grades and that including the assessment as part of students' grades may be less appropriate because these students often have difficulty with written tests. A teacher at Birch Park Secondary School explains:

I find that if you tell them it doesn't count at all, especially at the Applied level, well they're not going to put in any effort and they're going to leave it blank but if they know at least there is some purpose, there is something hanging over them, that it does count for something then they put in a little bit more effort and they do try harder on it. . . . but EQAO is not designed for Applied students. The whole thing with Applied students is that they're not test writers, most of them, you know, I find just the whole idea of this big two day test for them is a bit overwhelming and so I'm definitely more lenient with theirs [as compared with Academic students] because I don't feel it is the right way to assess those students. (Margot interview, p.2)

The department head at Scottsdale Collegiate also expressed his concern and stated that students in the Applied courses "are not motivated by any kind of assessment...We have to do all kinds of funny things to try to motivate them to come to write the assessment like offer free food and things like that" (Mark interview, p.6). These comments illustrate the particular challenges teachers expressed with regard to the use of the assessment as part of students' grades for Applied courses.

The department head at Scottsdale Collegiate suggested an additional reason for using the assessment as part of students' grades was to ensure that *teachers* took the assessment more seriously. This department head also said that there was value in getting

teachers to think more carefully about the assessment items they include on their classroom assessments and that marking the EQAO assessments might help to achieve that because the items on the assessment have been carefully developed and are of good quality (Mark interview, p.5).

Participants in the school-level case studies identified a number of benefits which result from this use of the EQAO assessment. Several teachers indicate that including the assessment as part of students' grades provides a more accurate measure of students' mathematics understanding both for EQAO's accountability purposes and for classroom teachers attempting to get a sense of their students' grasp of the curriculum *because* students take the assessment more seriously. These teachers expressed the view that when students do not put effort into completing the EQAO assessment items both the teachers and EQAO get a less accurate measure of students' understanding of the curriculum. A related benefit expressed by a number of teachers is that they consider the assessment to be closely aligned with the curriculum and believe that using the assessment as part of students' grades helps teachers to get a better sense of the extent to which they are addressing the curriculum in their teaching. The principal at Birch Park Secondary School summarized this view in explaining that using the assessment as part of students' grades helps teachers to answer the question "are we teaching the standards that are expected?" (Allen interview, p.2) and a teacher at the same school describes the assessment as providing "a good overall overview of what the Ministry says that they [students] should know" (William interview, p.4). A teacher at Scottsdale Collegiate expressed a similar view by stating "It's a really good summative assessment of the whole curriculum, the whole course" (Leanne interview, p.5).³⁰

³⁰ These sorts of comments are further examples of the varied meanings which participants ascribe to the EQAO scores, as described in the previous section of this chapter.

Some participating teachers expressed the view that using the assessment as part of students' grades is beneficial because they become more familiar with the items developed by EQAO. This greater familiarity was seen as beneficial both in helping teachers to prepare their students to complete the assessment and because the quality of the items created by EQAO was described by some teachers as being better than some of the items these teachers create for their classroom assessments. For example, one teacher states "a lot of the questions are actually quite good. There's a lot of nice thinking involved" (Whitney interview, p.2) and a mathematics department head described the items as providing an exemplar for classroom teachers to follow in creating their own items (Mark interview, p.5). A number of teachers indicated that they try to develop items for their unit tests that are similar to those on the EQAO assessment and they indicated that they thought that this was also beneficial. It seems likely that in large-scale assessments where teachers are not marking the items, teachers would not look as closely at students' responses or have a chance to review the items on the assessment; they would be more likely to simply bundle up the completed assessments to be shipped to the assessment agency for scoring.

Some teachers indicated that the assessment was beneficial in providing them with an opportunity to see how students could complete items that were not created by their own teacher. These teachers felt that the externally developed items on the EQAO assessment provide a more objective measure of students' mathematics performance and that they could get a better sense of students' performance on these items by scoring the items and including them in students' grades rather than by reviewing the official score reports which are sent to schools the following year. One teacher describes this benefit as follows:

Because it's, of course, a different style of test. You know, I didn't write it. I didn't specifically prepare them to answer that type of question or that way of thinking and

I just think it's really neat that you teach them to think in one way but they can apply it to other styles of questions. I think that's really, really neat. (Whitney interview, p.4)

Similarly, another teacher states:

At times it [marking the EQAO assessment] has seemed like a bit of a burden. But on the other hand, I guess it's nice to do something that's identical to everybody else so you can see how your kids stack up. And when you get positive results, it's reassuring. When you get negative results, I guess it causes you to do a little reflecting and introspection. (Owen interview, p.8)

Teachers who had used the assessment items in lieu of an in-school final exam or had considered adopting this practice indicated that another benefit of this use was that the end-of-course assessment load on students would be reduced to include only the performance task and the EQAO assessment. Teachers' workload would also be decreased if the EQAO assessment was used in lieu of the in-school exam because teachers would not need to spend time preparing or scoring the in-school exam in addition to the EQAO assessment items.

One additional benefit identified by a number of teachers is that using the assessment as part of students' grades helps to shorten the relatively long feedback loop associated with this assessment. That is, official EQAO results from the Grade 9 assessment are not received by classroom teachers until the fall of the following school year. Several participants indicated that receiving the results in the following school year makes it difficult for teachers to use the results to improve their mathematics programs. By marking the EQAO assessments immediately after students complete them, teachers can get a sense of students' strengths and weaknesses and they can use this information to make changes to their mathematics program as they plan for the following semester's courses. For example, one teacher states "it helps us to decide how we want to change the program, if there are certain things that need to be adapted for the next semester or anything like that" (Adrian interview,

p.5). As one of the EQAO personnel who had formerly been a mathematics teacher stated “that's one good reason why teachers would like to use this as a part of their marks because it gives them right away an idea of how their students have performed on that test, even though it's not necessarily the same way it is going to be marked during the summer” (Mitchell interview, p.13).

Tensions emerging from teachers' use of the EQAO Grade 9 assessment.

While participants described a number of benefits associated with using the assessment as part of students' grades, many tensions with this practice were also identified. In this section I summarize the tensions identified by the participants under a number of categories and I illustrate each category with one or two examples.

Tensions related to weighting.

Teachers identified many tensions related to deciding how much to weight the assessment when including it as part of students' grades. Some teachers indicated that they were trying to weight the assessment as part of students' grades heavily enough to ensure that students would take it seriously but not so heavily that completing the EQAO Grade 9 assessment would become overly stressful for students. Observations from the school-level case studies suggest that finding a weighting for the assessment which balances these two goals is not a straightforward task. Several teachers indicated that they did not want the assessment to count for too great a portion of students' grades because they were uncertain about the wording of the items and about students' familiarity with the items. One teacher at Birch Park Secondary School indicated that including students' EQAO responses as part of grades was particularly problematic for students who do not perform well on tests. He described the way that he had decided not to count the EQAO items as part of students'

grades for two students in his Grade 9 Academic course who had had difficulty successfully completing tests throughout the semester (William interview, p.5).

Although most of the teachers who participated in this study had decided how they would weight the assessment for students' grades, many expressed their uncertainty about the actual impact of this practice on their students' motivation. Some indicated that students in their Academic courses tended to take the assessment fairly seriously whether or not it counted while others indicated that they felt that many students in the Applied courses would not be motivated by having the assessment contribute to their grades even if the weighting was quite substantial. Thus, while getting the students to take the assessment seriously is offered by the participants in this study as the principal reason for using the EQAO Grade 9 assessment as part of students' grades, most teachers are not certain of the extent to which this technique is effective in achieving this goal.

Tensions related to scoring.

Many teachers expressed concerns with regard to scoring the EQAO assessments. In general, scoring multiple-choice items was viewed by all participants as more straightforward than scoring the open-response items. However, tensions with scoring both item types were described by teachers. For example, as noted, one teacher was concerned that those field test multiple-choice items that differ from one student's assessment to another would not be identified during the scoring process and this oversight could result in responses mistakenly scored as incorrect because they would not match the scoring template teachers had developed. Several teachers expressed their concern that the classroom teacher's approach to marking open-response items might be different from the way that EQAO marks the items. While those teachers who had decided to mark the open-response

items indicated that they would mark these items in the same way that they mark similar items on unit tests, some of these teachers also felt that this approach might create confusion for students who are not certain whether they should structure their response so that it conforms to the teachers' expectations or so that it is aligned with the marking criteria for EQAO. For example, one teacher explained the concern she has with the criteria for a Level 4 response at EQAO being different from the criteria for a Level 4 response she uses in her class. Shannon states:

I have shown the students what [EQAO] says is a Level 4 answer and one of the big things that has come out is that the student can answer the long answer question and get the wrong answer and still get a Level 4. And so, me personally, I do put a little bit more emphasis on the right answer . . . when I go to mark it they have to be correct. (Shannon interview, p. 2)

Shannon also indicated that students are sometimes confused with regard to what their responses to the open-response items need to include because of the difference between her marking criteria and the one used by EQAO.

Tensions related to providing feedback to students and parents.

Several teachers expressed the view that using the assessment as part of students' grades was problematic in terms of the lack of feedback they can provide to students and parents. Because teachers cannot photocopy students' responses to the EQAO items they are not able to provide useful feedback to their students. In addition, teachers do not have a record of the marking guide they used to score the open-response items to share with students and parents. These teachers indicate that providing feedback to students about their responses to specific test items and being able to substantiate grades that students receive with concrete evidence of students' performance are key aspects of classroom assessment.

For example the mathematics department head at Taylor Brook High School where the assessment is used as 5% of the summative grade states:

They [the students] don't even know what three out of five means because I can't show them, I can't say these are the questions you got right and these are the questions you got wrong because we can't photocopy them. And so it's very abstract for them [the students]. (Diane interview, p.16)

Tensions related to item types, content and wording.

Many teachers expressed concern with regard to the item types, content and wording on the assessment. These tensions became apparent as teachers were marking the items to include them as part of students' grades. For example, two of the mathematics department heads, Diane and Mark, expressed their concern with the increased emphasis on multiple-choice items in recent versions of the Grade 9 assessment. They were concerned that multiple-choice items can be difficult for students to answer and provide less information about students' problem solving skills than open-response items. In addition, they both felt that the increased emphasis on this type of item was resulting in teachers tending to use more multiple-choice items in creating their classroom assessments.

A number of teachers indicated that while marking the assessments they sometimes found items that they had not covered in their teaching or that they felt were not a part of the mathematics curriculum. Teachers also expressed concern that the wording of the items on the assessment was sometimes different from the wording they would use in creating items for their classroom assessments and that students sometimes found the wording unclear. These concerns were substantial enough for one teacher describing her approach in using the assessment in a Grade 9 Applied course to state:

And I have even, sometimes I have gone and said that it will count as quiz marks but the EQAO that year really did not match with whatever, so yes, I did mark them but I didn't count them because it wasn't indicative of what my students had been doing, of the way I had been teaching them. (Margot interview, p.1)

Tensions regarding the use of the assessment in Applied mathematics courses.

An important finding which surfaced in this research is a fairly widely held view among teachers, department heads and principals that the EQAO Grade 9 assessment is less appropriate and informative as an assessment for students in the Applied mathematics course than it is for the Academic course. All three department heads expressed the view that the use of the assessment as part of students' grades is less appropriate for students in the Applied course than in the Academic course because the use of manipulatives and hands-on approaches is more frequent in teaching Applied courses and this is not how the students are assessed on the EQAO Grade 9 assessment. One teacher at Birch Park Secondary School indicated that she believes that students in the Applied mathematics courses often struggle with completing tests and that when test items are used in an Applied course they should be broken down into smaller steps so that students do not become overwhelmed. She indicated that she did not want to use the EQAO assessment as part of students' summative grades in her Applied course because she did not want the assessment to have too great an influence on her students' grades. She stated "I don't find EQAO is geared to the way I teach my students. The questions aren't worded the way that I word my questions" (Margot interview, p.1). A number of other teachers expressed similar views with regard to the use of the assessment as part of students' grades in their Applied mathematics courses. As indicated in the previous section, these concerns are particularly problematic given the perceived need for motivating students in the Applied courses which these teachers also noted.

The tensions observed in this study regarding the use of the assessment in Applied mathematics courses are also interesting in light of the persistent gap in achievement on these two assessments (see Table 2) and the concerns expressed by Ross and Gray (2005) and Wolfe et al. (2004) with regard to the Grade 9 Applied assessment. As noted in chapter 4, Ross and Gray found the alignment of scores on the EQAO Grade 9 Applied assessment with students' report card grades lower than the alignment of the scores on the Grade 9 Academic assessment with students' report card grades. On the basis of this finding, Ross and Gray suggest that the Grade 9 Applied assessment may be less aligned with classroom assessment practices than the Grade 9 Academic assessment. Wolfe et al. (2004) recommended that EQAO conduct research to investigate the appropriateness of the Grade 9 assessment for Applied courses. Thus, the tensions described by participants in this study add to and further elaborate on concerns that have been expressed in previous research.

Tensions related to time constraints and test security.

One of the most pervasive areas where tensions arise related to teachers' use of the assessment as part of students' grades is with regard to time constraints. This tension was identified at all three participating schools. One way that time constraints create tension is with regard to the selection of items. Teachers are not permitted to open the sealed assessments until the morning that the assessment is to be administered (EQAO, 2007). This means that they do not have very much time to review the items, identify those that they will mark and then mark students' responses before the assessments must be returned to EQAO. Although EQAO indicates that the assessment can be administered during any two class periods within the 15-day testing window at the end of each semester, most schools administer the assessment towards the end of this time period as they are struggling to finish

teaching the mathematics course content. This means that the completed assessments must usually be returned to EQAO one or two days after they are administered, making the time available to select and mark students' responses quite limited. Some teachers expressed their concern that the time constraints they face in selecting and scoring the items to include in students' grades might mean that these procedures are not as well thought out as they should be. These time constraints are particularly problematic for teachers who have more than one Grade 9 mathematics class. For example, one teacher indicated that she had taught three classes of Grade 9 mathematics during one semester in the previous academic year and it was very difficult for her to find time to mark even the multiple-choice items from the EQAO assessment.

The requirements for storing the completed assessments in a secure space augment the tensions related to time. In some schools completed assessments are stored in a secure location and teachers are not permitted to take them home for grading purposes. For example, teachers at Taylor Brook High School indicate that the assessments are stored in the school vault which is locked at 4:00 PM each day. This leaves only about one hour after school each day for teachers to mark the assessments.

Another time related constraint indicated in all three schools is the number of assessments taking place near the time period when the EQAO Grade 9 assessment is being administered. In all three schools students are completing their summative performance assessment task and an in-school exam within a week or two of the time when the EQAO assessment is administered. Teachers expressed the view that this creates considerable stress for students completing these three assessments as well as a substantial time commitment for teachers who must mark all three tasks. This heavy end-of-term marking load was identified

by several teachers at Taylor Brook as one reason they had decided to mark the multiple-choice items rather than the open-response items.

Summary of How Teachers' Use of the Assessment Takes Place

In this section I have provided many observations related to how teachers' use of the assessment as part of students' grades takes place. Before moving on to the third research question, I offer a summary of my response to the second research question - *How does teachers' use of the EQAO Grade 9 Assessment of Mathematics as part of students' grades take place?*

The questionnaire results suggest that most teachers across the province are using the EQAO Grade 9 assessment in their students' grades. Analysis of the responses by school district indicates that in most boards at least some teachers use the assessment as part of students' grades but there are many boards where some schools within the board use the assessment as part of students' grades and others do not.

More is learned about teachers' practices and views of the use of the EQAO Grade 9 assessment from the school-level case studies. These studies provide evidence that the approach to weighting, item selection and scoring of the EQAO Grade 9 assessment as part of students' grades can differ within a classroom, within a school, and within a school board. Participants describe a range of weighting from schools where the assessment is not included in students' grades to schools where the EQAO assessment is used in lieu of an in-school exam and counts for as much as 20% of students' grades. With regard to item selection and scoring, in some cases only the multiple-choice items are scored while in others a selection

of multiple-choice and open-response items is scored.³¹ The selection of items to be scored is based on a wide range of factors such as time constraints, how well the items on the assessment match the approach teachers have taken in the course and the content of items included on other summative assessments being administered. Scoring of multiple-choice items is seen as more straightforward than scoring of open-response items. In scoring the open-response items, teachers in this study indicate that they base their approach on how they would score similar items on the assessments they give their students rather than on the scoring approach that is used by EQAO.

Looking at teachers' views of this practice, the main benefit which teachers associate with the use of the EQAO Grade 9 assessment as part of students' grades is that the increased motivation students may have as they complete the assessment items may mean that the assessment produces a more accurate measure of students' mathematics understanding. Teachers felt that this more accurate measure is beneficial for EQAO, teachers and the students themselves. However, many tensions regarding this use of the assessment were described. The teachers who participated in this study indicated that they experience difficulties with deciding on an appropriate weighting for the assessment, developing scoring methods for open-response items, providing adequate feedback to students and with the constraints related to time limits and test security. Many of the tensions which were identified were seen as particularly problematic for the use of the assessment as part of students' grades in the Grade 9 Applied Mathematics course.

These observations suggest that teachers' use of the EQAO Grade 9 assessment is very much a local practice and many teachers, whether acting individually or in

³¹ Since this research was completed I have also learned of schools who have decided to score only the open-response items in an effort to avoid the difficulties which many Grade 9 students experience with multiple-choice items.

collaboration with other teachers in their department, adjust their use of the assessment to suit their own classroom or school. That is, they decide on the weighting, select the items they wish to mark, decide not to include it for certain students, choose to mark only the geometry items, or give students the choice of whether to include the EQAO assessment or the in-school final exam in their grades. In many ways these teachers are treating the EQAO assessment as a classroom assessment – making dynamic and emerging decisions based on aspects of their particular classroom such as the content they covered that semester or the other assessment tasks they are using. These observations suggest some of the ways that the distinction between large-scale and classroom assessment is blurred when teachers use the EQAO assessment items as part of their students' grades. Further discussion of the distinction between large-scale and classroom assessment is included in subsequent chapters.

Research Question 3 – EQAO's Perspective

How is teachers' use of the EQAO Grade 9 Assessment of Mathematics as part of students' grades viewed by EQAO?

The perspectives of EQAO personnel are presented by focusing on the evolution and rationale for policies regarding this use, the tensions identified by EQAO personnel, and the issue of the comparability of official EQAO results given teachers' use of this assessment as part of students' grades. The data sources which are drawn on for this description include analysis of the EQAO documents as well as the interviews and email inquiries with EQAO personnel.

Rationale & Evolution of EQAO Policy Regarding this Use

A description of the origin, rationale and evolution of EQAO policies regarding this use of the assessment was developed using information from interviews, email inquiries and my analysis of EQAO documents. These sources indicate that EQAO permitted the use of

the Grade 9 Assessment of Mathematics as part of students' grades from the first operational year of the assessment which was the 2001-02 school year. As Susan, who was working at EQAO at that time, states "they were allowed right from the start. If you can find a copy of the first admin guide there is a statement in there." (Susan interview, p.15). This was also confirmed in several other interviews and in my email communications with a director at EQAO (personal communication, October 9, 2008). In addition, this finding is consistent with Lock's (2001) dissertation describing the views of teachers who included the assessment as part of students' grades during the first year that the assessment was administered (see chapter 4).

The interviews I conducted also provide information about how this use of the assessment came about. Several participants indicated that in the year before the first operational year of the EQAO Grade 9 assessment, EQAO administered a sample Grade 9 assessment in the schools and received feedback from teachers indicating that students had not taken the assessment seriously. As stated by Susan "one of the suggestions from a number of math teachers was that we provide the opportunity for teachers to let the test count and that way the kids would take it more seriously" (Susan interview, p.1). To further contextualize EQAO's decision to let the teachers use the assessment in this manner, it is important to recognize that the Grade 9 assessment was being developed at the same time as the OSSLT. According to Susan, the Ministry of Education had already decided that the OSSLT would be a requirement for students' to graduate but no similar requirement was attached to the Grade 9 Assessment of Mathematics. Thus, as observed by teachers during the sample assessment, students had little motivation for completing the assessment to the best of their ability. Susan also indicated that the decision to allow teachers to use the

assessment as part of students' grades was carefully considered and discussed at length before it was made including being taken to the CEO of EQAO for approval (Susan interview, p. 15).

