DOES THIS COUNT FOR MARKS? A MIXED METHODS ANALYSIS OF TEACHER OPINION OF COUNTING A LARGE-SCALE ASSESSMENT TOWARD CLASS GRADES

By

Sherry King

For

Advisor: Dr. Christina van Barneveld

Committee Member: Dr. Ann Kajander

A thesis submitted in partial fulfillment of

the requirements for the degree of

MASTER OF EDUCATION

in

The Faculty of Graduate Studies (Education)

Lakehead University

Thunder Bay, ON

August, 2011

Acknowledgements

I would like to express my most sincere thanks to my thesis supervisor, Dr. Christina van Barneveld, for her great generosity, wisdom, guidance, encouragement, honesty and commitment to my academic journey. It has been a journey of personal growth and learning that has taught me as much about rigor, perseverance and teamwork as it has broadened my scope of understanding of the topics of this research. I would also like to thank Dr. Ann Kajander, Dr. Ruth Childs, and Dr. Joan Chambers for their careful consideration of my work and most helpful recommendations for development and improvement. Thank you all for not allowing me to go half way with anything.

Thank you to the Education Quality and Accountability Office for their assistance with this research and for providing the data that made this study possible.

It was a great pleasure to work with the Ottawa team, and I would like to thank Dr. Marielle Simon, Catherine Nadon, and the rest of the team who offered ideas and support throughout the development and writing of this thesis. Here at home, I will always appreciate the care and concern shown by Carlos Zerpa, Melissa Hill, and of course Diana Mason. I doubt few graduate students would survive without her. Thank you to Dr. Constance Russell and the Graduate Studies Office at Lakehead University for their assistance with processing the final document at the end of a long, hot summer.

Thank you to my friends and colleagues at Confederation College for their overwhelming support and encouragement. Finally but not last, thank you to God, my mother, Shirley, and to my dogs for guarding my spirit throughout.