A number of EQAO personnel indicate that in the first few years that the assessment was administered it was not used for grades by many teachers. Further, several personnel that I interviewed indicated that EQAO was initially fairly neutral with regard to whether or not teachers actually used the assessment items as part of students' grades. They noted that in the early years, students, particularly those completing the Applied assessments, left quite a few of the items blank. Many of the EQAO personnel indicated that using the assessment as part of students' grades was beneficial in the sense that motivating students means that the results from both the teacher-derived and the official scores are a better reflection of students understanding of the curriculum. These individuals also indicated that in recent years EQAO has begun to more actively encourage teachers to use the assessment as part of students' grades. While no changes in policy or in the assessment design have been made related to teachers' use of the assessment for grades, these participants indicate that EQAO encourages the practice by providing more resources to support teachers as they mark the EQAO assessment items. As noted, EQAO does not provide an item-specific scoring guide for the current year's assessment but they do provide generic scoring rubrics for open-response items as well as sample scoring guides and anchor responses for released items from previous assessments on the EQAO website. Teachers may use these resources to create marking guides for the current year's open-response items. Several of the EQAO personnel that I interviewed felt that EQAO's decision to provide these resources had been made to support and encourage teachers to use the assessment in their students' grades. In addition,

while the number of assessment booklets is tightly controlled and all booklets must be returned to EQAO, EQAO currently provides each teacher with an assessment booklet that they can use to record their answers to the items they intend to mark. This booklet must be returned to EQAO along with students' completed test booklets. According to some of the EQAO interview participants, this procedure began two or three years ago and was adopted to encourage and support teachers in their use of the assessment as part of students' grades.

Despite these encouragements, each of the EQAO personnel emphasized that teachers' use of the assessment as part of students' grades remains optional. Several participants pointed out that classroom assessment policies are mandated by the Ministry of Education, not by EQAO and for this reason EQAO is not in a position to require all schools to use the assessment as part of students' grades, to establish a policy about how much the assessment should contribute to grades or to specify the approach that teachers should use to mark the open-response items.

Tensions Identified by EQAO

Interview participants from EQAO identified several tensions which result from teachers' use of the assessment. A few participants noted that discrepancies between the official scores and the teacher-derived scores can be a problem. The teacher-derived scores create expectations in teachers, students and parents who anticipate that the official scores will be similar to the teacher-derived scores and then question the value of the official EQAO scores when they are eventually received. Both Mitchell and Susan explained that when there is a difference between the two EQAO scores, parents sometimes call EQAO and ask for an explanation but since EQAO has no knowledge of the marking approach that was used by the teachers there is little information they can provide to parents.

Several EQAO participants indicated that they had concerns with the appropriateness of using items developed for large-scale assessments as part of students' grades. One explained that large-scale assessments typically do not include items which all or nearly all students would be expected to get correct (Marie interview, p.7). Items are only included on the EQAO assessment if they help to discriminate among test-takers. However, Marie noted that teachers often include items on their classroom assessments that they expect all or nearly all students to answer correctly. She states:

So some people question that because we [EQAO] will never put something that is 98% success [on the assessment]. And they say 'why not' and it's because we have to discriminate, we have to put the kids in four levels. That is different than a classroom because in the classroom you can have questions that everybody will [do correctly] and you want to have that. (Marie interview, p.7)

Thus, according to Marie, while large-scale assessments seek to discriminate among groups of students, classroom assessments do not have this goal. Other participants expressed the view that while multiple-choice items are the most valid and reliable item type for large-scale assessments they are not ideal for classroom assessment as many students struggle with them and for this reason using these items as part of students' grades might not be ideal (Mitchell interview, p. 2, Melanie interview, p.4).

Another tension which was identified by EQAO personnel is that teachers sometimes contact the assessment developers when they are marking the assessments to question the connection between specific items on the assessment and the Grade 9 mathematics curriculum. One of the EQAO participants indicated that she had been asked on several occasions to demonstrate the relevance of EQAO assessment items to the mathematics curriculum (Dana interview, p.6) and that these questions arise because teachers look at the

items more closely than they would if they were not using the assessment for students' grades. A similar view was expressed by Susan.

As was observed by some of the school-level case study participants, another tension identified by EQAO personnel is with teachers deciding how to mark the open-response items. Two EQAO participants suggest that teachers should mark these items in a manner which is consistent with what they normally would do in their class. However, another EQAO participant indicates that teachers continue to ask for item-specific scoring guides for the items they are scoring but EQAO has no intention of providing those item-specific scoring guides (Susan interview, p.1).

One other tension which was identified is that in recent years teachers of Grades 3 and 6 have begun to ask if they can use the EQAO assessments which are administered in those grades as part of their students' marks. Marie states "I was asked just last week for Grades 3 and 6 if they could use it [the EQAO assessment] for their marking . . . but it's not set up for them at all. Like it's set up really for the Grade 9 teachers to do the marking because the teacher receives the booklet so they can do their own test, and decide what questions to mark and all that" (Marie interview, p. 1).

The Issue of Comparability across Schools and Districts

When asked to describe tensions associated with teachers' use of the EQAO Grade 9 assessment as part of students' grades, none of the EQAO personnel I interviewed suggested that the varied ways that the assessment is used as part of students' grades across the province creates an issue with regard to the comparability of official EQAO results across schools and school districts. I was surprised by this since from the outset of the project this seemed to me to be an obvious and serious concern. I asked all EQAO participants directly

about the issue of comparability of results during the interviews I conducted. When I asked the consultant, Carol, about this issue she expressed the opinion that “the whole question of comparability across schools never seems to have been something that EQAO has paid a whole lot of attention to...it never seems to have been a huge concern” (Carol, p.3). My interviews with other EQAO personnel also suggest that comparability across schools or districts is not a serious concern for EQAO. While these participants acknowledged that the different ways that teachers use the assessment as part of students’ grades creates different testing conditions as some students might be more motivated than others when they complete the assessment, they also indicated that EQAO is not overly concerned with this issue. A range of reasons for this lack of concern were offered. These reasons are explored in a little more detail to better understand EQAO’s view.

Several participants indicated that as individual school districts interpret their results they are able to do so with the knowledge of whether or not the assessment was used by teachers as part of students’ grades. These participants indicated that it is up to districts to be aware of teachers’ use of the assessment rather than EQAO. For example, one participant states that when districts interpret their official EQAO scores they can consider “if they are low because they didn’t motivate the kids then they have to decide if they want to motivate the kids ” (Marie, p.7). Another participant pointed out that comparability across schools and districts and the different testing conditions created by teachers’ use of the assessment in students’ grades are not serious concerns because EQAO does not encourage the use of their results to rank schools (Susan, p.4). In addition, three participants expressed the view that EQAO is more concerned with other aspects of comparability such as ensuring that the

results of the assessment can be compared from year-to-year rather than with comparability across schools or boards.

Further evidence of the lack of concern with comparability across schools and boards is found in the statement that the EQAO Grade 9 assessment provides “provides reliable and comparable year-to-year data on student achievement” (EQAO, 2009c, p.6). This quote suggests that the focus is on comparing results from year to year by equating subsequent versions of the assessment rather than comparing results from school to school. In addition, the framework document includes a chapter entitled “Maintaining Comparability” which refers only to the issue of comparability from year to year. I suggest that EQAO’s lack of concern with comparability across schools and boards is also found in the fact that teachers’ use of the assessment as part of students grades has been permitted since the first administration of the assessment in 2001 but the only research EQAO has conducted on this practice is the Pang (2004) study described in chapter 4 which attempted to investigate the impact of telling students that the assessment would be used in their grades. As noted by the EQAO director, EQAO has not done any studies on the impact of the different weightings of teacher-derived scores in students’ grades on students’ motivation or on the official EQAO scores (personal communication, October 16, 2008). Given that EQAO has permitted this practice to occur for nearly a decade but has conducted very little research on this practice, comparability across schools and districts is not a central concern for EQAO.

While all EQAO personnel acknowledged that the purpose of including the assessment in students’ grades was to ensure that students take the test seriously and they also seemed to feel this was an effective way to motivate students, none was particularly concerned about the impact of this practice on the comparability of scores across schools or

districts. For me, this lack of concern remains a puzzle especially in light of the fact that many other aspects of the administration of the Grade 9 assessment are standardized. I suggest that further research into this issue is needed.

Research Question 4 – Identifying Interactions

Is there evidence to suggest that there are interactions between teachers' use of the assessment as part of students' grades and EQAO's use of the assessment for accountability purposes?

In this study I define an interaction as the situation created when a practice associated with one use of the assessment has an impact on the validity of the interpretations that can be made for the other use(s). In an effort to find evidence of interactions, I reviewed the observations from all aspects of the case study focusing on EQAO's policy and perspective regarding this use as well as the procedures and perspectives indicated by teachers. Many of the interactions which I uncovered were evident in the tensions that participants described. Some tensions that were described principally involve only one use. For example, the challenges teachers face in marking the assessments in the time available before returning them to EQAO is a problem for teachers but does not have an impact on the use of the scores for accountability purposes. However, other tensions directly impact both uses of the assessment and these are the tensions I refer to as interactions. I found evidence of at least four ways that the use of the assessment items as part of students' grades interacts with the use of the assessment for accountability purposes. Each interaction is explained in this section.

Interaction 1: Impact of Differences in Weighting on Official EQAO Scores

As noted, participants from both the case studies and the EQAO interviews unanimously indicated that the main reason for using the assessment as part of students'

grades was to ensure that students took the assessment more seriously. They felt that students would be more motivated as they completed the assessment if they knew that it would count toward their grades. The case study results also clearly demonstrate that the weighting of the EQAO assessment items as part of students' grades varies considerably from schools where the assessment is not being counted at all to schools where the assessment counts for as much as 20% of a student's final grade. The differences in weighting observed in this study provide one example of the way that the multiple-use of the assessment as part of students' grades and for accountability may interact. Students in schools which weight the assessment items more heavily may have increased motivation to complete the items than students where the assessment counts for very little or not at all. In effect, allowing the teachers the option of including the assessment items as part of students' grades creates non-standard conditions of administration for this large-scale assessment. The differences in weighting observed in this study may seriously impact the comparability of results across schools and boards. No studies of the impact of including the assessment in students' grades or of differences in the weighting on students' performance have been conducted so that the implications of this interaction on the validity of inferences made from the assessment are not clear. Further, as mentioned in the previous section, at present, EQAO does not seem to be particularly concerned with the comparability of results from school to school or board to board.

Interaction 2: Scoring of Open-response Items

The scoring of open-response items on the EQAO assessment is another example of an interaction between the two uses of the assessment. As described in the previous section, some teachers expressed their concern with marking open-response items. They

felt that the most appropriate way to mark these items was to be consistent with the way they marked similar items on unit tests during their courses. However, some of the teachers were also aware that their marking approach was not necessarily aligned with the generic marking criteria at EQAO or with the item-specific scoring guides EQAO develops. A teacher at Birch Park Secondary School explained this issue as follows:

I have shown the students what [EQAO] says is a Level 4 answer and one of the big things that has come out is that the student can answer the long answer question and get the wrong answer and still get a Level 4. And so, me personally, I do put a little bit more emphasis on the right answer . . . when I go to mark it they have to be correct. (Shannon interview, p. 2)

Shannon went on to explain that if a student has the correct answer, she decides if there is enough detail and explanation in their response for the student to receive a Level 4. However, in Shannon's marking criteria, if a student does not have the correct answer they cannot receive a Level 4. While she feels that this approach, which is the same one she uses for unit tests, is the most appropriate way of marking the open-response items on the EQAO assessment, Shannon also indicated that she is aware that her approach creates confusion for students:

My students are now confused as to what is the difference between a Level 3 and a Level 4. So they are always saying "well, I don't really have to get the right answer" [to get a Level 4 in the EQAO marking approach] and I said be careful though because they [EQAO] are not saying "oh, you've got the wrong answer, we'll give you a Level 4". They're saying "you've got the wrong answer but there's a lot of stuff in there that we can see you know what you are doing". And so it's really hard for me to explain to the students the concept that you can get the wrong answer but still show you know what you're doing. (Shannon interview, p.4)

Shannon's description of the challenges she encounters in marking the open-response items suggests how the scoring of open-response items creates an interaction between the two uses of the assessment. That is, wherever teachers use an approach to

marking the open-response items which differs from that used by EQAO, students must decide if they should structure their response so that it conforms to the teachers' expectations or so that it is aligned with the marking criteria for EQAO. If students adjust the way they write their response to the open-response items to conform to teacher's expectations then they may score lower on the official EQAO grading. Alternatively, if they structure their response to be consistent with the EQAO criteria, they may score lower on the teacher-derived mark which is the one that counts toward their grades. In this way then, the marking approach used by teachers when they include the assessment items as part of students' grades can have an impact on the official EQAO scores and, as with the previous interaction I described, no empirical studies have been conducted to investigate the implications of this interaction.

Interaction 3: Discrepancies between EQAO Scores and Teacher-derived Scores

The case study observations demonstrate that teacher-derived scores and official EQAO scores may differ and that this discrepancy may result from a number of sources. Teachers typically select only some of the items to mark whereas the EQAO score includes all operational items on the assessment. At the same time, teachers may be including field-test items in their marking whereas these items are excluded from the official EQAO scores. In addition, many teachers reported that they do not include items that they feel they did not cover adequately in class or that are worded in a manner that is inconsistent with the way they approached the mathematics content in their teaching. The teacher-derived and official scores may also differ because teachers develop their own scoring approach for the open-response items.

The teacher-derived EQAO scores are available to teachers, students and parents several months before they receive the official EQAO scores. Thus, as noted by one of the EQAO personnel interviewed, the teacher-derived scores create an expectation of what the official scores will be. When there is a discrepancy between the two scores, the teacher-derived scores may undermine or alter the perceived meaning of the EQAO official scores. In addition, because all test materials must be returned to EQAO and teachers cannot photocopy the items, students' responses, or the marking key they have created, teachers cannot provide detailed feedback to students or retain documentation of students' work to show to parents. This means that they do not have the information they need to explain discrepancies between the two scores. Adding to the problem, EQAO has no knowledge of the scoring practices which classroom teachers use so that they also do not have the necessary information to explain the discrepancy between the scores. This discrepancy between the two scores is an example of an interaction between the two uses of the assessment because the use as part of students' grades has an effect on the way the official EQAO scores are understood by students, teachers and parents. Thus, activities associated with one use have an impact on the validity of the interpretations associated with the other use.

A series of comments made by Mitchell, one of the EQAO personnel I interviewed illustrate some aspects of this interaction. The first quotation describes how the difference between teacher-derived and official EQAO scores may undermine teachers' understanding of the meaning of EQAO scores. The second quotation suggests how students' and parents' understanding of the meaning of the official EQAO scores may be impacted.

One good reason why teachers would like to use this as part of their marks is because it gives them right away an idea of how their students have performed on the test even though it's not necessarily the same way it is going to be marked during the summer. It gives a good idea but that's a little dangerous also because usually what happens is that you get expectations by doing that . . . and then when the [official] results come you look at the results and say "wow, I thought they did much better than that" . . . and you might complain, you might argue, you might not accept it, you might say EQAO is not fair and that means a lot. So, that's tricky because if we encourage teachers to mark they initially don't have the possibility to mark it the same way we do so they will mark it in a subjective way, they will mark it by themselves and they will have some expectation of what their classroom has done. Then, when they see the result, that's the danger. (Mitchell interview, p.13)

And sometimes also what's a little dangerous is that you [the teacher] say that to the student and you say that to the parents and then they expect their child to have performed well on the test and then when the result arrives they say well, there are two possibilities the teacher didn't do a good job or didn't assess it properly or EQAO was too, they're really out of it. So there is a little thing there that might deserve a little attention. (Mitchell interview, p.14)

Mitchell's description of the impact of the discrepancy between teacher-derived and official EQAO scores suggests some implications of this interaction but as with the other interactions discussed in this section, no research into this issue has been conducted.

Interaction 4: Properties of Large-scale Assessment Items

As noted in previous sections, the EQAO Grade 9 assessment items have been developed for the intended purpose of the assessment which is as part of Ontario's accountability program. However, large-scale assessments have a number of characteristics which differ from classroom assessments and these differences can result in interactions when one assessment is used both as part of the accountability program and for classroom assessment purposes. Three examples of this sort of interaction emerged from this case study and are described in this section.

The first characteristic of large-scale assessment items which is at odds with classroom assessment is the need to keep large-scale assessment items secure. This feature of large-scale assessments precludes teachers from providing students with feedback on the test items they have completed. However, providing students with feedback is a key feature of effective classroom assessment, particularly from a sociocultural perspective (see chapter 1). This issue is described by Mitchell, one of the EQAO personnel I interviewed, as follows:

If you are using it [the EQAO assessment] as a part of the final exam, of course, students would like to have feedback, they would like to know what they got right, what they got wrong and why . . . but if you look at the Assessment Guide that we [EQAO] give to teachers it says that they are not allowed to discuss any of the items because some of these items are repeated and might come back in another assessment and so we don't want them to talk about it or even remember it. So that's one thing that goes into contradiction with the fact that teachers can use it as a classroom assessment. (Mitchell interview, p.5-6)

The two uses of the assessment interact in the sense that the use of the assessment for accountability purposes requires specific assessment practices which undermine the use of the assessment for classroom assessment purposes.

A second characteristic of large-scale assessment which differs from classroom assessments and leads to an interaction between the two uses is item types. The item types which are currently used on the EQAO assessment are multiple-choice and open-response because these item types can be administered in a reasonable amount of time and can be marked reliably. However, as noted by several participants, the use of multiple-choice items for classroom assessment is problematic. These items provide less specific information about what students know and can do and Grade 9 students frequently have difficulty responding to these items. Thus, the item types required by EQAO in the use of the assessment for accountability purposes may result in a less useful and less accurate

classroom assessment. The validity of the use of the assessment as part of students' grades is impacted by the item types required for the use of the assessment for accountability.

Another characteristic of large-scale assessment items which is problematic is that EQAO test developers typically do not include items which all or nearly all students are expected to get correct (Marie interview, p. 7). Rather, items are included on large-scale assessments because they discriminate among test-takers. However, as Marie, one of the EQAO personnel I interviewed, notes teachers often include items on classroom assessments that they expect all or nearly all students to answer correctly and this is considered to be good classroom assessment practice. Marie points out that while large-scale assessments seek to discriminate among students, classroom assessments often do not have this goal. She identified this as a tension resulting from teachers' use of the items as part of students' grades. Should the assessment developers include more items which can be successfully completed by all students in recognition of the fact that many teachers will use these items as part of their students' grades? Alternatively, teachers might choose not to include in their teacher-derived scores, those items that most students do not get correct. In either case, there is an interaction between the two uses because item characteristics suited to the use of the assessment for accountability have an impact on the validity of the interpretations that can be made for the use of the assessment as part of students' grades.

Discussion of Case Study Observations

In this chapter I have addressed the first four research questions in this inquiry. In doing so, I have presented a considerable volume of observations regarding various aspects of the multiple-use of the EQAO Grade 9 assessment. Before moving on to the two remaining research questions which focus on the process of validation for the EQAO Grade

9 assessment, I provide a brief overview of the observations in this chapter and discuss the importance of these observations for this inquiry.

In the first part of this chapter, information from the document analysis and interviews with EQAO personnel was used to describe the pattern of uses associated with the EQAO Grade 9 assessment. These data sources demonstrate that each administration of the EQAO Grade 9 assessment is used for many different purposes and that the practice of multiple-use is taking place for the EQAO Grade 9 assessment. I focus on two of these uses (i.e. the use of the assessment as part of Ontario's accountability program and the use of the assessment by teachers as part of their students' grades) to investigate the implications of the multiple-use of this assessment for the process of validation. To begin to understand the implications of these two multiple-uses, I developed a detailed description of how teachers' use of the assessment as part of students' grades takes place. The results from my province-wide questionnaire and two other studies provide evidence that the use of the assessment as part of students' grades occurs quite commonly but there are schools within various districts where this practice is not taking place. This finding is important in that it provides evidence that the conditions under which the EQAO assessment is administered are not consistent across all schools in the province. That is, in some schools the assessment counts for students' grades and in other schools the assessment has no such consequences for students.

Additional details regarding how teachers' use of the EQAO Grade 9 assessment takes place emerged from the school-level case studies which were conducted. The observations from these school-level case studies provide evidence that there is considerable variation in the weighting, item selection and scoring procedures teachers use and that the variations in the procedures teachers use take place at all levels of the inquiry including

across students within a classroom, across classes within a school, across schools within a district and across districts in the province. The observations show that the weighting of the EQAO assessment as part of students' grades ranges from schools that are not counting the assessment at all to schools where the assessment is being used in lieu of a final exam and counts for as much as 20% of students' final course grades, with many variations in weighting between these two extremes. With regard to item selection, I found examples of teachers using only the multiple-choice items, a selection of both multiple-choice and open-response items and only the open-response items. I also found a range of approaches to scoring the EQAO items. The observations from the school-level case studies also suggest some of the factors which influence teachers in their use of the EQAO Grade 9 assessment such as district policy, class size, the number of Grade 9 courses a teacher is teaching etc.

The degree of variation observed in teachers' practices is an important finding in that it provides further evidence that the conditions under which the EQAO assessment is administered are not consistent across the province. Together with the observations from the questionnaire, this study provides clear evidence that the EQAO Grade 9 assessment results cannot be considered to be comparable across classrooms and schools in the province. These observations also suggest that teachers have developed procedures for including the assessment as part of their students' grades to make the assessment more appropriate for their own classroom setting but these procedures may create difficulties for the use of the assessment for accountability purposes.

In addition to the observations about the procedures teachers use, in this chapter I present my observations of teachers' views of using the EQAO assessment as part of their students' grades. While ascertaining the meaning which teachers ascribe to the teacher-

derived and official EQAO scores is quite challenging, my observations clearly indicate that these teachers have decided to use the assessment as part of their students' grades in an effort to motivate students to put more effort into completing the assessment items. However, these observations also show that teachers struggle with deciding how much to weight the assessment. Another important finding with regard to participants' views of this practice is that most of the teachers, department heads and principals who participated in this case study felt that motivating students to complete the assessment was particularly important for students in the Applied course but, at the same time, these participants indicated that they felt that the use of the assessment as part of students' grades is not appropriate in Applied courses because the students taking these courses often struggle with completing pencil-and-paper tests. In addition to this challenge, a number of other tensions identified by participants in the school-level case studies are described.

With regard to EQAO's perspective on this practice, I present evidence that EQAO has permitted teachers to use the assessment as part of students' grades from the first operational year³² of the assessment program. I also recount my findings with regard to why EQAO has permitted this use of the assessment to take place. That is, both teachers and EQAO feel that it is important to motivate students to take the assessment seriously and that including the assessment as part of students' grades is one way to achieve this. However, EQAO is not in a position to mandate that teachers include the assessment as part of their students' grades. Thus, teachers are given the option of including the assessment in grades but cannot be required by EQAO to do so. Giving teachers the option of including the assessment in students' grades creates a number of tensions particularly as EQAO does not

³² The first operational year of the assessment program was 2001, which is when the assessment results were first reported to the public. The assessment was not used as part of students' grades during the years when pilot tests were administered.

provide teachers with item-specific marking guides. In addition, while providing teachers with the option of including the assessment in students' grades creates non-standard testing conditions, my interviews and inquiries with EQAO provide evidence that EQAO is not particularly concerned with the comparability of results across schools and that no studies of the impact of this practice had been conducted by EQAO at the time of this inquiry.

The detailed description of teachers' use of the EQAO assessment as part of students' grades which is presented in this chapter illustrates many of the tensions which may be encountered when a large-scale assessment developed for accountability purposes is used as part of a classroom assessment strategy. For this inquiry, analysing these tensions is particularly valuable as it provides evidence of the kinds of interactions which are taking place between the two uses of this assessment. In particular, in this chapter I provide evidence of four interactions in which the activities associated with one use of the assessment have an impact on the validity of the inferences which can be made for the other use of the assessment. These findings can be directly related to the theoretical discussion of multiple-use which is presented in chapter 2. That is, in chapter 2 I suggested that when multiple-use takes place we cannot assume that the uses are independent of one another. The observations in this chapter show that in the case of the EQAO Grade 9 assessment, interactions between the multiple-uses are taking place. Each of the interactions which is identified suggests an area where additional research needs to be conducted and since the interactions which are taking place undermine the validity of the interpretations which can be made for each use of the EQAO Grade 9 assessment research into these interactions should be included as part of the process of validation for this assessment.

How the process of validation might proceed for the EQAO Grade 9 assessment given these interactions is the focus of the fifth and sixth research questions in this inquiry which are addressed in chapter 7.

CHAPTER 7

VALIDATING THE MULTIPLE-USES OF THE GRADE 9 ASSESSMENT

In this chapter I consider how the process of validation might be undertaken for the two multiple-uses of the EQAO Grade 9 Assessment of Mathematics in light of the interactions between these uses that were identified in the case study. As outlined in chapter 3, I consider the process of validating the multiple-use of the EQAO Grade 9 assessment using two approaches. First, I investigate the adequacy of Kane's (1992, 2006) argument-based model of validation for this instance of multiple-use and then I consider how boundary objects and boundary encounters might contribute to the process of validation for this instance of multiple-use. These two approaches address the fifth and sixth research questions, respectively.

Research Question 5 - Applying Kane's Model

How can Kane's argument-based model of validation be applied to the multiple-use of the EQAO Grade 9 Assessment of Mathematics?

To address this research question, I begin by explaining my rationale for choosing Kane's model for validating this instance of multiple-use. I then describe my attempt to apply the model as well as the challenges which I encountered.

Rationale for Choosing Kane's Model

In choosing a measurement-based model to validate the multiple-use of the EQAO Grade 9 assessment, I had two main criteria. To begin with, I was looking for a model that seemed capable of addressing the practice of multiple-use. In reviewing various approaches to validation, I looked for connections between the characteristics of the model and the practice of multiple-use and tried to envision how the model might be applied to an

assessment where multiple-use is known to occur. My second criterion was to find a model that shares some common ground with the sociocultural view of assessment that I adopt. In this section, I summarize my reasons for choosing Kane's argument-based model of validation by highlighting the characteristics of the model that seem relevant to the multiple-use of the EQAO Grade 9 assessment and by describing those aspects of the model which I see as consistent with a sociocultural view of assessment.

Characteristics of Kane's Model that Relate to Multiple-use

The pragmatic orientation of Kane's model is one reason I chose this approach for validating the multiple-use of the EQAO assessment. Given the complex issues which multiple-use creates, I wanted to find a model that included explicit guidance regarding how to initiate and proceed with the validation process. Kane offers practical guidance for the validation process in his 2006 chapter by providing a thorough description of this approach and including detailed examples of interpretive arguments. In these examples Kane identifies the specific inferences, assumptions, warrants and backing for four different uses of educational assessments. Examining these examples enables researchers to more easily identify the inferences, assumptions, warrants and backing that might be required for their own validation endeavour. The merits of Kane's model with regard to articulating a useful framework for the practical validation of large-scale assessments have also been noted by Moss et al. (2006) and Shepard (1993).

A second characteristic that suggests Kane's model might be appropriate for validating the multiple-use of the EQAO assessment is that his approach centers on the proposed interpretation and use of an assessment whereas other measurement-based models of validation focus on the psychometric characteristics of an assessment (e.g. Lissitz &

Samuelson, 2007). As detailed in chapter 3, for Kane, the validation process begins with specifying the proposed use for an assessment and the interpretive and validity arguments relate directly to that use. Further, Kane indicates that an assessment may have more than one legitimate interpretation or use and notes that “the interpretations and uses of test results are complex, and validation efforts should recognize this complexity” (2006, p.42). For these reasons, Kane’s model has been described as being responsive to differences in proposed uses (Kane, 1992; Moss et al., 2006; Shepard, 1993). This characteristic has obvious relevance for validating multiple-use practices given that multiple-use, by definition, includes two or more uses of an assessment.

Another characteristic of Kane’s model which suggests that it would be appropriate for the multiple-use of the EQAO Grade 9 assessment is that Kane (2006) includes a discussion of the way his approach may be used for validating both large-scale and classroom assessments. Kane’s model is the only measurement-based approach to the process of validation I encountered that claims to be appropriate for both types of assessment. Since the instance of multiple-use for the EQAO Grade 9 assessment I am investigating involves the use of a large-scale assessment by mathematics teachers as part of their students’ grades as well as the use of the assessment by EQAO for accountability purposes, the applicability of Kane’s approach to both types of assessment contributed to my decision to use his approach.

A fourth characteristic that persuaded me to select Kane’s model is that this approach has been fairly broadly endorsed by the measurement community. Kane’s approach to validity as argument was selected to be included as the validity chapter in the most recent edition of *Educational Measurement* which, as noted earlier, is widely acknowledged as an

influential text in the measurement field. In addition, this model has been used in a number of published validation studies. For example, Haertel (1999) used this approach to develop a theoretical validity argument for a fictional assessment with multiple-uses (described in chapter 2) and Llosa (2008) used it to develop a validity argument for a test of English-language learning. These developments suggest that this model may be emerging as the state-of-the-art approach to validation (Brennan, 2006) and I wanted to get a sense of how such an approach might be used for validation in the context of multiple-use.

Coherence of Kane's Model with a Sociocultural View of Assessment

In choosing a model for validating the multiple-use of the EQAO assessment, I was also looking for an approach that is consistent with a sociocultural view of assessment. In my opinion, Kane's model is in keeping with some key aspects of this perspective. Specifically, in Kane's model context is recognized as key, the value-laden nature of assessment is explicitly identified, and validation is seen as a defeasible process rather than an unequivocal proof. A brief explanation of each characteristic is provided in this section.

With regard to context, Kane draws on Toulmin's discussion of argument and notes that the criteria for evaluating an argument are always dependent on the context in which the argument occurs. Kane states "a validity argument evaluates a particular interpretive argument in a particular context" (2006, p.29) and he points out that a validity argument developed for one context cannot be assumed to be applicable in another context. He also emphasizes the importance of context in his discussion of the use of the argument-based model of validation for classroom assessment. He notes that in interpreting classroom assessments, teachers engage in an active search for the meaning of their observations by considering the contextual factors that interact in a classroom and influence a student's

performance (p.47). While Kane seems to use context to signify a bounded setting or a backdrop in which assessment activities take place, he is clearly indicating that validating the inferences made on the basis of large-scale assessments must include a consideration of the particular setting where a large-scale assessment is being used. That is, for Kane, a validity argument is not something which can be standardized across settings, even if the validity argument is developed for a large-scale assessment with standardized procedures for administration, scoring and reporting. Thus, although Kane's use of context does not include the relational notion of contextualization advocated by some sociocultural theorists (as described in chapter 1), in my view, his repeated flagging of the importance of context is consistent with a sociocultural view of assessment.

As noted in the introductory chapter, the value-laden nature of assessment is a key characteristic of sociocultural perspectives of assessment. This aspect of assessment practice is also acknowledged and directly addressed in Kane's model. Kane (2006) provides a number of examples of the value judgments which are inherent in large-scale assessment processes. He describes the way that these value judgments influence the decisions which are made with regard to the development of assessment items, scoring procedures and guidelines for making decisions from assessment results. Kane also notes that the value judgments which are made in large-scale assessment contexts are often implicit and an important part of his model is the explicit identification of these value judgments.

In my opinion, Kane's view of interpretive arguments as tentative and defeasible (as explained in chapter 3) is also consistent with sociocultural perspectives of assessment. From a sociocultural view, assessment is closely intertwined with learning in a process whereby multiple sources of evidence are considered and the interpretations made on the basis of that

evidence evolve (Shepard, 2001; Moss, 2003). Similarly, Kane characterizes validation as a process which draws on multiple sources of evidence to build a validity argument. Kane notes that the process of validation does not unequivocally prove anything and he points out that as more evidence is collected the evaluation of an interpretive argument may change.

Given these characteristics, I decided to attempt to use Kane's argument-based model for validating the multiple-use of the EQAO Grade 9 assessment for accountability purposes and by teachers as part of students' grades.

Applying Kane's Argument-based Model to the EQAO Assessment

The first phase of Kane's (2006) approach to validation, as described in chapter 3, is to develop an interpretive argument. To begin this process, Kane indicates that the intended interpretation or meaning of the assessment must be established so that the relevance of the score interpretation to the proposed use can be considered. After this is done, the inferences and assumptions can be made explicit and the process of articulating the warrants and backing for those assumptions can take place. In attempting to apply Kane's model to the multiple-use of the EQAO Grade 9 assessment, I began with these initial steps.

Establishing the Meaning of EQAO Grade 9 Assessment Scores

For the two uses of the EQAO assessment that are the focus of this study, (i.e. the use for accountability and the use for students' grades) establishing the intended interpretation or meaning of the assessment scores is not a straightforward task and as I attempted to do this I encountered a number of difficulties. To begin with there are actually two EQAO Grade 9 scores which must be considered; the meaning ascribed to the official EQAO score must be determined as well as the meaning ascribed to the teacher-derived score. As observed in chapter 6, official EQAO scores may differ from the teacher-derived scores and the items on

which each of these scores is based may also differ. Accordingly, in this instance of multiple-use two different scores with two different meanings must be considered.

A second challenge in establishing the meaning of the scores in this instance of multiple-use is that different individuals may ascribe different meanings to each of the scores. The document analysis conducted in the case study provides a clear indication of the meaning which EQAO ascribes to the official test scores. The official scores are described as a measure of the knowledge and skills the Ontario curriculum expects students to demonstrate by the end of Grade 9 (EQAO, 2009c). The EQAO documents further qualify this interpretation by specifying that the connection between the scores and the curriculum is limited to those aspects of mathematics that can be assessed on a large-scale, pencil-and-paper assessment. However, the meaning which teachers, parents, and students ascribe to the official scores may not be the same as the meaning which EQAO ascribes.

With regard to the teacher-derived scores, establishing the meaning is even more problematic. As observed in the case study, individual teachers select the items that they include as part of students' grades according to a number of criteria. As a result the meaning ascribed to the teacher-derived scores varies from teacher to teacher and is very much locally determined. For example, as described in chapter 6, several teachers indicated that they felt that different meanings were associated with the Academic and Applied EQAO scores and others describe the way that they chose to mark specific EQAO items to gather assessment information for particular aspects of the curriculum such as geometry or their students' ability to apply mathematics ideas to new contexts. Further, the meaning of the teacher-derived EQAO scores is difficult for others such as parents or EQAO to get a sense of

because they have no knowledge of which items the score is based on or of the approach that was used to mark the items.

The differences in establishing the meaning of scores for large-scale accountability assessments and for classroom assessments are clearly identified in Kane's model (2006). As described in chapter 3, Kane indicates that for an accountability assessment an initial *semantic interpretation* of each student's performance is made. By semantic interpretation, Kane specifies that he is referring to the meaning assigned to the assessment results in terms of the standards or curriculum objectives on which the assessment is based. For the EQAO Grade 9 assessment the semantic interpretation of the official EQAO scores is clearly identified in the EQAO documents in terms of the curriculum expectations that are assessed by the items. However, as Kane explains, teachers do not begin with this sort of semantic interpretation to establish the meaning of a score. Rather, teachers use their own conceptual frameworks, including their views concerning the subject area, appropriate pedagogical theories and effective instructional techniques, as well as their experience to integrate the available evidence into a coherent view of their students (2006, p. 51). Kane suggests that this process is an iterative one with the teacher developing an initial view of their students based on their conceptual framework and on the evidence they gather as they interact with students during instruction and assessment activities and then refining that view based on ongoing interactions. In this way, Kane sees the meaning teachers ascribe to classroom assessments as dynamic and situated. My observations of the meanings teachers ascribe to the teacher-derived EQAO scores are consistent with this view. In fact, in retrospect, I suspect that this was the main reason for my difficulty in conducting my analysis of the meaning which teachers ascribe to the EQAO assessment scores (as described in chapter 6).

Given these observations, one of the challenges I encountered in validating the two multiple-uses of the EQAO Grade 9 assessment was trying to reconcile the semantic interpretation of the official EQAO scores with the dynamic and situated meaning ascribed to the teacher-derived scores in order to develop the interpretive argument.

A third difficulty which arises in attempting to ascertain the meaning of the official and teacher-derived EQAO Grade 9 assessment scores is that the meaning of each score may actually change as a result of the other use taking place. That is, the meaning which would be ascribed to the official score would differ on the basis of whether or not the assessment is also being used as part of students' grades. For example, the meaning associated with the official EQAO scores in jurisdictions where the assessment is not included in students' grades may differ from the meaning that can be ascribed to official EQAO scores in jurisdictions where the assessment counts as 20% of students' grades. In this way, the interaction between the two uses adds considerably to the difficulty of establishing the meaning of either score.

Identifying Inferences and Assumptions

Unpacking the intended interpretation or meaning of the scores is only the preliminary step in creating an interpretive argument. According to Kane, the series of inferences which are inherent in each use must be made explicit and the assumptions within each inference must be identified. In my attempt to begin this process for the two uses of the EQAO Grade 9 assessment, I re-visited the examples of interpretive arguments Kane (2006) provides for an accountability assessment and for a classroom assessment. These examples suggest some of the sources of evidence that would need to be collected to establish the warrants and backing for each use. While I found these examples very helpful in suggesting

how an interpretive argument might be developed for the use of the EQAO Grade 9 assessment for accountability *or* for the use as part of students' grades, they do not provide any guidance regarding how to specify inferences and develop an interpretive argument when two uses interact. For the EQAO Grade 9 assessment, I found that as I attempted to specify the inferences inherent in one of the uses, it became evident that those inferences and the assumptions supporting them would be different given the existence of the other use. In other words, the inferences that are inherent in using the assessment for accountability purposes are different than the inferences which are inherent if a single administration of the assessment is being used both for accountability *and* as part of students' grades. Thus, I found that making the inferences and assumptions explicit was problematic in this instance of multiple-use because I was unable to find a means of accounting for the impact of one use on another within the interpretive argument.

The difficulties in establishing meaning and in specifying the inferences and assumptions for the interpretive argument for the EQAO Grade 9 assessment preclude the development of a validity argument for this instance of multiple-use because, in Kane's model, the validity argument can only be developed after the interpretive argument has been established.

Challenge of Using Kane's Model for Validating Multiple-use

Kane's (2006) approach to validation is based on making the interpretations and uses of an assessment explicit and laying out the inferences and assumptions inherent in each use. My study demonstrates that the practice of multiple-use is difficult to address using this approach because the model ostensibly treats each use as independent of the other uses. Nonetheless, Kane's model does specify some parameters that could be applied to the

practice of multiple-use. In particular, the defeasible nature of the interpretive argument and the assumption of *ceteris paribus*, both of which are described in chapter 3, could be used to account for multiple-use. In this section I explain how each of these characteristics could be used to address the practice of multiple-use.

Kane (2006) maintains that an interpretive argument that has survived all reasonable challenges to its coherence can be provisionally accepted but he points out that new evidence may undermine the credibility of an interpretive argument at any point in time. As noted in chapter 3, this is what he means in referring to an interpretive argument as defeasible. In a sense, the multiple-use of the EQAO Grade 9 assessment may be viewed as a practice that undermines the credibility of the established interpretive argument. That is, teachers' use of the assessment as part of students' grades may undermine the credibility of the interpretive argument for the use of the assessment by EQAO for accountability. Following this approach, when validating an assessment with multiple-uses one could indicate that whenever there is evidence of interactions arising from the multiple-uses, the interpretive argument for the intended use may not hold. This sort of condition could be explicitly identified as a limitation on the interpretive argument for the intended use.