Table of Contents

Acknowledgements……..... 2

Table of Contents..... 3

Abstract……..... 6

List of Tables..... 7

List of Figures..... 8

CHAPTER 1: Introduction..... 9

1.1. Overview of the Issues..... 9

1.2. Theoretical Framework…..... 13

1.3 Statement of the Problem..... 18

1.4. Purpose of the Study..... 20

1.5. The Research Questions..... 22

1.6. The Importance of the Study..... 23

1.7 The Scope of the Study..... 24

1.8 Limitations of the Study..... 25

1.9 Definitions of Key Terms..... 26

CHAPTER 2: Review of the Literature..... 28

2.1 Validation of Standardized Assessments..... 28

2.2 Achievement Motivation Theories..... 34

2.3 Student Motivation, Test Stakes and CIV …..... 36

2.4 Implications for the EQAO Grade 9 Assessment of Mathematics ……..... 45

CHAPTER 3: Research Methods ……..... 49

3.1 Research Design …….....…. 49

3.2 Data Collection …….....……. 51

3.3 Participants …….....…… 52

3.4 Instrument …….…….…..…..... 53

3.5 Data Analysis ……..….....…. 54

3.6 Ethical Considerations …….…….....…. 56

CHAPTER 4: Results ……..……..... 58

4.1 Descriptive Statistics..... 58

4.2 Formulated Meanings for Teacher Responses to the Constructed-Response Part of Question..... 60

4.3 Coding and Qualitative Analysis of Teacher Comments …… 63

4.3.1. Yes – Teacher response to Q 22 with comment …….……..… 66

4.3.2. Undecided – Teacher response to Q22 with comment ……..…. 70

4.3.3. No – Teacher response to Q22 with comment ……..… 72

CHAPTER 5: Discussion ……. 74

5.1 Frequencies of Responses..... 74

5.2 Variations in Opinion by Response..... 75

5.3 Informing Expectancy-Value Theory of Motivation..... 77

5.4 Informing Validity Theory in the Context of Large-Scale Assessments..... 80

5.5 Convergence of Data..... 82

CHAPTER 6: Summary and Conclusions ……..... 83

6.1 Recommendations for Further Study ……..... 84

REFERENCES ……..... 85

APPENDIX

Abstract

The objective of this research was to gain insights on teacher opinions related to the practice of counting a portion of a large-scale educational assessment (LSA) for students‟ grades. I analyzed the responses of 1,203 Grade 9 teachers who responded to a question on a Teacher Questionnaire administered as part of a grade 9 LSA of mathematics. Teacher opinions were clustered according to two dimensions (1) whether the teacher opined that counting the LSA for student marks motivated the students to take the assessment more seriously, and (2) whether the teacher comment related to characteristics of the student, student behaviours, characteristics of the test, or the value students placed on the LSA. The results were that most of the teachers (85%) opined that counting the LSA for student marks motivated students to take the test more seriously because it raised the value of the LSA for the students, and that translated into improved test preparation and/or effort on the test. When teachers opined that counting the LSA did not motivate students, or when the teachers were undecided, teacher comments tended to focus on characteristics of the students such as program (academic or applied), ability, and anxiety.

List of Tables

Table 1. Frequencies for selected and constructed-response questions (Q22)……59

Table 2. Selected examples of significant comments by teachers and related formulated meanings…….....60

Table 3. Frequency of theme clusters as a percentage of total for each selected-response ……..……..……..62

Table 4. Examples of theme clusters with their formulated meaning by teacher responses ……...……..……64

List of Figures

Figure 1. Conceptual model of test-taking motivation: Themes and Codes.....16

Figure 2. Question 22 from the EQAO 2010 Teacher Questionnaire.....54

Chapter 1

Introduction

It is the goal of educators and test designers working in the field of educational measurement to ensure that test design and administration result in student performance outcomes, as measured by assessment instruments, which are useful indicators of students‟ actual proficiency in the subject domain. We do this with the assumption that students put forth effort on the assessment; however, some students may not fit this generalization, and test performance may not accurately represent students‟ actual proficiency. This is of particular concern for low-stakes, large-scale assessments (LSAs) where the value or importance of the assessment may be inconsequential for the student.

This thesis reports on teacher opinion of counting all or part of a large-scale mathematics assessment toward students‟ final grades as a way to motivate students to take the assessment more seriously. This chapter is organized in the following sections: (1) overview of the issues, (2) statement of the problem, (3) theoretical framework, (4) purpose of the study, (5) research questions, (6) importance of the study, (7) scope of the study, (8) limitations of the study, and (9) definition of key terms.

1.1 Overview of the Issues

In 1996, in response to recommendations from its 1995 Royal Commission on Learning, the Ontario government established the Education Quality and Accountability Office (EQAO) as a Crown agency. The organization‟s mandate is to develop, administer, mark and report on province-wide tests of student achievement and to ensure that its tests provide “credible evidence of student learning based on The Ontario Curriculum” (EQAO, 2010a, p. 3). In addition to reporting results to students, schools, and boards, the data collected from the assessment results

and student and teacher questionnaires are used by the Ministry of Education and the province‟s school boards for student learning improvement planning (Office of the Auditor General of Ontario, 2009). The results of the Grade 9 Assessment of Mathematics may also be used by teachers in the calculation of students‟ final grades. The decision of whether to count the assessment toward class grades and what parts of the assessment to count is made by the board, the principal, the principal and department heads, or individual teachers depending on the board and the school.

The following examples demonstrate some of the ways this assessment is used for multiple purposes: for measuring student achievement, to drive curriculum, in the calculation of students‟ grades, and to guide planning and improvement processes at the school, board, and ministry levels. When assessment scores are used for multiple purposes, the interests of multiple stakeholders are potentially affected by the outcome of the assessment (Koch, 2009). As a result of the multiple uses described, several groups of stakeholders can be identified for this assessment including students, parents, teachers, schools, boards, and the provincial government. There are potential consequences for each identified group, so it is very important that the assessment effectively measure student achievement and that the intended use(s) can be defended with sound arguments for validity.

In the years since the first administration of the EQAO assessments, validity arguments have been called into question. Validity concerns included narrowing the curriculum by teaching to the test, the effects of stress on student performance, and concerns over students with special educational needs and students receiving specialized programming being placed at a disadvantage due to the difficulty, literacy demands or formatting of the assessment (Cheng, Klinger, Zheng, 2007; Fairbairn & Fox, 2009; Johnstone, 2003; Wolf, Smith, & Birnbaum,

1995). Since 2006, the EQAO has developed and expanded the options for accommodations and special provisions to better address the needs of English second language (ESL) learners and students with Individual Education Plans (IEPs) in an effort to make the administration of the assessment fair to all students and to reduce the number of student exemptions (EQAO, 2008).