Another way of addressing the practice of multiple-use using Kane's model is by using the assumption of *ceteris paribus* (i.e. all else being equal). As noted in chapter 3, Kane (2006) sees this assumption as applicable to all validity arguments. According to Kane, as part of the validation process the exceptions or conditions where the inferences do not hold must be specified. Using this approach, separate interpretive arguments would be developed for each use. For multiple-uses where evidence of interactions between uses is found, the assumption of *ceteris paribus* would not hold; all else is not equal because one use

interferes with the interpretations that are made for the other use and in these cases, the interpretive argument would be overturned. This approach would effectively mean that the interpretive argument would be overturned for all multiple-uses where interactions between uses are known to take place. For the EQAO Grade 9 assessment, invoking the assumption of *ceteris paribus* would mean that the interpretive argument for the use of the assessment for accountability would not hold because of the interactions between the use for accountability and the use as part of students' grades.

To validate assessments where multiple-use takes place using Kane's approach, the defeasible nature of the interpretive argument and/or the assumption of *ceteris paribus* would provide ways of overturning the validity argument whenever interactions between uses are known to occur. Essentially this approach amounts to stating that whenever multiple interacting uses are known to occur, the validity argument may not hold. Adopting such an approach is arguably preferable to current measurement-based approaches to the process of validation in that the potential for interactions among multiple-uses is at least being acknowledged. However, given the prevalence of the multiple-use of large-scale assessments described earlier and the likelihood of interactions among multiple-uses taking place, I do not think that invoking these aspects of Kane's approach within an argument-based model is an adequate way to approach the process of validation. I am left wondering if there are ways that the argument-based approach to validation might be expanded or revised to more effectively address the practice of multiple-use? Alternatively, what other approaches to validation might be considered for large-scale assessments where the practice of multiple-use is known to occur? In the next section, I begin to address this issue by considering the contributions to the process of validation that may come from the use of boundary concepts.

Research Question 6 – Using Boundary Concepts

In what ways can the concepts of boundary objects and boundary encounters contribute to the process of validation given the multiple-use of the EQAO Grade 9 Assessment of Mathematics?

In this section, I describe my application of the concepts of boundary objects and encounters to the multiple-use of the EQAO Grade 9 assessment. I begin by providing a description of these concepts and explaining my rationale for using these concepts to analyse this instance of multiple-use. I then describe my analysis of the school-level case studies, EQAO documents and EQAO interviews which provides evidence that the EQAO Grade 9 assessment functions as a boundary object when teachers use it as part of their students' grades. I also identify some boundary encounters associated with this assessment. In the final section of this chapter I discuss how these analytic tools contribute to the process of validation for the EQAO Grade 9 assessment and might be beneficial for other instances of multiple-use.

Boundary Objects and Encounters

As noted in chapter 3 where I introduced these concepts, boundary objects (Bowker & Star, 1999; Star & Griesemer, 1989) and boundary encounters (Cobb et al., 2003) are analytic tools which are situated within sociocultural theory. The concept of a boundary object³³ was developed by Star and Griesemer (1989) as they examined the ways that different groups of people work co-operatively toward a common goal while, at the same time, maintaining their own distinctive perspectives and agendas. Star and Griesemer first identified the phenomenon of a boundary object as they were considering the different meanings that amateur bird watchers and professional biologists associated with a collection

³³ In Bowker and Star (1999), credit for coining the term *boundary object* is given to Star.

of bird specimens as the two groups worked collaboratively toward the establishment of a natural history museum. The same specimens were being used by the two groups and Star and Griesemer noticed that while the specimens had separate meanings for each group they also became shared objects with a common meaning. Moreover, Star and Griesemer found that the common meaning associated with the specimens served to facilitate communication across the two groups. In further developing this idea, Star and Griesemer observed that different groups of people often share tools, artifacts or techniques and that these shared objects can have both a local meaning for each community as well as a meaning which is shared by both communities. They describe boundary objects as “both plastic enough to adapt to local needs and the constraints of the several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly structured in common use, and become strongly structured in individual-site use” (Star & Griesemer, p.393).

The term boundary object can be used to refer to a concrete object but this term may also signify an abstract idea. For example, Bowker and Star (1999) maintain that systems of classification are boundary objects in the sense that classifications are typically devised to meet the needs of one or more distinct and identifiable communities while at the same time functioning as “objects of cooperation across social worlds” (p.15). Whether concrete or abstract, boundary objects arise over time as two or more communities work cooperatively (Bowker & Star, 1999). As conceived by Bowker and Star, boundary objects emerge over time and once they have emerged they are relatively stable but are typically not permanent. That is, while these objects do not have a fleeting or temporary function, they can, as circumstances change, cease to function as a boundary object. In addition, because they are shared objects, boundary objects act as bridges or translators across different groups of

people who are working together. As Bowker and Star observe, “boundary objects are one way that the tension between divergent viewpoints may be managed” (1999, p. 292).

However, it is important to recognize that boundary objects enable communication across different communities but do not require consensus among those communities. A boundary object is a “working arrangement” (p. 297) that does not impose the views of one group on the other. For this reason, Bowker and Star (1999) suggest that boundary objects are particularly useful for analysing cooperative situations where the status of the groups who share the object is relatively equal.

Cobb et al. (2003) suggest the related concept of a boundary encounter in which members of different communities engage in activities together and these activities function to bridge gaps in interpretation and meaning between the communities. Boundary encounters have the same characteristics as boundary objects but refer specifically to events or activities rather than concrete objects or abstract ideas that are shared by two or more communities.

Rationale for Using Boundary Concepts in this Inquiry

While the notion of boundary objects originated in sociology and information systems theory, both boundary objects and encounters are concepts which are increasingly being used by educational researchers (Cobb et al., 2003). For example, these concepts have been used in theorizing aspects of teaching and learning in mathematics and science education research (Buxton, Carlone, & Carlone, 2005; Venkat & Adler, 2008) and in examining practices in the area of lifelong learning (Edwards, 2005). As noted in chapter 3, I first encountered the notion of boundary objects in reading the work of Moss et al. (2006) who suggest boundary objects as a tool which might be used in validating large-scale assessment practices. Moss maintains that “a large-scale assessment would function as a

boundary object when actors in the local context are able to cooperate in providing necessary information to outsiders while maintaining a productive level of authority and agency over their own practice” (Mislevy et al., 2009, p.97). The similarities I found between Moss’s description of how a large-scale assessment might function as a boundary object and the activities surrounding the multiple-use of the EQAO Grade 9 assessment suggested to me that using this analytic tool as part of the validation process might be worthwhile. To explore the idea that the EQAO assessment functions as a boundary object and consider whether some of the activities which EQAO and teachers engage in may be considered to be boundary encounters, I decided to look for evidence of these concepts in my data sources. The process which I followed to do this is described in the following section.

A Boundary Concept Analysis of the EQAO Grade 9 Assessment Observations

Identifying Groups and Goals

Boundary objects and encounters, by definition, relate to a situation where two or more groups each have their own agenda but also share a common goal. Given this definition, in order to use boundary concepts as an analytic tool for the EQAO Grade 9 assessment I began by identifying two groups, specifying a common goal for these two groups and identifying some aspects of the separate agendas of each group. For the two multiple-uses of the EQAO Grade 9 assessment which are the focus of my inquiry, I consider classroom teachers as one group and EQAO as a second group. Many other groups with an interest in this assessment could be identified such as school administrators, parents or students but since the information I gathered in my study was primarily from teachers and EQAO personnel these are the two groups I focused on for this analysis. In addition, as noted in chapter 6, many of the participants in this study have belonged to both of these

groups at different points in time. That is, two of the department heads have participated in EQAO Grade 9 item development and scoring activities and four of the EQAO personnel were classroom teachers who had had the opportunity to use the assessment as part of their own students' grades. Thus, while the two groups can be seen as distinct in some senses, they also overlap and membership in each group changes over time.

In terms of a common goal, on the basis of my observations I suggest that both groups are interested in accurately measuring the extent to which students have mastered the Grade 9 mathematics curriculum. This goal is explicitly stated by EQAO in two sentences in the framework document: "The purpose of the Grade 9 Assessment of Mathematics is to assess the level at which students are meeting curriculum expectations in mathematics up to the end of Grade 9" and "EQAO provides the Ontario school system with valid, reliable and comparable year-to-year data on student achievement" (EQAO 2009c, p.6). Thus, EQAO intends the Grade 9 assessment to be a means of accurately measuring students' mastery of the Grade 9 mathematics curriculum.³⁴ Evidence that teachers share the goal of accurately measuring students' understanding of the mathematics curriculum is suggested in the comments of several teachers who participated in the school-level case studies. For example, with reference to including the assessment in students' grades, Adrian states:

When they know that it's for marks they tend to take it more seriously and I think that guarantees a better overall indication of the students' performance. Because it could be that a student who is achieving well just decides not to take it seriously and doesn't get a true result on the EQAO test.

(Adrian interview, p.2)

³⁴ As noted earlier, EQAO does clearly indicate that aspects of the mathematics curriculum such as expectations which require investigations are not measured by the items on the Grade 9 assessment.

Adrian's comment, along with similar comments made by other participants, suggests that teachers also want to get an accurate measure of the extent to which students have mastered the Grade 9 mathematics curriculum.

According to Star and Griesemer (1989), while the groups using a boundary object have a shared goal, each group also maintains their own distinct agenda. For the EQAO Grade 9 assessment, as demonstrated in chapter 6, EQAO's agenda includes the intended use of the assessment for accountability purposes as well as the other benefits they identify such as providing data for improvement planning and target setting. My school-level case study observations suggest that teachers also have an agenda as they use the Grade 9 assessment items in their students' grades. For example, Diane, the department head at Taylor Brook High School, indicates that her department is trying to increase the EQAO Grade 9 Applied assessment scores in an effort to safeguard the school's academic reputation. She explains:

This school has a reputation for being an academic school and so EQAO scores that are low attached to [Taylor Brook High School] look bad on paper . . . so they needed something to be done. All I could do was come in and say 'okay, I know what the curriculum is supposed to look like, I know what it's supposed to sound like, I know what it's supposed to feel like. Let me give it a try and see if we can move our EQAO scores in a different direction' . . . and so I've made a start . . . and I need to keep doing it . . . and if they're still not improving [the EQAO scores] then frankly I have to rethink a lot of what I'm doing because they're judging the program by EQAO scores and I know it's just one measure, but it is a measure that's fairly important.

(Diane interview, p.4)

Diane's comment on her efforts to improve the EQAO scores for students taking Applied mathematics at Taylor Brook not only illustrates some of the reasons teachers might have for using the assessment as part of students' grades, it also suggests the agendas that school administrators and school districts might have for this practice.

Given the observations from this case study, I think an argument can be made that there are at least two groups using the EQAO Grade 9 assessment who share a common goal of obtaining an accurate measure of the extent to which students have mastered the Grade 9 mathematics curriculum. In addition, each group has their own agenda with regard to the use of the Grade 9 assessment as part of students' grades. Having established these parameters, I set about looking for evidence that the EQAO assessment functions as a boundary object for these two groups.

Finding Evidence of the EQAO Assessment as a Boundary Object

For Star and Griesemer (1989), an object is considered a boundary object if it is used by two or more groups and the use of it helps each group gain a better understanding of the other group's perspective with regard to their shared goal. In this section I provide some evidence to support the contention that teachers' use of the assessment as part of students' grades effectively turns the EQAO Grade 9 assessment into a boundary object. Before presenting this evidence, I summarize the analysis process I used.

To find evidence that the EQAO assessment functions as a boundary object when it is used as part of students' grades, I analysed the transcripts from the school-level case studies and the interviews with EQAO personnel. To conduct this analysis, I read the transcripts from each participant to identify those comments which suggest that teachers' use of the assessment items as part of students' grades helped either teachers or EQAO to better understand the view of the other group. As was done for the analyses described in chapter 5, I created a table to record these observations with a separate entry for each comment. Each entry included the participant's pseudonym, a summary of the comment and/or cut-and-pasted text from the transcript as well as the page number where the comment is located. In

this way I summarized the observations related to the concept of a boundary object across the three school-level case studies and the EQAO interviews. Through this process I was able to find a number of ways that teachers' use of the assessment as part of students' grades leads to the EQAO Grade 9 assessment functioning as a boundary object. In this section I explain three aspects of teachers' use of the assessment that emerged from this analysis: teachers' increased familiarity with EQAO items, creating marking guides for open-response items, and the emergence of the assessment as a boundary object over time.

Teachers' increased familiarity with EQAO items.

The school-level case study observations show that teachers carefully examine the items on the EQAO Grade 9 assessment when they choose which items they will mark to include in their students' grades. In one school this process was done individually, while in the other two schools the process was collaborative with teachers meeting to discuss the items on the assessment. In all three schools, the examination of the EQAO assessment to choose items to include in students' marks makes teachers much more aware of the items on the assessment and the way those items connect to the curriculum expectations. As Margot states "sometimes we [mathematics teachers in the department] have to read over a question two or three times" to understand the intent of the item (Margot interview, p.1). I suggest that this increased knowledge of the EQAO Grade 9 assessment would not occur if teachers were not marking the assessment items. That is, if teachers were not including the assessment in their students' marks they would be more likely to simply bundle up the completed assessments and send them to EQAO without spending much time looking at the items or their students' responses to those items. Some evidence supporting the contention that teachers become more familiar with the items because they are marking them is found in

comments made by participating teachers. For example, Whitney explains that she only looks at her student's responses to those EQAO items that the teachers in her department have decided to mark. She states "I never look at the other items . . . I mean, I guess I could take the time to do that but I haven't" (Whitney interview, p.6).

As these teachers examine the items on the assessment, they question how the items align with their understanding of the curriculum as well as the extent to which the items enable students to demonstrate their understanding of mathematics concepts. For example, Whitney states "this year there was a question on it [the EQAO Grade 9 assessment] that isn't even in our curriculum so we didn't count it as part of what we are marking" (Whitney interview, p.5). Another example of this are the many comments made by participating teachers that the multiple-choice items on the assessment do not provide a particularly good indication of students' understanding of the mathematics concepts in the curriculum. These teachers are more aware of the items on the assessment as well as the connection between these items and the curriculum because they are selecting items to mark, even if they do not agree with EQAO with regard to the value of those items.

These observations demonstrate that teachers' use of the assessment as part of students' grades turns the Grade 9 assessment into a boundary object in that teachers get a better understanding of the efforts which are made by EQAO to get an accurate measure of students' understanding of the Grade 9 mathematics curriculum. While teachers have a better understanding of the items EQAO includes on the assessment, they may not necessarily agree with the approach that is used. Thus, the EQAO Grade 9 assessment functions as a boundary object because it enables one group to better understand the perspective of the other group with respect to their shared goal of getting an accurate measure of students'

understanding of the mathematics curriculum. In this example it is teachers who are gaining a better understanding of the perspective of EQAO³⁵ and while they may not necessarily agree with the approach used by EQAO, their comments suggest that they do have a better understanding of EQAO's approach.

Creating marking guides for open-response items.

Another aspect of teachers' use of the assessment which increases teachers' understanding of EQAO's perspective is the development of marking keys for the open-response items. The school-level case studies demonstrate that as teachers create marking keys for the open-response items they have decided to mark, they become more familiar with the approach that EQAO uses for scoring these items. For example, Shannon explains her approach to scoring the open-response items:

For the long answer questions I would assess them based on how I've been assessing on all my tests . . . so, when I go to choosing the longer answer questions I would base it on what I would expect a Level 4 student to do . . . I have shown the students what [EQAO] says is a Level 4 answer and one of the big things that has come out is that the student can answer the long answer question and get the wrong answer and still get a Level 4 . . . but me personally, I put a little bit more emphasis on the right answer. (Shannon interview, p.2)

This description not only demonstrates that Shannon has considered the scoring approach which is used by EQAO, she has decided to score the open-response items using her own criteria even though her criteria (as described in chapter 6) differ from that which is used by EQAO. A number of other teachers also indicate that they mark the open-response items using the same approach they use to mark similar items on the unit tests they create for students rather than according to the EQAO approach. Thus, teachers' development of marking guides for the open-response items increases teachers' understanding of EQAO's

³⁵ Examples of EQAO gaining an understanding of teachers' perspectives are provided in the discussion of boundary encounters later in this chapter.

efforts to accurately measure students' understanding of the Grade 9 mathematics curriculum and this constitutes another example of how teachers' use of the assessment as part of students' grades turns the assessment into a boundary object. Once again, teachers have a better understanding of the approach used by EQAO even though they may not agree with that approach.

Emergence of the assessment as a boundary object over time.

Bowker and Star (1999) indicate that an object does not start out functioning as a boundary object but acquires this status over time as two communities continue to share one tool. The observations from this case study suggest that the EQAO Grade 9 assessment only began to function as a boundary object as more teachers began to use it as part of their students' grades. As described in chapter 6, in all three school districts classroom teachers were initially distrustful of the EQAO Grade 9 assessment. They questioned its value for classroom assessment purposes and most were not inclined to use it as part of their students' grades. These comments provide evidence that the EQAO Grade 9 assessment did not function as a boundary object during the first few years that it was administered; very few teachers were using the assessment so it could not be described as a shared tool. However, over time, in all three districts, participants in this study indicate that more and more teachers have begun to use the assessment as part of their students' grades. In two of the school districts, participants indicated that this shift took place as teachers became more familiar with the assessment items and could see the connection between those items and the curriculum documents.³⁶ As teachers became more familiar with the assessment items, their

³⁶ In the third school district use of the assessment as part of students' grades was mandated by the school district. While the observations I have gathered in this inquiry do not provide sufficient detail to comment, it would be interesting to investigate if the boundary functions of the assessment differ for this district because the use of the assessment as part of grades (i.e. the sharing of the tool) has been mandated.

increased familiarity made them more inclined to use the assessment as part their students' grades. In turn, including the assessment in their students' grades means that teachers are actively involved in selecting and marking the items and thereby become more and more familiar with EQAO's approach to assessment.

However, the observations from this case study also demonstrate that teachers' increased familiarity with the assessment does not lead to a consensus between teachers and EQAO with regard to all aspects of this assessment. This point is particularly evident with respect to the Grade 9 Applied assessment. That is, most of the participants in this study who use the EQAO Grade 9 assessment as part of their students' grades continue to express the view that the assessment is not well suited to students who are taking the Applied mathematics course. These observations suggest that the EQAO Grade 9 assessment has, over time, begun to function as a boundary object but that teachers are more in agreement with the appropriateness of the assessment as a measure for the Grade 9 Academic course than as a measure for the Grade 9 Applied mathematics course. Teachers and EQAO share the tool, teachers are better able to understand the perspective of EQAO as a result of using the tool as part of their students' grades, but teachers and EQAO continue to retain their own views with regard to various aspects of mathematics assessment.

Summary of Observations Related to Boundary Objects

The observations collected in the school-level case studies and the interviews with EQAO personnel suggest that if the assessment were not being used as part of students' grades it would not be a tool shared by two communities and would not function as a boundary object. I expect that in schools where the assessment is not used as part of students' grades teachers are less aware of EQAO's efforts to get an accurate measure of

students' understanding of the Grade 9 mathematics curriculum. Teachers in these schools would receive official score reports the following fall but their understanding of the assessment would likely be more limited than that of teachers who are using the assessment as part of their students' grades. For teachers who do use the assessment in grades, the assessment has become a shared tool used by EQAO for accountability purposes and by teachers to get additional information about how well students have grasped the mathematics curriculum. While each group has a distinct purpose in using the assessment, they share a common goal in their desire to get an accurate measurement of the extent to which students understand the mathematics curriculum. At the same time, it is also clear that teachers do not always agree with EQAO with regard to the items on the assessment, particularly with regard to the use of multiple-choice items and to the use of the assessment in the Grade 9 Applied mathematics course. This is the sort of enhanced understanding without consensus which is described as characteristic of groups using a boundary object (Star & Griesemer, 1989). The observations from this study also suggest that the emergence of the assessment as a boundary object has taken place over time. Whether or not this assessment will continue to function as a boundary object is unknown and could be the subject of further research.

Boundary Encounters as an Analytic Tool for the EQAO Grade 9 Assessment

In this section I explore how the concept of boundary encounters might contribute to a better understanding of the two uses of the EQAO Grade 9 assessment. As noted earlier, a boundary encounter is an activity which members of different communities engage in together that bridges the gaps in interpretation and meaning between the communities (Cobb et al., 2003). To gather evidence of boundary encounters for the EQAO Grade 9 assessment, I used the same analysis process I used for boundary objects. I searched the transcripts from

the school-level case studies as well as the EQAO interviews for examples of activities that might be considered boundary encounters and created a table to summarize these examples. In addition to the interview transcripts, I also considered responses to an open-ended item on the CIIM questionnaire (see chapter 5) where teachers describe their participation in EQAO activities. Through this analysis process, I found three kinds of activities which involve both EQAO and Grade 9 mathematics teachers and which seem to function as boundary encounters: teachers' participation in item development and official EQAO scoring, feedback from classroom teachers gathered by EQAO, and district-wide workshops for teachers led by EQAO personnel. In this section, I describe each of these activities and indicate the sense in which they function as boundary encounters.

Teachers' Participation in Item Development and Official EQAO Scoring

As noted in chapter 5, mathematics teachers are involved in the development of EQAO items and scoring keys as well as in the official marking of the open-response items which takes place during the summer. Several observations from this study suggest that these activities function as boundary encounters which help EQAO and classroom teachers better understand one another's perspective with regard to assessing students understanding of the Grade 9 mathematics curriculum.

From the perspective of EQAO, several interview participants indicate that the involvement of classroom teachers in item writing and their knowledge of students and classrooms helps EQAO ensure that the assessment is consistent with the Grade 9 mathematics courses. For example, Melanie states "our developers are teachers throughout the province who have developed the questions so they understand students and they're involved in the classroom" (Melanie interview, p.4). The benefits of teachers' involvement

in these activities in terms of enhancing the validity of the assessment are also described in some of the technical reports EQAO has produced (EQAO 2001, 2008).

Teachers who participated in this study also indicate that their experiences with item development and summer marking help them to better understand the mathematics curriculum, enhance their assessment practices and help them to better understand the approach EQAO uses for marking the Grade 9 assessment items. Some evidence of this is found in teachers' responses to an open-ended item on the CIIM questionnaire (described in chapter 5). This item asked teachers to describe a professional development experience that had positively influenced their approach to mathematics teaching. In response to this item, several teachers identified their participation in EQAO summer marking, item writing and the development of EQAO scoring guides as beneficial. Examples of the benefits they associate with this experience are:

Marking the Grade 9 EQAO assessment during the summer gave me a better understanding of the four categories of knowledge and understanding, application, problem solving and communication. This allowed me to better design my assignments and tests in order to assess students in the categories.
(Participant # 03400)

Marking EQAO in the summer. It made me really think about how different students can give different answers and it can still be in the same level. It also made me try to analyze whether or not the student can demonstrate their learning by explaining their thinking in their answers.
(Participant # 03970)

Having been a marker/group leader of EQAO, I am of the opinion that teachers are more cognizant of the abilities of their students. (Participant #24206)

In a similar vein, one teacher from the school-level case studies, Shannon, who had struggled with the criteria used by EQAO to assign a Level 4 to student's responses commented that she was not comfortable with the approach taken by EQAO and indicated that she would like to have a better understanding of EQAO's reasons for developing this assessment criteria. I

asked her if she had ever been involved in EQAO summer marking and she responded that she had not and added “but actually maybe if I was it would probably make more sense to me and then I would understand their rationale” (Shannon interview, p.6).

In my interviews with EQAO personnel I also learned that during the summer marking session a number of other activities take place, beyond training the teachers to mark and the actual marking of the open-response items, which function as boundary encounters. For example, Carol and Susan explain that during some summer marking sessions EQAO has held “lunch-and-learn” activities at the marking site where mathematics teachers who have been hired to mark the open-response items can listen to explanations of the psychometric procedures which are used to develop the EQAO Grade 9 assessment items, to equate the assessments from year-to-year and to ensure that the assessments are scored reliably. Carol and Susan explained that a series of 5 or 6 half-hour sessions were held during which psychometric tools such as item response theory (IRT), an approach used with the EQAO Grade 9 assessment, were explained. They also indicated that these voluntary information sessions were well attended by the teachers, that the teachers seemed quite interested in the ideas presented and that they asked in-depth questions during the sessions.

Susan describes these lunch-and-learn sessions as follows:

It was great to see their interest and the fact that they understood probability at some level or another allowed them to understand what IRT was doing . . . and they understood all this stuff so they went away feeling ‘yeah, this sounds okay’ . . . everyone seemed to understand it to a certain extent. (Susan interview, p. 9)

Susan’s description suggests that these lunch-and-learn sessions provided the teachers with a much better understanding of the EQAO Grade 9 assessment process and, for this reason, I consider these sessions to be another example of a boundary encounter.

The benefits of teachers' involvement in item development and marking activities have also been observed in other studies of the EQAO Grade 9 assessment. Suurtamm, Lawson and Koch (2008) identify some of the ways that the involvement of mathematics teachers in item development and marking is beneficial for the EQAO Grade 9 assessment. They maintain that the involvement of teachers in the item development process helps to ensure that the items on the assessment match the curriculum and acknowledges the important contribution made to the Grade 9 assessment by teachers who are implementing the curriculum and are familiar with the nature of activity in Grade 9 mathematics classrooms in Ontario. The benefits of teachers' involvement with developing and marking the EQAO for both classroom teachers and EQAO are also discussed by Kitto (2006), Lock (2001) and Nagy (2000).

Feedback from Classroom Teachers Gathered by EQAO

The second activity involving EQAO personnel and classroom teachers that functions as a boundary encounter is the feedback that EQAO seeks from classroom teachers with regard to the EQAO Grade 9 assessment. One of the principals I interviewed explained that "there was a lot of feedback from teachers, actual practitioners in the field to EQAO, how to make it more teacher-friendly, more student relevant, there was lots of feedback there" (Allen interview, p.2). Allen felt that teachers' feedback had been instrumental in areas such as EQAO's decision to reduce the number of days that the assessment is administered. Another participant in the study, Owen, indicated that EQAO's decision to reduce the administration time from 70-minute sessions to 50-minute sessions made the administration of the assessment much more feasible for teachers (Owen interview, p.2). Seeking feedback from teachers may be considered a boundary encounter in the sense that both groups

participate in the activity and it enables EQAO to better understand the perspective of teachers with regard to the EQAO Grade 9 assessment.

District-wide Workshops for Teachers

The third activity that seems to function as a boundary encounter which was described in one of the interviews I conducted is that EQAO personnel are sometimes invited by school districts to conduct workshops with mathematics teachers. During these workshops, EQAO provides teachers with sample student responses to some of the released items from the EQAO Grade 9 assessment and teachers have an opportunity to score these student responses and to discuss the scoring criteria with other teachers as well as with the EQAO personnel who conduct the workshops (Dana interview, p.3). Dana expressed the view that this activity helps teachers to better understand EQAO's approach to scoring at the same time as it provides EQAO personnel with insights into teachers' perspectives with regard to the Grade 9 assessment. She indicated that through working with teachers in these district-wide sessions she has become familiar with a range of approaches mathematics teachers use to score the open-response items and she described these sessions as "a really enriching conversation" (Dana interview, p.3).

Summary on Boundary Encounters

My observations in this case study suggest that there are a number of activities involving EQAO personnel and classroom teachers that seem to function as boundary encounters. Teachers' participation in developing and reviewing EQAO items and in summer marking teams, EQAO's efforts to seek feedback about the assessment from classroom teachers, and district-wide workshops for teachers are three examples of boundary encounters I identified in this study. The comments made by teachers and

EQAO personnel who participated in this study suggest that these encounters facilitate communication between teachers and EQAO personnel and enable both groups to gain a better understanding of one another's perspective regarding the Grade 9 assessment.

In the time period since I conducted this case study, EQAO has begun to make changes to their assessment development approach and is moving toward contracting this process out to other assessment development agencies. This approach to assessment development would reduce teachers' involvement in the item development process. While teachers would still be involved in the official marking of the open-response items on the Grade 9 assessment, the boundary encounter functions which are associated with teachers' participation in the item development process would likely be diminished. Thus, while there is evidence to suggest that there are boundary encounters associated with the Grade 9 assessment, changes made by EQAO in the item development process might reduce the opportunity for teachers and EQAO to continue to learn more about one another's perspectives. This is a clear example of how the boundary functions associated with an object may change over time.

The Value of Boundary Concepts for Validating this Instance of Multiple-use

What then, is the value of using boundary concepts as part of the process of validation for the EQAO Grade 9 assessment? The case study revealed interactions between the use of the assessment for accountability and the use of the assessment as part of students' grades. As indicated earlier in this chapter, using Kane's model enables the validation of one use of this assessment and then the other but does not seem to provide a means of accounting for interactions between the uses. In light of the interactions which were found, one approach which could be taken is to use Kane's concepts of defeasibility and the assumption of *ceteris*

paribus to overturn the validity argument for one or both uses of the EQAO Grade 9 assessment. Thus, we might conclude that given the interactions between the two multiple-uses discovered in the case study, the use of the assessment for accountability and/or as part of students' grades may not be warranted. However, simply overturning the validity argument is a limiting approach in that it does little to help us better understand the significance of the interactions that are taking place. Some interactions between the two uses might seriously undermine the validity argument while others might not.

In contrast with this approach, using boundary objects and encounters as analytic tools contributes to the process of validating the multiple-use of the EQAO Grade 9 assessment in several ways: by unpacking the interactions which are taking place in multiple-use, by including multiple perspectives in the process of validation, by acknowledging the hierarchical relationships which are present in the multiple-uses and by facilitating the contextualization of the practice of multiple-use. In this section, each of these ways that boundary concepts contribute to the process of validation for the EQAO Grade 9 assessment are discussed using examples from the observations in chapters 6 and 7. Some comments on how the benefits of using boundary concepts in the process of validation observed in this case study may extend to other instances of multiple-use are also offered.

Unpacking the Interactions among Multiple-uses

Using the analytic tools of boundary objects and encounters helps to unpack or better understand the interactions which were found. Conducting an analysis using boundary concepts provides a sense of which interactions are reasonably well negotiated and which interactions may be untenable. To illustrate how interactions are better understood using boundary concepts, two examples from this case study are described. The first example is an

aspect of an interaction that seems to be reasonably well negotiated while the second example is an aspect of an interaction which continues to be problematic for the multiple-use of the EQAO Grade 9 assessment.

One of the interactions identified in the case study is that some of the item characteristics which are associated with the use of the EQAO Grade 9 assessment for accountability purposes may undermine the use for classroom assessment purposes (see chapter 6 for a detailed description of this interaction). For example, while the EQAO Grade 9 assessment has moved toward greater reliance on multiple-choice items which can be quickly and reliably scored, many teachers who participated in the school-level case studies question the suitability of these items for measuring Grade 9 students' mathematical understanding. However, the information gathered in this case study also suggests that teachers are able to negotiate this interaction in the sense that they can choose not to include the multiple-choice items in the portion of the assessment that they include in students' grades. Alternatively, some teachers include the multiple-choice items in what they are marking but have chosen to weight the assessment as a lower proportion of their students' grades because they have concerns about the multiple-choice items. In these ways, teachers are able to reduce the impact this interaction has on their use of the assessment as part of students' grades.

A second example of an interaction which is better understood as a result of the boundary concept analyses is the impact of the different scoring approaches used by teachers and EQAO for the open-response items. As explained in chapter 6, evidence from this case study suggests that teachers who score the open-response items may use a different assessment criteria than that which is used by EQAO with the result that students may be

unsure as to how to structure their responses. The EQAO assessment functions as a boundary concept in that teachers who use the assessment as part of their students' grades or who participate in the official marking process have a better understanding of the scoring approach which is used by EQAO. However, some of these teachers disagree with EQAO's marking criteria and continue to use their own criteria for these items. Thus, this interaction is better understood through the boundary analyses (i.e. we better understand the perspectives of both teachers and EQAO with regard to scoring) but the interaction remains problematic.

The information about the interactions which are taking place which comes from these boundary analyses can be incorporated into the evaluative judgement made when validating these two multiple-uses. Those interactions which remain problematic may lead to a judgment that this combination of uses (as part of students' grades and for accountability) cannot be warranted. Thus, conducting the in-depth case study helps to identify the pattern of uses and locate interactions between those uses and analysing the case study observations using boundary concepts enables a better understanding of the impact those interactions may have. We get a better sense of which interactions are being effectively negotiated and which remain problematic. In this manner, the concepts of boundary objects and encounters contribute to the process of validation for this instance of multiple-use.

A similar approach may be worthwhile for other instances of multiple-use where evidence of interactions among the uses has been found. Looking for evidence of boundary objects and encounters may reveal valuable insights about the interactions and these insights can be used along with other evidence in making the evaluative judgments that are a part of the process of validation.

Including Multiple Perspectives in the Process of Validation

Using the concepts of boundary objects and encounters to analyse the multiple-use of the EQAO Grade 9 assessment is also beneficial in that it provides a means of ensuring that multiple perspectives are included in the process of validation. Looking for evidence that the assessment functions as a boundary object and that activities which involve both teachers and EQAO personnel function as boundary encounters, requires a consideration of the perspectives of teachers and EQAO personnel as well as a sense of the extent to which the two groups understand one another. Since a defining characteristic of all boundary concepts is that they facilitate communication and understanding between the groups who are using them, applying these concepts ensures that multiple perspectives are considered. For example, in the analysis of the case study observations it became clear that many teachers do not share EQAO's perspective with regard to the suitability of this assessment for use in the Grade 9 Applied mathematics course. The concerns expressed by teachers with regard to the Applied assessment are important evidence which should be included in the process of validation for this assessment.

Similarly, for other large-scale assessments where multiple-use takes place, looking for evidence of boundary objects and encounters would require a careful consideration of the perspectives of the various groups using the assessment. As noted in chapter 1, from a sociocultural perspective, the recognition and inclusion of multiple perspectives is an important aspect of understanding assessment practice.

Acknowledging Hierarchical Relationships

Using the concepts of boundary objects and encounters to analyse the multiple-use of the EQAO Grade 9 assessment also provides some insights about the relative status of teachers and EQAO with regard to mathematics assessment. Most of the evidence for boundary concepts in my observations demonstrate ways that teachers gain a better understanding of EQAO's perspective with regard to mathematics assessment as a result of using the assessment as part of their students' grades. There are markedly fewer examples of instances where EQAO gains a better understanding of teachers' perspectives of mathematics assessment from this shared tool. As noted earlier in this chapter, Bowker and Star (1999), suggest that boundary objects are particularly useful as analytic tools when the groups who share the object have relatively equal status. For the EQAO Grade 9 assessment, the status of EQAO is arguably greater than that of classroom teachers especially given the value which is currently placed on accountability assessments (as described in chapter 2). The unequal status between the two groups is a factor which merits further consideration in the application of boundary concepts to the multiple-use of the EQAO Grade 9 assessment. Nonetheless, an additional benefit which comes from applying boundary concepts to this instance of multiple-use is that the relative status of the two groups is made more evident.

From a sociocultural perspective, reducing the hierarchical relationships among those involved in assessment practice is seen as valuable. Thus, the use of boundary concepts in this instance of multiple-use helps to reveal the hierarchical relationships which exist leading to a better understanding of the assessment practices that are

taking place and thereby contributing to the process of validation. Highlighting the hierarchical relationships among assessment uses may also be beneficial in the process of validating other large-scale assessments where multiple-uses involving a variety of stakeholders are taking place.

Context and Contextualization in Validating Multiple-uses

Using the concepts of boundary objects and encounters to analyse the multiple-use of the EQAO Grade 9 assessment is a means of contextualizing the assessment practices which are taking place. Boundary objects and encounters are particularly useful tools for understanding how a social practice, in this case assessment, takes place across different settings or contexts. In the case of the EQAO Grade 9 assessment, the assessment practice being examined is both a large-scale assessment used for accountability purposes and a teacher-constructed classroom assessment used as part of students' grades. Rather than look at the assessment practice as taking place in two distinct and static contexts, analysing the practices which are taking place using boundary concepts helps us better understand how the contexts co-emerge with the assessment practices. Teachers who decide to use of the EQAO assessment as part of their students' grades are changing their classroom assessment practices but, at the same time, these teachers' classroom assessment practices have an impact on the use of the assessment for accountability purposes. Boundary concepts are key aspects of contextualization because they focus on the relationships between various individuals and groups, the tools shared by those groups and the mediated activities that the groups are engaged in. The use of boundary concepts is one way of making the sometimes intangible process of contextualizing the use of an assessment more explicit. I suggest

that gaining a better understanding of these dynamic assessment activities is an important aspect of the process of validation for large-scale assessments where multiple-use takes place.

CHAPTER 8

IMPLICATIONS OF MULTIPLE-USE FOR THE PROCESS OF VALIDATION

As observed in chapter 2, despite evidence of the frequent occurrence of the multiple-use of large-scale assessments, the implications of multiple-use for the process of validation have not been broadly explored in the assessment literature. The limited research published on this topic suggests that identifying the uses associated with an assessment is an important step in the process of validation (Haertel, 1999; Stobart, 2009) and demonstrates that when assessments are used for multiple purposes a large number of inferences and assumptions may need to be warranted (Haertel, 1999). However, as observed by Stobart (2009), a framework for prioritizing and organizing the process of validation given the practice of multiple-use is needed. While I concur with the observations of Haertel and Stobart, as I observed in chapter 2, both authors seem to make a tacit assumption that multiple-uses are independent of one another. Because multiple-uses all stem from a single administration of an assessment, I contend that multiple-use increases the stakes associated with an assessment and that interactions among multiple-uses are likely to occur. Both these issues must be considered in the process of validation.

In this final chapter, I step back to consider what this inquiry has shown about the implications of the multiple-use of large-scale assessments for the process of validation. Drawing on my theoretical discussion of the problem of multiple-use and my in-depth case study of the multiple-use of the EQAO Grade 9 assessment, I suggest a number of aspects of the process of validation that must be reconsidered in order to better address the multiple-use of large-scale assessments. I also identify the contributions this dissertation makes to the

field of educational assessment, acknowledge some of the limitations of this inquiry and suggest some areas for further research.

Characteristics of the Process of Validation that Address Multiple-use

In this case study of the multiple-use of the EQAO Grade 9 assessment, evidence of interactions between two of the multiple-uses was found and the limitations of using Kane's argument-based model for validation given these interactions was shown. I explored the contributions to the process of validation that come from the use of boundary objects and boundary encounters as analytic tools for the case study observations. The observations from this case study demonstrate that there is a need for models of validation that can address the complex ways assessments are actually being used. The study also provides some insights into characteristics of the process of validation that could better address the multiple-use of large-scale assessments. In this section, five such characteristics are described.

Focusing on the Pattern of Uses Rather than the Intended Use

Most measurement-based views of validity are centered on the idea that there is an intended use for an assessment (e.g. AERA et al., 1999). The notion of an intended use predisposes researchers to focus validation efforts on that use. Focusing on the intended use of the assessment may be appropriate during the development phase of validation which is typically carried out by the assessment developers. However, given the complex ways that many large-scale assessments are used in practice (i.e. intended and alternative proposed uses, multiple scoring and multiple-use), I suggest that continuing to focus solely on the intended use during the appraisal stage of validation is problematic. To begin with, determining the intended use for an assessment is not always straightforward. As discussed in chapter 2, some assessment developers do not explicitly identify an intended use and

many others indicate a range of uses for their assessment (Cronbach, 1971; Koretz & Hamilton, 2006). Further, even when the intended use is clear, there is a tendency for the uses of large-scale assessments to shift and expand over time (Klinger et al., 2008; Nagy, 2000; Wolfe et al., 2004). In addition, a review of the literature suggests that the multiple-use of large-scale assessments is neither a recent phenomenon nor a rare occurrence. Examples of multiple-use include province and state-wide assessments used in various US states (Phoham, 2002; Wilson, 2007), England (Stobart, 2009), Canada (Klinger et al., 2008) and a number of other countries (de Lange, 2007) as well as international comparative assessments such as PISA and TIMMS (Phelps, 2000). These examples of the multiple-use of large-scale assessments across a wide variety of jurisdictions disrupt the notion of a single intended use around which validation efforts should focus. For these reasons, I contend that the process of validation, particularly during the appraisal stage, must begin with identifying the pattern of uses associated with an assessment. Identifying the pattern of uses must include a consideration of the ostensible intended use as well as any additional uses or intended consequences proposed by the assessment developer, and those uses of the assessment which are carried out by other individuals or groups. Even those uses which are not sanctioned by the assessment developer and are viewed as misuses should be included in the pattern of uses for an assessment because they may increase the stakes associated with an assessment and interact with other more legitimate uses. I see establishing the pattern of uses for an assessment as a critical step that sets the parameters for the process of validation to be undertaken.

Looking for Evidence of Interactions

Once the pattern of uses has been identified, rather than proceeding to validate each use separately, I maintain that there is a need to gather evidence to see if there are interactions among the multiple-uses. One way of finding evidence of interactions among the uses is by carefully examining the procedures and activities which take place within each multiple-use. In this study, I followed this approach for two uses of the EQAO Grade 9 Assessment of Mathematics and found evidence that interactions between the two uses of the assessment are taking place. For any large-scale assessment, once the interactions are identified, any additional evidence that might need to be gathered to understand the impact of the interactions on the interpretations being made from the assessment can be specified. Such evidence makes an important contribution to the process of validation for an assessment. While it may not be feasible to identify or pursue every interaction, particularly with assessments that have a large number of multiple-uses, those uses which are most likely to have an impact on other uses should be the most rigorously reviewed. In ignoring the interactions between uses, as is presently the case, we run the risk of overlooking evidence that might change the evaluative judgments that are made in the process of validation.

Developing Flexible Approaches to Validation

The practice of multiple-use highlights the need for flexible models of validation which can be employed in a variety of settings. With the EQAO Grade 9 assessment, a single assessment is being used both as a large-scale assessment and as a classroom assessment (i.e. for accountability and as part of students' grades). As noted, the use of large-scale assessments as part of students' grades also takes place in other educational jurisdictions in Canada (CMEC, 2009; Klinger et al., 2008). This intersection of classroom

and large-scale assessment practices requires an approach to the process of validation that can address each type of assessment as well as the intersection of the two. Similarly, instances of multiple-use may take place in a variety of settings. For example, the use of assessments for ranking purposes is often conducted by independent research organizations and then distributed by the media. Thus, approaches to validation must include the capacity to gather and evaluate evidence from the wide range of settings where large-scale assessments may be used.

Developing a Collaborative Approach

The practice of multiple-use reveals the need for increased collaboration in the process of validation. In chapter 3, I discussed the idea that validation is a shared responsibility (Cronbach, 1971; Kane, 2006, 2008; Messick, 1989; Shepard, 1997). The commonly held view is that assessment developers should validate the intended use while those who propose other uses should have the responsibility for validating those uses. This notion of shared responsibility amounts to a splitting or division of responsibilities and this is not the same as adopting a collaborative approach to the process of validation. I think that dividing the responsibilities for the process of validation is not an adequate approach for assessments where multiple-uses are known to occur. Rather, I suggest greater collaboration between assessment developers and the various assessment users is needed to enable the collection of validation evidence regarding the impact of interactions among the uses on validating the inferences which are being made. As was noted for some of the interactions observed in the multiple-use of the EQAO Grade 9 assessment, neither the assessment developers nor the classroom teachers acting in isolation had access to the information needed to better understand the interactions that are taking place. Teachers and EQAO

personnel need to work collaboratively to collect and share information in order to better understand the implications of the interactions for validating these two uses of the assessment.

The need for greater collaboration in the process of validation is also evident in the case of international assessments with multiple-uses. For instance, both the PISA and TIMSS assessments are used to compare the performance of schools across countries but the results of these assessments are also used within some countries to monitor the performance of teachers and to assess the effectiveness of educational programs (Phelps, 2000). To address the multiple-uses of international assessments, the process of validation would require the collaborative efforts of a broad range of individuals and groups including assessment developers and administrators as well as individuals at various levels in the education systems of countries where these assessments are being used. Dividing responsibility for the process of validation among these users would result in a much more fragmented judgment of the inferences which are made from the assessment. I suggest that the multiple-use of these large-scale assessments necessitates a more holistic and collaborative approach.

At the same time, I recognize that developing a collaborative approach to the process of validation across diverse groups of individuals who have different perspectives and agendas represents a considerable challenge. One theoretical framework which could contribute to the development of a more collaborative approach to validation is philosophical hermeneutics. As described in chapter 3, Moss suggests that using a hermeneutic approach to validation could provide a means of integrating the variety of types of evidence which are gathered when developing an interpretation (Moss et al., 2006). Moss also indicates that

adopting a hermeneutic approach facilitates the appreciation of multiple interpretations and helps make differing interpretations mutually comprehensible to various stakeholders. The use of hermeneutic approaches can result in a more holistic and integrative approach to validation because the focus is not on arguing against the other person but on questioning to bring out the strengths in the other's argument. For example, the hermeneutic circle is a recursive process that considers individual interpretations in relation to group interpretations. The process continues as new information from a variety of stakeholders is brought forward. Using a hermeneutic circle approach, interpretations that are appropriate to the available evidence and that are situated within a specific setting can be made and the process through which the interpretations were created can be described along with any remaining insufficiencies in evidence. In such an approach, the various individuals and groups involved in the assessment process can make their own evaluative judgment of the interpretations that are made on the basis of the assessment. Using an approach such as the hermeneutic circle to draw together information from a variety of stakeholders could be a means for the kinds of collaboration that is needed in the process of validation.

Thus, this inquiry suggests that much more extensive collaboration among users must be an essential aspect of the process of validation where multiple-use is known to occur. The allocation of responsibility for specific pieces of the validation process to specific users cannot adequately address the interactions that take place among users and approaches situated within hermeneutics may facilitate and encourage this sort of collaboration.

Using Case Studies in the Process of Validation

This study demonstrates that in-depth case studies are a valuable framework for the process of validation. Using a case study approach facilitates the identification of multiple-

uses and enables gathering detailed information about how each use take place. From this detailed information and by considering the perspectives of various individuals, interactions among the uses may be found and the additional evidence needed to investigate the significance of these interactions can be identified. For example, information gathered in this case study shows an interaction between the two uses that is related to the weighting of the EQAO Grade 9 assessment and indicates that studies of the impact of weighting on students' performance are needed for the validation of this assessment. In addition, an interaction between the two uses related to the scoring of the open-response items was identified and suggests that studies of the impact of scoring methods on students' performance are needed as part of the validation of this assessment. Thus, case studies facilitate the identification of interactions and also reveal the kinds of additional evidence that must be gathered in the process of validation. Further, in-depth case studies provide detailed information about a specific assessment practice. This information can be analysed using a variety of tools, such as boundary concepts, which then contribute to the process of validation. Accordingly, while case study is not a method that has been widely used in measurement-based approaches to validation, I think it has tremendous potential as a framework for validating assessments which are used in complex ways.

Contributions to Educational Assessment

This inquiry makes a substantial contribution to the field of educational assessment. One important area of contribution is the identification and description of the practice of multiple-use itself. While a few articles have mentioned that multiple-use takes place, the characteristics of this practice have not been previously explored. Factors such as the time frame over which multiple-uses occur, the range of individuals and groups who may be

involved in multiple-use, and the impact of test users and test takers knowing about multiple-use at the time the assessment is administered had not been previously considered and are explored in my discussion and in the visual representations of multiple-use included in chapter 2. In addition, I have described some early instances of multiple-use as well as a number of more recent examples. I have explored the ways that multiple-use relates to other practices in the use of large-scale assessments such as multiple scoring and intended and alternate proposed uses. I have also suggested a number of factors which may facilitate multiple-use and contribute to an increase in its occurrence, especially for accountability assessments where assessment results are widely available.

Another key contribution of this dissertation is that I identify and demonstrate two specific problems that the practice of multiple-use creates for the process of validation. This is the first study which points out that increased stakes are associated with multiple-use practices. That is, the stakes that are associated with an assessment are likely to increase as the number of uses of an assessment increase. As noted in chapter 2, the process of validation for assessments with high stakes often differs from that associated with lower stakes assessments (Koretz & Hamilton, 2006). Accordingly, even where there are no interactions between uses, the multiple-use of an assessment raises issues which must be addressed in the process of validation. In addition, while other articles identify the plethora of assumptions to be considered when validating an assessment which is used for multiple purposes (Haertel, 1999), this study is the first to suggest the potential for interactions among uses to occur and the importance of investigating those interactions in the process of validation. I maintain that the separate validation of each use as though the other use were

not taking place is not an adequate approach. Multiple-use is a practice which must be acknowledged and addressed throughout the process of validation.

A third area of contribution is that this dissertation shows how the in-depth study of one assessment practice can contribute to the current dialogue on the process of validation. Studying the practice of multiple-use in a particular situation has suggested many ways that the process of validation may need to be reconsidered. For example, the importance of collaboration among the various stakeholders involved in the multiple-uses of a large-scale assessment was revealed in this study when it became apparent that neither EQAO nor classroom teachers has the necessary information to explain the differences between official and teacher-derived scores on this assessment. In-depth, situated studies of other practices in the use of large-scale assessments might also contribute to re-conceptualizing the process of validation.

The fourth key area of contribution this dissertation makes is in demonstrating how drawing on areas such as sociocultural studies can help to re-conceptualise the process of validation. While Moss suggested that such approaches might be useful, this is the first study I am aware of where the concepts of boundary objects and encounters have been applied to the process of validation. The contributions which these approaches can make to the process of validation particularly in understanding those interactions which are reasonably well negotiated as well as in demonstrating the activities which various stakeholders may engage in to better understand one another's perspectives are significant. The information gathered from an analysis using boundary concepts can be incorporated into the validation process and into the evaluative judgments that are made in validating multiple-uses. As noted, those

interactions that remain problematic may, in some instances, lead to a judgment that a particular combination of uses cannot be warranted.

Limitations of this Inquiry

In this section I suggest some of the limitations of this inquiry with regard to understanding the implications of multiple-use for the process of validation. The limitations I identify are of two types; some are characteristics of the case study which limit our understanding of the multiple-use of the EQAO Grade 9 assessment while others are aspects of the research that may limit the way this study can inform other instances of multiple-use.

Limits on Understanding the Multiple-use of the EQAO Assessment

Three aspects in the design of this study that may limit our understanding of the multiple-use of the EQAO Grade 9 assessment are the limited number of perspectives that could be included, the reliance on self-reported data, and the fact that the boundary analysis was conducted by a single researcher. Each factor is briefly described.

In this study, I focused principally on the perspectives of teachers who have decided to use the EQAO Grade 9 assessment as part of their students' grades and of EQAO personnel who have been involved in the development and administration of this assessment. Considering the perspectives of these two groups in some detail enabled the identification of interactions between the uses which was one of the key research questions in this inquiry. However, valuable insights about this practice might also come from conducting case studies in schools where teachers have chosen not to include the EQAO Grade 9 assessment as part of their student's grades. For example, asking these teachers to describe their reasons for this decision might reveal additional tensions related to this use of the assessment that teachers who have decided to use the assessment are not aware of. In addition, some other

perspectives that are not included in this inquiry are those of Grade 9 mathematics students, parents, school board administrators and the Ontario Ministry of Education. Understanding these perspectives would enrich our understanding of teachers' use of the EQAO Grade 9 assessment as part of students' grades. For example, understanding why many school board administrators and/or the Ministry of Education have not established policies with regard to the amount that teacher-derived EQAO scores can contribute to students' grades would be very informative. In any inquiry a limited number of perspectives can be included but, as noted earlier, in interpretivist research it is important for the researcher to acknowledge those perspectives which have been considered as well as those which have been left out.

Another characteristic of this inquiry that limits our understanding of the multiple-use of the EQAO Grade 9 assessment is that the research design relies mostly on self-reported data. For example, responses to the CIIM questionnaire provide information about what teachers report in terms of their use of the EQAO Grade 9 assessment as part of students' grades but what teachers report doing may differ from what they actually do. Similarly, in the school-level case studies I draw on teachers' descriptions of the procedures they follow but I do not have any observations or artifacts of what they actually do as they use the EQAO Grade 9 assessment items as part of their students' grades. Because teachers are not permitted to keep any documents related to the use of the EQAO assessment, they do not have any artifacts, such as the marking keys they develop, that would provide evidence of the approach they take as they mark students' responses. Conducting observations in schools at those times when teachers are selecting the items to mark, creating the marking guides and completing the marking process would be very informative. However, for reasons of test security, EQAO might not consent to having a researcher present during this

process. Certainly their current policies prohibit teachers from sharing the content of the assessment with any individuals other than their students.

A third limitation of this inquiry is that the boundary analysis was conducted by a single researcher. If the project had been conducted by a research team, rather than as a doctoral dissertation, a second researcher could have conducted an independent analysis of the transcripts to look for evidence that the assessment functions as a boundary object and that boundary encounters are taking place. Such an approach would provide additional support for the findings reported herein.

Limits on the Ways the Study Informs Other Instances of Multiple-use

Two aspects of this study may limit the ways it can inform other instances of multiple-use: this study focused on the multiple-use of items and scores rather than the multiple-use of assessment scores only and this study focused on only two of the multiple-uses that actually take place. Each aspects is described in a little more detail.

In the instance of multiple-use investigated in this dissertation the additional use by teachers is of the assessment items rather than the test scores. That is, in the use of the EQAO scores for accountability purposes interpretations are made based on the test scores. In using the assessment as part of students' grades, teachers do not use the official test scores, they use the test items. In a sense teachers construct their own assessment by selecting the items they wish to score. Thus, the interactions related to item selection and scoring observed in this instance of multiple-use but might not be a factor in other multiple-use cases. As noted, the multiple-use of the EQAO Grade 9 assessment is also an instance of multiple-scoring. In all of the other examples of multiple-use that I uncovered in the assessment literature (see chapter 2) it is the simultaneous use of the *test scores* for two or

more purposes that takes place. For example, in the multiple-use of an international comparative assessment (for comparisons of student achievement across countries *and* as a measure of program effectiveness within some of those countries) both users would be using the same test score results. Thus, the interactions would likely differ from those found with the multiple-use of the EQAO assessment. Given these observations, much can be learned from research focusing on examples of multiple-use where both uses relate to the same assessment score.

This study showed several multiple-uses of the EQAO Grade 9 assessment (for example, five uses are shown in figure 7, p. 175) but only two of those uses are closely examined. Considering the other uses and the interactions which might be taking place would be beneficial for the process of validation for the EQAO Grade 9 assessment as well as for understanding other instances of multiple-use where several uses are known to take place. From such studies, ways of modeling the interactions among three or more uses might be developed.

Suggestions for Further Research

While the central question of this inquiry emerges from a measurement-based view of the process of validation, I wanted to move away from a technical and psychometric view of multiple-use to a more sociocultural view of the process of validation for large-scale assessments where multiple-use takes place. This study of the multiple-use of the EQAO Grade 9 Assessment of Mathematics has suggested some ways to do this. I have shown that the practice of teachers' using large-scale assessments as part of students' grades is quite widespread not only in Ontario with the EQAO Grade 9 assessment but with other province-wide assessments in Canada. Multiple-use is fairly common and yet a way of dealing with

validation that addresses the interactions among uses is not obvious. Continued research on this issue is needed and in this section I suggest three areas of focus for that research.

More studies into multiple-use practices are needed. In particular, studies looking at different patterns of multiple-use and uncovering the interactions which might be occurring between those uses would be beneficial. One challenge which needs to be addressed is a means of looking at the interactions between three or more uses. For instance, in this study I considered only the use of the assessment for accountability and the use of the assessment by teachers as part of their students' grades but I think that looking at how the use of the EQAO Grade 9 assessment by the Fraser Institute to rank schools impacts these two uses would be very informative. While the full range of interactions may not be able to be considered, I suggest that some effort to understand those interactions which are most likely to affect the validity of interpretations made for other uses must be made. I contend that the powerful impact of publicly released ranking tables, which takes place for many large-scale assessments in a variety of educational jurisdictions should be better understood – even if the use for ranking is identified as a misuse by the assessment developers.

Another area of research is to continue to identify useful analytic tools. I used boundary objects and encounters as a way of looking at the interactions and found some evidence that the assessment can function as a boundary object and that boundary encounters do occur. The use of this analytic approach contributes to the validation process because it is a way of understanding which interactions are being negotiated and which interactions remain. It also provides insights into the relative status of the individuals and groups who are using an assessment and ensures that multiple perspectives are included in the process of validation. However, boundary objects and encounters are just one tool that might be

considered in a validation effort. Validity evidence concerning the interactions between multiple-uses can be gathered using other analytic tools. Where multiple-use is known to occur, and where we want to consider three or more uses, a variety of tools may be needed to collect and analyse validity evidence. Accordingly, research which helps to find more analytic tools for the process of validation would be very valuable.

Finally, there is a need for further theoretical exploration of multiple-use. This study contributes to discussions of the re-conceptualization of validity theory currently taking place in the measurement community. The multiple-use of large-scale assessments is a frequently occurring practice. This study highlights the implications of this practice for the process of validation and demonstrates the limitations of an argument-based approach to validation where multiple-use takes place. The study points to a need to expand existing approaches to validation to better address this practice and suggests some ways to begin this process. While this inquiry is an initial foray into these issues, there is a need for further dialogue regarding the practice of multiple-use and for additional theoretical considerations of how the process of validation might be undertaken given this practice.

Closing Remarks

Before I began this inquiry I looked at teachers' use of the EQAO Grade 9 assessment as part of their students' grades solely from a measurement-based perspective and I considered the practice to be misguided and inappropriate. Since that time, I have gained a much greater awareness of the complex ways that large-scale assessments are frequently used as well as an appreciation of the benefits of looking at these complex practices through a sociocultural lens. In this study I focused on the ways that boundary concepts might overcome some of the limitations of an argument-based approach to the

process of validation for assessments where interacting multiple-uses are known to occur. I have found a number of ways that the use of boundary concepts in the analysis of large-scale assessment practices can contribute to the process of validation in this study and expect that using these concepts may contribute in other ways beyond what has been described in this dissertation. Accordingly, I look forward to continuing to explore the use of boundary concepts in future research.

Conducting this inquiry has also provided me with a much greater understanding of how adopting a hermeneutic stance might contribute to the process of validation. Earlier in the inquiry process I could not envision how hermeneutics could be applied to the multiple-use of large-scale assessments. However, as I have become a little more familiar with hermeneutics and much more familiar with the activities that take place as teachers use the EQAO Grade 9 assessment as part of their students' grades, I have begun to see specific ways that approaches such as the hermeneutic circle could contribute to the process of validation. Thus, through this inquiry I have gained a better understanding of the implications of multiple-use for the process of validation as well as a clear indication of some directions for my future research.

Epilogue

As I was in the process of completing this manuscript in January 2010, five years after the curriculum night at my son's school where I first learned of teachers' use of the EQAO Grade 9 assessment as part of students' grades, my daughter was preparing to complete this assessment as part of her Grade 9 Academic mathematics course. Her teacher provided copies of released items from the EQAO website to help students prepare for the assessment and he encouraged them to look at the anchor responses for these items posted on the EQAO website. As my daughter began to prepare for the assessment, she learned that her teacher, and all the teachers in the mathematics department at her school, had decided to use the EQAO assessment in lieu of the in-school final examination they had previously used as part of students' summative grade. She learned that the teachers would be marking all the items on the EQAO assessment and that those items would count as 20% of her final grade. The other 10% of her summative grade would be based on a common performance task.

My daughter expressed a number of concerns with regard to completing the EQAO assessment. She was concerned that her final course grade would be negatively affected by her performance on this large-scale assessment. In addition, as she reviewed the sample items and the anchor responses on the EQAO website, she told me that she was sure that her teacher would use a different approach to scoring the open-response items than the approach which is used by EQAO. Her understanding of the scoring approach which is used by EQAO is based on reviewing the anchor responses to EQAO open-ended items at the four levels of achievement. As she looked at these anchor responses she indicated that her teacher would not mark the responses in the same manner. Her understanding of the scoring approach used by her teacher was based on the approach he had used on unit tests throughout the term. In

particular, she expressed concern that while her teacher acknowledges the importance of the process students use in responding to open-response items, he does not place nearly as much value on this as EQAO does. She felt that he would place more value on having the correct answer than EQAO does. Thus, while I encouraged her to show as much of her mathematical process as she could as she completed the items on the EQAO assessment, she indicated that because she does not want her teacher's scoring of her EQAO responses to negatively affect her course grade she intended to write her responses to the items according to the way she believes her teacher will score them rather than the approach used by EQAO. She is caught in one of the interactions which I identified in this research and her response, understandably, is to focus on her grades in the course.

As a researcher, I wonder what impact her decision will have on the official EQAO score that she will receive and if many students make this same choice, what impact it will have on the information that EQAO gets from the official scores. As a parent, I am sympathetic to her concerns about the approach that her teacher is taking with regard to this assessment.

With these preparations in place, my daughter set out to take the two-day assessment. She found the booklet on the first day to be challenging both in terms of completing the items in the allowed time and in terms of understanding the intent of the items. She found the booklet on the second day to be even more challenging and was feeling quite discouraged by the negative impact that she fears this assessment will have on her final grade. She indicated that she found the wording of the EQAO items to be confusing and that it was quite different than the wording of the questions she had worked on during the semester, particularly for the multiple-choice items. She would prefer to have an

examination that was more in keeping with the unit tests she has been completing throughout her mathematics course. Still, since she had achieved a score of 94% on the summative performance task for this course just a week earlier, I encouraged her to be positive.

A few days later my daughter returned from school rather dismayed. Her teacher-derived score on the EQAO assessment was only 61%. In addition, her teacher informed the class that many students had not done very well on the EQAO Grade 9 assessment. He also told the students that the mathematics department finds that students often do not do well on the EQAO Grade 9 assessment and is reconsidering the decision to count the assessment for 20% of the final grade. In a conversation I had with him later that week, the teacher explained that all the Grade 9 teachers in his department had marked all the items on the assessment even though they found some of the items to be problematic. He also indicated that given the low scores of many students on the assessment, he and the other teachers had decided to reduce the weighting on the EQAO assessment items in students' grades from 20% to 15%. They had also decided to increase the weighting given to the summative performance task which students had completed a week earlier from 10% to 15%.

As a parent I am confused and dismayed. Why does a student with 94% on the summative performance task get a teacher-derived score of 61% on the EQAO Grade 9 assessment? I wonder how this teacher-derived score will compare with the official score that we will receive from EQAO next fall. Further, if students in this school seem to do poorly on the EQAO Grade 9 assessment most years, why has this mathematics department chosen to use the EQAO assessment in lieu of an in-school final exam? More and more questions come to mind as I consider how my daughter's experience relates to the findings from this inquiry.

My daughters' experience with taking the EQAO Grade 9 assessment and my reaction as a parent have reinforced my interest in conducting further research into this practice and in attempting to better understand other students' and parents' views of the use of this large-scale assessment as part of students' mathematics grades.

REFERENCES

- Abu-Alhija, F. N. (2007). Large-scale testing: Benefits and pitfalls. *Studies in Educational Evaluation, 33*, 50-68.
- American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. (2nd ed.). Washington, DC: Author.
- American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin Supplement, 51*(2), 1-38.
- Baker, E. L., & Linn, R. L. (2002). *Validity issues for accountability systems* (CSE Technical Report. No. 585). Los Angeles, CA: Center for Study of Evaluation; National Center for Research on Evaluation, Standards, and Student Testing; Graduate School of Education and Information Studies; University of California, Los Angeles.
- Barnes, M., Clarke, D., & Stephens, M. (2000). Assessment: the engine of systemic curricular reform? *Journal of Curriculum Studies, 32*(5), 623-650.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 7-74.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.
- Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

- Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1-16). Westport, CT: American Council on Education/Praeger.
- Brookhart, S. M. (2001). Successful students' formative and summative uses of assessment information. *Assessment in Education: Principles, Policy & Practice*, 8(2), 153-169.
- Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues & Practice*, 22(4), 5-12.
- Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record*, 106(3), 429-458.
- Buxton, C. A., Carlone, H. B., & Carlone, D. (2005). Boundary spanners as bridges of student and school discourses in an urban science and mathematics high school. *School Science and Mathematics*, 105(6), 302-312.
- Canadian Psychological Association Working Group on Test Misuse (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*. Ottawa, ON: Canadian Psychological Association.
- Chaiklin, S., & Lave, J. (Eds.). (1996). *Understanding practice: Perspectives on activity and context*. Cambridge, UK: Cambridge University Press.
- Childs, R., Falenchuk, O., & Dunn, J. (2004). *Score reliability for the Grade 9 assessment*. Research report for the EQAO Assessment Review Project.
- Clarke, M. M., Madaus, G. F., Horn, C. L., & Ramos, M. A. (2000). Retrospective on educational testing and assessment in the 20th century. *Journal of Curriculum Studies*, 32(2), 159-181.

- Cobb, P. (1994). Where is the mind? Constructivist and sociocultural perspectives on mathematical development. *Educational Researcher*, 23(7), 13-20.
- Cobb, P., McClain, K., Lamberg, T., & Dean, C. (2003). Situating teachers' instructional practices in the institutional setting of the school and district. *Educational Researcher*, 32(6), 13-24.
- Commission on Instructionally Supportive Assessment. (2001). *Building tests that support instruction and accountability: A guide for policymakers*. Washington, DC: Author.
- Council of Ministers of Education Canada (CMEC). (2009). *PCAP-13 Contextual report on student achievement in reading*. Toronto, ON: Author.
- Cowley, P. & Easton, S. (2008a). *Report card on Ontario's elementary schools*, Toronto, ON: Fraser Institute.
- Cowley, P. & Easton, S. (2008b). *Report card on Ontario's secondary schools*, Toronto, ON: Fraser Institute.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.
- Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302.
- Crundwell, R. M. (2005). Alternative strategies for large scale student assessment in Canada: Is value-added assessment one possible answer [Electronic Version]. *Canadian Journal of Educational Administration and Policy*, 41, 1-7. Retrieved October 31, 2007 from <https://www.umanitoba.ca/publications/cjeap/currentissues.html>

- Cureton, E.E. (1951). Validity. In E.F. Lindquist (Ed.) *Educational measurement* (pp. 621-694). Washington, DC: American Council on Education.
- Davis, B., & Sumara, D. (2006). *Complexity and education: Inquiries into learning, teaching and research*. Mahwah, NJ: Lawrence Erlbaum.
- de Lange, J. (2007). Large-scale assessment and mathematics education. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (2nd revised ed., Vol. 2, pp. 1111-1142). Reston, VA: National Council of Teachers of Mathematics.
- Delandshere, G. (2001). Implicit theories, unexamined assumptions and the status quo of educational assessment. *Assessment in Education: Principles, Policy & Practice*, 8(2), 113-133.
- Delandshere, G. (2002). Assessment as inquiry. *Teachers College Record*, 104(7), 1461-1484.
- Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher*, 27(2), 14-24.
- Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471-515). Westport, CT: American Council on Education/Praeger.
- Dunn, J., Childs, R., & Xu, Y. (2004). *Incomplete student work and scoring approaches for the Grade 9 Assessment of Mathematics*. Research report for the EQAO Assessment Review Project.
- Earl, L. M. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin Press.

Earl, L. M., & Torrance, N. (2000). Embedding accountability and improvement into large-scale assessment: What difference does it make? *Peabody Journal of Education*, 75(4), 114-141.

Education Quality and Accountability Office. (2001). *The development, scoring and reporting of the Grade 9 Assessment of Mathematics, 2000-2001* [Explanatory Paper]. Retrieved January 28, 2008, from <http://www.eqao.com/Parents/Secondary/09/Explanatory.aspx?Lang=E&gr=09&Aud=Parents&gb=true>

Education Quality and Accountability Office. (2006). *Framework: Grade 9 Assessment of Mathematics*. Retrieved July 5, 2008 from <http://www.eqao.com/Educators/Secondary/09/09.aspx?Lang=E&gr=09&Aud=Educators&App=Educators>

Education Quality and Accountability Office. (2007). *Administering the Grade 9 Assessment of Mathematics (for teachers and principals) winter and spring 2007*. Toronto, ON Canada: Queen's Printer.

Education Quality and Accountability Office. (2008). Technical report for the 2006-07 assessments. Toronto, ON Canada: Queen's Printer.

Education Quality and Accountability Office. (2009a). About the EQAO. Retrieved July 13, 2009, from <http://www.eqao.com/AboutEQAO/GeneralQuestions.aspx?Lang=E>

Education Quality and Accountability Office. (2009b). Grade 9 Assessment of Mathematics Frequently Asked Questions. Retrieved July 5, 2010 from <http://www.eqao.com/Educators/Secondary/09/09FAQ.aspx?Lang=E&gr=09>

- Education Quality and Accountability Office. (2009c). *Framework: Grade 9 Assessment of Mathematics*. Toronto, ON Canada: Queen's Printer.
- Education Quality and Accountability Office. (2009d). Technical report for the 2007-2008 assessments. Retrieved October 5, 2009 from <http://www.eqao.com/Educators/Secondary/09/09.aspx?Lang=E&gr=09&Aud=Educators&App=Educators>
- Education Quality and Accountability Office. (2009e). Ontario student achievement English language students 2008-2009 [Web Version]. Retrieved Oct 3, 2009 from http://www.eqao.com/pdf_e/09/Cpr_r_PJ9e_0909_web.pdf
- Edwards, R. (2005, September). *Contexts, boundary zones and boundary objects in lifelong learning*. Paper presented at the British Educational Research Association Annual Conference, University of Glamorgan.
- Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, 36(8), 449-455.
- Emenogu, B. (2003). *Teachers' instructional practices, students' opportunity to learn and differential item functioning in the Grade 9 Assessment of Mathematics*. Research report for the EQAO Assessment Review Project.
- Emenogu, B., Pang, X., & Falenchuk, O. (2004). *Incomplete assessments and the Grade 9 Assessment of Mathematics*. Research report for the EQAO Assessment Review Project.
- Engeström, Y. (1993). Developmental studies of work as a testbench of activity theory: The case of primary care medical practice. In S. Chaiklin & J. Lave (Eds.) *Understanding practice: Perspectives on activity and context* (pp. 64-103). Cambridge, UK: Cambridge University Press.

- Engeström, Y. (1999). Activity theory and individual and social transformation. In Y. Engeström, R. Miettinen, & R. Punamäki (Eds.), *Perspectives on activity theory* (pp. 19-38). Cambridge, UK: Cambridge University Press.
- Engeström, Y. (2001). Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work*, 14(1), 134-156.
- Ercikan, K. (2002). Scoring examinee responses for multiple inferences: Multiple scoring in assessments. *Educational Measurement: Issues and Practice*, 21, 8-14.
- Ercikan, K., & Roth, W. M. (2006). Constructing data. In C. Conrad & R. C. Serlin (Eds.), *Sage handbook for research in education: Engaging ideas and enriching inquiry* (pp. 451-475). Thousand Oaks, CA: Sage.
- Erickson, F. (1986). Qualitative methods in research on teaching. In M. Whittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 119-161). New York: Macmillan.
- Erickson, F. (2004). Demystifying data construction and analysis. *Anthropology and Education Quarterly*, 35(4), 486-493.
- Falenchuk, O., Childs, R. & Dunn, J. (2003). *Scaling of the Grade 9 assessment: Guidelines for calibration*. Research report for the EQAO Assessment Review Project.
- Falenchuk, O., Childs, R. & Dunn, J. (2004). *An evaluation of the alignment approach for the Grade 9 Assessment of Mathematics*. Research report for the EQAO Assessment Review Project.
- Fullan, M. J. (2001). *The new meaning of educational change* (3rd ed.). New York: Teachers College Press.
- Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education*, 24, 355-392.

- Gipps, C. V. (2002). Sociocultural perspectives on assessment. In G. Wells & G. Claxton (Eds.), *Learning for life in the 21st century* (pp. 73-83). Malden, MA: Blackwell.
- Gipps, C. V. (2003). *Educational accountability in England: The role of assessment*. Paper presented at the American Educational Research Association Annual Conference. Retrieved February 11, 2006.
- Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher*, 36(8), 456-462.
- Government of Ontario. (1996). Education Quality and Accountability Office Act. Retrieved 20 October 2006 from <http://www.e-laws.gov.on.ca/DBLaws/Statures/English/96e11>
- Greene, M. (1994). Epistemology and educational research: The influence of recent approaches to knowledge. In L. Darling-Hammond (Ed.), *Review of Research in Education* (Vol. 20, pp. 423-464). Washington, DC: American Educational Research Association.
- Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, 18(4), 5-9.
- Herbert, M., Dunn, J., & Luthra, V. (2004). *The validation of psychometric procedures: Teachers' perceptions and opinions of the Grade 9 Assessment of Mathematics*. Research report for the EQAO Assessment Review Project.
- Herman, J. L., Baker, E. L., & Linn, R. L. (2004). Accountability systems in support of student learning: Moving to the next generation. *Cresst Line - Newsletter of the National Center for Research on Evaluation, Standards & Student Testing*, 1-7.
- Howe, K. R. (2001). Qualitative educational research: The philosophical issues. In V. Richardson (Ed.), *Handbook of research on teaching* (pp. 201-208). Washington, DC: American Educational Research Association.

- Howe, K. R. (2004). A critique of experimentalism. *Qualitative Inquiry*, 10(1), 42-61.
- Johnson, D. D., & Johnson, B. (2006). *High stakes: Poverty, testing and failure in American schools* (2nd ed.). Lanham, MD: Rowman & Littlefield.
- Joint Committee on Standards for Educational Evaluation (JCSEE). (2003). *The student evaluation standards: How to improve evaluations of students*. Thousand Oaks, CA: Corwin Press & E.T.S. Educational Policy Leadership Institute.
- Joint Committee on Testing Practices (JCTP). (2004). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.
- Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement*, 58(5), 736-753.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: American Council on Education/Praeger.
- Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher*, 37(2), 76-82.
- Kaplan, A. (1997). Research methodology: Scientific methods. In L. J. Saha (Ed.), *International encyclopedia of the sociology of education* (pp. 234-241). New York: Pergamon.
- Kitto, R. J. (2006). *Teacher perceptions of the validity of a large-scale mathematics assessment instrument*. Unpublished doctoral dissertation. University of Western Ontario, London, ON, Canada.

- Klinger, D. A., DeLuca, C., & Miller, T. (2008). The evolving culture of large-scale assessments in Canadian education. *Canadian Journal of Educational Administration and Policy*, 76, 1-34.
- Klinger, D., Falenchuk, O., & Pang, X. (2004). *Simultaneous versus separate calibration of the Grade 9 Assessment of Mathematics*. Paper commissioned by the EQAO Assessment Review Project.
- Kohn, A. (1999). *The schools our children deserve: Moving beyond traditional classrooms and "tougher standards"*. New York: Houghton Mifflin.
- Koretz, D., & Hamilton, L. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531-578). Westport, CT: American Council on Education / Praeger.
- Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio program: Findings and implications. *Educational Measurement: Issues & Practice*, 13(3), 5-16.
- Lather, P. (1993). Fertile obsession: Validity after poststructuralism. *The Sociological Quarterly*, 34(4), 673-693.
- Lather, P. (2001). Validity as an incitement to discourse: Qualitative research and the crisis of legitimation. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 241-250). Washington, DC: American Educational Research Association.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.
- Levin, B. (1998). An epidemic of education policy: (What) can we learn from each other? *Comparative Education*, 34(4), 131-141.

- Lincoln, Y. S. (1995). Emerging criteria for quality in qualitative and interpretive research. *Qualitative Inquiry*, 1(3), 275-289.
- Lincoln, Y. S., & Guba, E. G. (2000). Paradigmatic controversies, contradictions, and emerging confluences. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 163-188). Thousand Oaks, CA: Sage.
- Lissitz, R.W. (2009). *The concept of validity: Revisions, new directions and applications*. Charlotte, NC: Information Age Publishing.
- Lissitz, R. W., & Samuelsens, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437-448.
- Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice*, 27(3), 32-42.
- Lock, C. L. (2001). *The influence of a large-scale assessment program on classroom practices*. Unpublished doctoral dissertation, Queen's University, Kingston, ON.
- Luce-Kapler, R., & Klinger, D. A. (2005). Uneasy writing: The defining moments of high-stakes literacy testing. *Assessing Writing*, 10, 157-173.
- Mason, J. (2006). Mixing methods in a qualitatively driven way. *Qualitative Research*, 6(1), 9-25.
- Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16(2), 16-18.
- Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30(10), 955-966.

- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education / Macmillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.
- Messick, S. (1998). Test validity: A matter of consequence, *Social Indicators Research*, 45, 35-44.
- Mishler, E. G. (1990). Validation in inquiry-guided research: The role of exemplars in narrative studies. *Harvard Educational Review*, 60(4), 415-442.
- Mislevy, R. J. (2007). Validity by design. *Educational Researcher*, 36(8), 463-469.
- Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 83-108). Charlotte, NC: Information Age Publishing.
- Mislevy, R. J., Moss, P. A., & Gee, J. P. (2009). On qualitative and quantitative reasoning in validity. In K. Ercikan & W.-M. Roth (Eds.), *Generalizing from educational research: Beyond qualitative and quantitative polarization* (pp. 67-100). New York: Routledge.
- Moran-Ellis, J., Alexander, V. D., Cronin, A., Dickinson, M., Fielding, J., Sleney, J., et al. (2006). Triangulation and integration: Processes, claims and implications. *Qualitative Research*, 6(1), 45-59.
- Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62, 229-258.
- Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, 23(2), 5-12.
- Moss, P. A. (1998). Recovering a dialectical view of rationality, *Social Indicators Research*, 45, 55-67.

- Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice*, 22(4), 13-25.
- Moss, P. A. (2005). Toward "epistemic reflexivity" in educational research: A response to Scientific Research in Education. *Teachers College Record*, 107(1), 19-29.
- Moss, P. A. (2007). Reconstructing validity. *Educational Researcher*, 36(8), 470-476.
- Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education*, 30, 109-162.
- Nagy, P. (2000). The three roles of assessment: Gatekeeping, accountability, and instructional diagnosis. *Canadian Journal of Education*, 25(4), 262-279.
- National Council of Teachers of Mathematics (NCTM). (1995). *Assessment standards for school mathematics*. Reston, VA: Author.
- Ontario Ministry of Education (OME). (2005). *The Ontario curriculum Grades 9 and 10 mathematics revised*. Toronto, ON Canada: Queen's Printer.
- Ontario Ministry of Education (OME). (2009) The Ontario curriculum. Retrieved October 1, 2009 from <http://www.edu.gov.on.ca/eng/curriculum/secondary/index.html>
- Ottawa-Carleton District School Board (OCDSB). (2008). Assessment, evaluation and reporting of student achievement (PR.584.CUR). Retrieved December 12, 2009 from http://www.ocdsb.edu.on.ca/au_pp_main.asp#curriculum
- Pang, X. (2004). Grade 9 Assessment of Mathematics, 2003-04 *The Impact of informing students that part of the assessment would be used to determine final grades in their mathematics class*. Internal report of Education Quality & Accountability Office.
- Phelps, R. P. (2000). Trends in large-scale testing outside the United States. *Educational Measurement: Issues & Practice*, 19(1), 11-21.

- Popham, W. J. (1997). Consequential validity: Right concern - wrong concept. *Educational Measurement: Issues and Practice*, 17(2), 6-12.
- Popham, W. J. (2002). Right task, wrong tool. *American School Board Journal*, 189(2), 18-22.
- Rabinow, P., & Sullivan, W. M. (1979). The interpretive turn: Emergence of an approach. In P. Rabinow & W. M. Sullivan (Eds.), *Interpretive social science: A reader* (pp. 1-21). Berkeley, CA: University of California Press.
- Ricci, C. (2004). The case against standardized testing and the call for a revitalization of democracy. *The Review of Education, Pedagogy, and Cultural Studies*, 26(4), 339-361.
- Ross, J. A., & Gray, P. (2005). *Cross-validation of classroom testing and mandated external assessments*. Paper presented at the Joint Meeting of the American Evaluation Association and Canadian Evaluation Society.
- Schwandt, T. A. (1996). Farewell to criteriology. *Qualitative Inquiry*, 2(1), 58-72.
- Seidman, I. (2006). *Interviewing as qualitative research* (3rd ed.). New York: Teachers College Press.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405-450.
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5-8,13,24.
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4-14.
- Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 1066-1101). Washington, DC: American Educational Research Association.

- Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623-646). Westport, CT: American Council on Education/Praeger.
- Sicoly, F. (2002). What do school-level scores from large-scale assessments really measure? *Educational Measurement: Issues & Practice*, 17-26.
- Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, 36(8), 477-481.
- Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19-37). Charlotte, NC: Information Age Publishing.
- Slomp, D. (2008). Harming not helping: The impact of a Canadian standardized writing assessment on curriculum and pedagogy. *Assessing Writing*, 13, 180-200.
- Slomp, D. H., & Fuite, J. (2005). Following Phaedrus: Alternate choices in surmounting the reliability/validity dilemma. *Assessing Writing*, 9(3), 190-207.
- Stake, R. E. (1995). *The art of case study research*. Thousand Oaks, CA: Sage.
- Stake, R. E. (2000). Case studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 435-454). Thousand Oaks, CA: Sage.
- Stake, R. E. (2008). Qualitative Case Studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *Strategies of qualitative inquiry* (3rd ed., pp. 1-43). Thousand Oaks, CA: Sage.
- Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, 19(3), 387-420.
- Stevenson, R. B. (2004). Constructing knowledge of educational practices from case studies. *Environmental Education Research*, 10(1), 39-51.

- Stiggins, R. (2005). From formative assessment to assessment for learning: A path to success in standards-based schools. *Phi Delta Kappan*, 87(4), 324-328.
- Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. London: Routledge.
- Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research*, 51(2), 161-179.
- Suurtamm, C., & Graves, B. (2007). *Curriculum Implementation in Intermediate Math (CIIM) [Research Report, Year 1]*. Ottawa, ON, Canada: Faculty of Education, University of Ottawa.
- Suurtamm, C., Lawson, A., & Koch, M. J. (2008). The challenge of maintaining the integrity of reform mathematics in large-scale assessment. *Studies in Educational Evaluation*, 34(1), 31-43.
- Suurtamm, C., & Moisan, P. (2004). *Examining the multiple versions of the Grade 9 Assessment of Mathematics*. Paper commissioned by the EQAO Assessment Review Project.
- Suurtamm, C., Moisan, P., & Luthra, V. (2004a). *Does the mathematics being measured by the Grade 9 Assessment of Mathematics represent the “important” mathematics to be learning in the Grade 9 curriculum?* Paper commissioned by the EQAO Assessment Review Project.
- Suurtamm, C., Moisan, P., & Luthra, V. (2004b). *How well does the Grade 9 Assessment of Mathematics reflect the Ontario curriculum?* Paper commissioned by the EQAO Assessment Review Project.
- Taylor, C. (1979). Interpretation and the sciences of man. In P. Rabinow & W. M. Sullivan (Eds.), *Interpretive social science : A reader* (pp. 25-71). Berkeley, CA: University of California Press.

- Tierney, R. D. (2006). Changing practices: influences on classroom assessment. *Assessment in Education: Principles, Policy & Practice*, 13(3), 239-264.
- Tittle, C. K. (1989). Validity: Whose construction is it in the teaching and learning context? *Educational Measurement: Issues and Practice*, 8(1), 5-13, 34.
- Toulmin, S. E. (2003; 1958). *The uses of argument*. Cambridge: Cambridge University Press.
- Venkat, H., & Adler, J. (2008). Expanding the foci of activity theory: Accessing the broader contexts and experiences of mathematics education reform. *Educational Review*, 60(2), 127-140.
- Volante, L. (2007). Education quality and accountability in Ontario: Past, present, and future. *Canadian Journal of Educational Administration and Policy*, 58, 9.
- Wertsch, J.V. (1995). The need for action in sociocultural research. In J.V. Wertsch, P. del Rio, & A. Alvarev (Eds.). *Sociocultural studies of mind* (pp. 56-74). Cambridge: Cambridge University Press.
- Wertsch, J.V. (1998). *Mind as action*. Oxford: Oxford University Press.
- Wiliam, D. (2004). *The validity and impact of educational assessments*. Background paper prepared for Discussion Group 12 at the 10th International Congress on Mathematical Education, July 2004, Copenhagen, Denmark.
- Wiliam, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (2nd revised ed., Vol. 2, pp. 1053-1098). Reston, VA: National Council of Teachers of Mathematics.

- Wilson, L. D. (2007). High-stakes testing in mathematics. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (2nd revised ed., Vol. 2, pp. 1099-1110). Reston, VA: National Council of Teachers of Mathematics.
- Wilson, L.D., & Kenney, P.A. (2003). Classroom and large-scale assessment. In J. Kilpatrick, W.G. Martin & D. Shifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 53-67). Reston, VA: National Council of Teachers of Mathematics.
- Wolcott, H. F. (1990). On seeking - and rejecting - validity in qualitative research. In E. W. Eisner & A. Peshkin (Eds.), *Qualitative inquiry in education: The continuing debate* (pp. 121-152). New York: Teachers College Press.
- Wolfe, R., Childs, R. A., & Elgie, S. (2004). *Final report of the external evaluation of EQAO's assessment processes [Ensuring quality assessments: A project to refine and affirm assessment processes]*. Toronto, ON, Canada: Education Quality and Accountability Office & Ontario Institute for Studies in Education, University of Toronto.
- Yin, R. K. (2003). *Case study research: Design and methods* (3rd ed.). Thousand Oaks, CA: Sage.
- Yin, R. K. (2006). Case study methods. In J. L. Green, G. Camilli, P. B. Emore, A. Skukauskaite, & E. Grace (Eds.), *Handbook of complementary methods in education research* (pp. 111-122). Washington, DC: American Educational Research Association/Lawrence Erlbaum.
- Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol 26: Psychometrics* (pp. 45-79). Amsterdam: North Holland.

Appendix A
EQAO Multiple-choice Items (Grade 9 Applied)

Multiple-Choice

Grade 9 Assessment of Mathematics, Winter 2007

- 1** The ratio of girls to boys in Mr. Dilworth's music class is 18:12. What percent of the students are boys?

A 12%
B 40%
C 60%
D 67%

- 2** Which of the following expressions is equal to 49?

F $(4 + 3)^2$
G $4 + 3^2$
H $(3 + 4^2)$
J $3^2 + 4^2$

- 3** Herman uses his scientific calculator to determine the value of the expression $\frac{4}{3}\pi(3.5)^3$.

What is the **approximate** value of this expression?

A 175.0
B 175.1
C 179.0
D 179.6

- 4** Ivan shows his steps in solving the following equation for x :

$$2x + 3 = 7$$

Step 1: $2x + 3 - 3 = 7 - 3$

Step 2: $2x = 4$

Step 3: $x = 4 \times 2$

Step 4: $x = 8$

In which step has Ivan made an **error**?

F Step 1
G Step 2
H Step 3
J Step 4

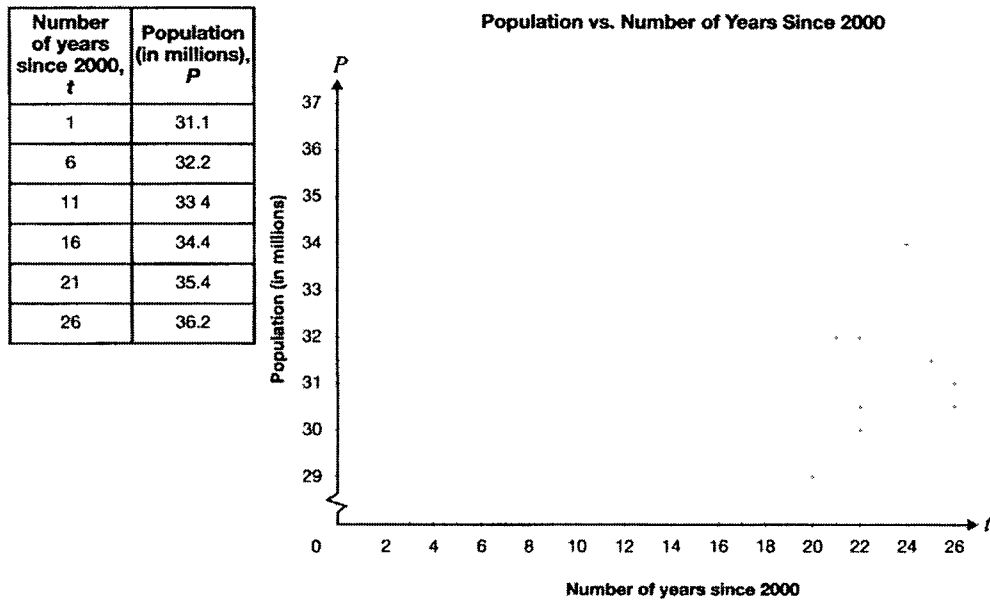
Appendix B
EQAO Open-response Item (Grade 9 Academic)

Open-Response

Grade 9 Assessment of Mathematics, Winter 2007

10 Population Plans

Alvin is researching the population of Canada. He finds data for the year 2001 and predictions for every 5 years after that, as shown below.



Determine an algebraic model for Alvin's data, and use it to make a reasonable prediction for the population of Canada in 2036.

Justify your answer.

Appendix C
EQAO Generic Rubric for Open-Response Items

Code	Descriptor
Blank	blank nothing written or drawn in response to the question
Illegible/ Off topic	- 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
Code 10	- demonstration of limited understanding of concepts and/or procedures - application of knowledge and skills shows limited effectiveness due to - misunderstanding of concepts - incorrect selection or misuse of procedures - problem-solving process shows limited effectiveness due to - minimal evidence of a solutions process - limited identification of important elements of the problem - too much emphasis on unimportant elements of the problem - no conclusions presented - conclusion presented without supporting evidence
Code 20	- demonstration of some understanding of concepts and/or procedures - application of knowledge and skills shows some effectiveness due to - partial understanding of the concepts - errors and /or omission in the application of the procedures - problem-solving process shows some effectiveness due to - an incomplete solutions process - identification of some of the important elements of the problem - some understanding of the relationships between important elements of the problem - simple conclusions with little supporting evidence
Code 30	- demonstration of considerable understanding of concepts and/or procedures - application of knowledge and skills shows considerable effectiveness due to - an understanding of most of the concepts - minor errors and/or omissions in the application of the procedures - problem-solving process shows considerable effectiveness due to - a solution process that is nearly complete - identification of most of the important elements of the problem - a considerable understanding of the relationships between important elements of the problem - appropriate conclusions with supporting evidence
Code 40	- demonstration of thorough understanding of concepts and/or procedures - application of knowledge and skills shows a high degree of effectiveness due to - a thorough understanding of the concepts - an accurate application of the procedures (any minor errors and/or omissions do not detract from the demonstration of a thorough understanding) - problem-solving process shows a high degree of effectiveness due to - a complete solution process - identification of all important elements of the problem - a thorough understanding of the relationships between all of the important elements of the problem - appropriate conclusion with thorough and insightful supporting evidence

Source EQAO (2009b), p 32

Appendix D

EQAO Item Specific Rubric for Population Plans Item

Assessment of Mathematics Grade 9 Academic Program Specific Open-Response Scoring Guide Population Plans (Winter 2007)

- B = Blank nothing written or drawn in response to the question
 I = • 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" "I don't know")
 • Off topic no relationship of written work to the question

Codes	Description
10	Problem-solving process involving an algebraic model to determine a prediction of the population in 2036 shows limited effectiveness due to • minimal evidence of a solution process • limited identification of important elements of the problem • too much emphasis on unimportant elements of the problem • no conclusions presented • conclusion presented without supporting evidence
20	Problem-solving process involving an algebraic model to determine a prediction of the population in 2036 shows some effectiveness due to • an incomplete solution process • identification of some of the important elements of the problem • some understanding of the relationships between important elements of the problem • simple conclusions with little supporting evidence
30	Problem-solving process involving an algebraic model to determine a prediction of the population in 2036 shows considerable effectiveness due to • a solution process that is nearly complete • identification of most of the important elements of the problem • a considerable understanding of the relationships between important elements of the problem • appropriate conclusions with supporting evidence
40	Problem-solving process involving an algebraic model to determine a prediction of the population in 2036 shows a high degree of effectiveness due to • a complete solution process • identification of all important elements of the problem • a thorough understanding of the relationships between all of the important elements of the problem • appropriate conclusions with thorough and insightful supporting evidence

Appendix E
EQAO Item Population Plans Anchor Response

ANCHOR

Academic Winter 2007 Q10

Name: Population Plans

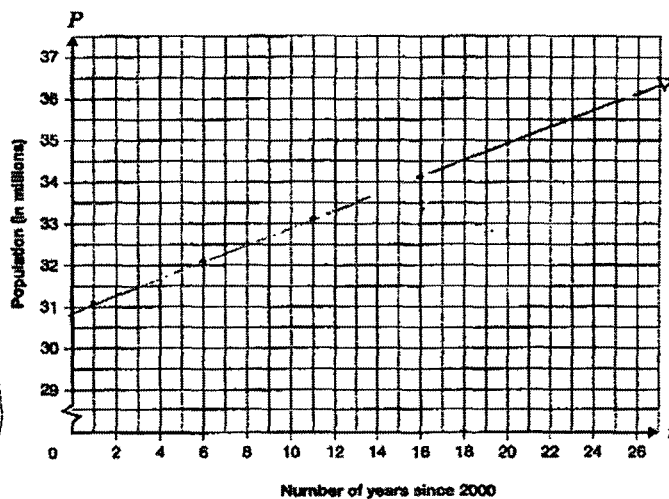
CODE: 30

Nina is researching the population of Canada. She finds data for the year 2001 and predictions for every 5 years after that, as shown below.

Population vs. Number of Years Since 2000

Number of years since 2000, t	Population (in millions), P
1	31.1
6	32.2
11	33.4
16	34.4
21	35.4
26	36.2

31 37
36 38



Determine an algebraic model for Nina's data, and use it to make a reasonable prediction for the population of Canada in 2036.

Justify your answer.

A (6, 32.2) B (26, 36.2)

∴ the algebraic model is $y = \frac{1}{5}x + 31.1$.

$$\text{Slope AB} = \frac{\Delta y}{\Delta x} = \frac{32.2 - 36.2}{6 - 26} = \frac{-4}{-20} = \frac{4}{20} = \frac{1}{5}$$

$$\begin{aligned} y &= \frac{1}{5}x + 31.1 \\ &= \frac{1}{5}(36) + 31.1 \\ &= 38.3 \end{aligned}$$

∴ the population of Canada in 2036 will be approximately 38.3 million. because every 5 years, the population increases about 1 million.

Rationale:

Student's problem-solving process demonstrates a considerable understanding of the relationships between all the important elements of the problem; student determines an algebraic model and plots graph; shows evidence of points used for the slope (on the line), but uses value for y-intercept when $t = 1$, not $t = 0$. Student uses their algebraic model used to determine population.

Appendix F
Interview Protocol for Department Heads & Teachers

A. Getting started

Before we talk about your use of the Grade 9 EQAO assessment, would you please tell me a little about your current teaching assignment and your past experience in teaching mathematics?

- What courses are you teaching this year?
- How long have you been at this school?
- How many years have you been teaching?
- How many years have you taught grade 9 math?

B. Open-ended prompt

As you know from the description on the consent form, the main focus of my research is to get a better understanding of the use of the grade 9 EQAO assessment as part of students' grades. Can you describe how this is done with your class? Please provide as much detail as you can.

C. Additional prompts

Depending on the level of detail offered in response to the open-ended prompt ask participant "Can you tell me a little more about"

How you choose the items you will score?

- Are different kinds of items chosen for the Academic and Applied courses?
- What about the field-test items, how do you deal with them when choosing items to score?

How you score the items you have chosen?

- Does EQAO provide any materials to guide you in scoring the items?
- Do you score for correct answer only or do you look at the process used?
- Does each teacher score their own students' work or is scoring done collaboratively in your department?

How does your use of the EQAO Grade 9 assessment fit with the other assessment strategies you are using for the Grade 9 Academic and/or Applied courses?

- How much does the EQAO assessment contribute to students' grades?
- Do you use the EQAO assessment for final grades or as part of term grades?
- Does it count for the same amount for both Grade 9 Academic & Grade 9 Applied courses?
- What information do you get about students' math learning from scoring the EQAO items?

- What information do you give to students about how you will use of the EQAO assessment as part of their grades?

What are your reasons for using EQAO items as part of students' final grades?

- In what ways do you think this use of the EQAO assessment is beneficial?
- Do you see any difficulties related to this use of the EQAO assessment?

How are decisions regarding this use of the EQAO assessment made in this school?
Does the department as a whole decide or does each teacher decide for themselves?

- Are there any school policies which are related to this use of the assessment?
- Do you know of any board policies regarding this use of the assessment?

D. Closing Prompt

Are there any other comments you would like to make about using the EQAO assessment as part of students' grades?

E. Final Remarks

Thank you for participating.

After I leave you may think of other comments you would like to make about this topic. If you do please feel free to get in touch with me using the contact information which I have provided on the consent form.

Appendix G

Interview Protocol for Principals

A. Getting started

Before we talk about the use of the Grade 9 EQAO assessment as part of students' grades, could you please tell me a little about your experience as a Principal?

- How many years have you been a principal?
- How long have you been the principal at this school?

B. Open-ended prompt

As you know from the description on the consent form, the main focus of my research is to get a better understanding of the use of the grade 9 EQAO assessment as part of students' grades. I am interested in learning more about how this is done in schools and about any policies which may have been established related to this practice? Can you describe what happens here at "name of school"? Please provide as much detail as you can.

C. Additional prompts

Depending on the level of detail offered in response to the open-ended prompt ask participant "Can you tell me a little more about"

- When the use of the EQAO Grade 9 assessment as part of students' grades first started in this school?
- What reasons do you see for using the EQAO items as part of students' grades?
- In what ways do you think this use of the EQAO assessment is beneficial?
- Do you see any problems or difficulties related to this use of the EQAO assessment?
- Are the decisions regarding this use of the EQAO made by the department as a whole or does each teacher decide for themselves?
- Are there any school policies related to this use of the assessment? Are they any board policies? Are you aware of any Ministry of Education policies which relate to this use?
- Do you know of other schools which use the EQAO items in this way?
- Do you know of schools with a different approach to using the EQAO grade 9 assessment?

D. Closing prompt

Are there any other comments you would like to make about using the EQAO assessment as part of students' grades?

E. Final Remarks

Thank you for participating. After I leave you may think of other comments you would like to make about this topic. If you do please feel free to get in touch with me using the contact information which I have provided on the consent form.

Appendix H

Interview Protocol for EQAO Personnel

A. Background questions

- What is your current position at EQAO?
- In what ways are you involved with the grade 9 assessment?
- How long have you been doing work related to the grade 9 assessment?

B. Open-ended prompt

As you know from the description on the consent form, the main focus of my research is to get a better understanding of teachers' use of the EQAO Grade 9 assessment as part of students' grades. I am interested in learning more about EQAO's view of this practice. Can you describe the policies which have been established related to this practice? What is your sense of the reasons for EQAO establishing these policies? Please provide as much detail as you can.

C. Additional prompts

Depending on the level of detail offered in response to the open-ended prompt ask participant "Can you tell me a little more about"

- In the 2007 guide for teachers and principals administering the EQAO Grade 9 assessment there is a section which describes teachers' use of EQAO assessment items (refer to p.22.) Do you know when this policy was established?
- Does EQAO provide any materials for teachers to guide their scoring of these items?
- In what ways do you think this use of the EQAO assessment for final marks is beneficial?
- Do you see any problems or difficulties related to this use of the EQAO assessment?
- I understand that the use of the assessment items as part of students' grades was not a part of the initial intended use for the assessment. Do you think that there are elements of the design of the assessment which may need to be reconsidered given this use?
- What does the use of the EQAO assessment items as part of students' grades do to the comparability of results across schools and boards?

Closing prompt

Are there any other comments you would like to make about using the EQAO assessment as part of students' grades?

E. Final Remarks

Thank you for participating. After I leave you may think of other comments you would like to make about this topic. If you do, please feel free to get in touch with me using the contact information which I have provided on the consent form.