In the 2006-2007 report of results, the EQAO states that exemptions are no longer permitted, yet there remain provisions for exemptions in the policy documents, and exemptions are still granted by the principal of the school in consultation with the student, parents/guardian and teacher. Since 2007, the reporting of exemptions has changed, and eligible students who are granted exemptions are now placed in the "no data" category on the EQAO report on results (EQAO, 2010b, p. 2). This change in reporting makes it difficult to account for all of the eligible students in the no data category, and that number ranges from 1% for students in the Grade 9 academic math course to 5% in the Grade 9 applied mathematics course (EQAO, 2010b). Based on the data provided by the EQAO, it is not possible to determine what percentage of these students deferred writing, were exempt, or did not attend to write the test.

There is an additional group of students considered non-eligible to write the assessment, and whether this group is exempt or ignored is not clear. Eligible students are defined by the EQAO as all students working toward an academic or applied Grade 9 mathematics credit. The two Grade 9 mathematics credit courses, designated academic and applied, differ in their degree of difficulty and the extent to which attention is given to theory and abstract thinking (Ontario Ministry of Education, 2005, p. 6). Successful completion of either the academic or applied course is a credit toward the Ontario Secondary School Diploma, and all students enrolled in academic or applied Grade 9 mathematics courses are considered eligible and are required to participate in the EQAO assessment.

Students considered non-eligible to write the assessment are students not enrolled in either a Grade 9 applied or academic mathematics course. These students may be completing locally developed math courses, and they do not write the EQAO Grade 9 Assessment of Mathematics (EQAO, 2011) because they are not receiving the curriculum that is assessed by the test. How the EQAO accounts for these students is not clear. Exemption applies to eligible students, and these students do not fall in that category. If they were exempt and included in the reports on results, they would be included in the overall results for each school and assessed as not achieving the provincial standard, but because they are considered non-eligible, it is unlikely that they are included in the reports on results, but again, this is not clearly communicated.

An additional anomaly was discovered while reviewing the 2009-2010 enrolment data provided by the Ontario Ministry of Education (2011). There appears to be approximately 5% of the total number of Grade 9 students enrolled in English speaking public schools missing from the EQAO report on results for that year. In 2010, the ministry reported that 156,952 students were enrolled in Grade 9 English language schools in the province; however, the EQAO report for the same year indicates a total of 148,834 students, so there is a discrepancy of about 5%. These may be non-eligible students, but with the information available, it was not possible to determine the source of the discrepancy. If the students not reported are non-eligible, it is important to consider if they should be included in the EQAO report in some way.

Other students that may not be included in the EQAO data includes students enrolled in independents schools or students receiving their education while in treatment or correctional facilities. Depending on the policy of the independent school, the institution or the circumstances of the student, these students may or may not participate in the EQAO assessments. Reporting on the actual number of Grade 9 students in Ontario who are not included in the EQAO Grade 9

Assessment of Mathematics, if well documented, does not appear to be reported in the policy documents or reports on results that are made available to educators and the general public, and this omission may have implications with regard to the use of assessment data in measuring student achievement in Ontario and should be addressed.

The ongoing revisions to the assessment and administration policies since the initial running of the assessments show that the EQAO is concerned with issues of validity, reliability and fairness. More recently, EQAO has expressed an interest in examining how factors may influence student motivation as a potential threat to the validity of the Grade 9 Assessment of Mathematics. Variables like motivation can affect student performance on the test and reduce the effectiveness of the measurement of student achievement derived from the test because these kinds of variables represent potential sources of construct-irrelevant variance (CIV) (Lam & Bordignon, 2001; Suurtamm, Lawson, & Koch, 2008; Wolfe, Childs, & Elgie, 2004). This is variance in the test scores that is irrelevant to the interpreted construct (Wise, Bhola, & Yang, 2006; Wise & DeMars, 2010). If sources of CIV such as student motivation are found to be significant, the assessment may fail to satisfy validity arguments for its intended use(s) (Haladyna & Downing, 2004).

1.2 Theoretical Framework

This study is grounded in two theoretical frameworks. First, I will use expectancy-value model of achievement motivation (Wigfield & Eccles, 2000) to guide my understanding of student achievement motivation. Expectancy-value theories of achievement motivation posit that a person‟s choice, persistence and performance can be explained by their expectancy for success on a task and how much they value the task, either because they have interest in it or perceive success at the task may be useful or important at the time of performance or at some time in the

future (Bong, 2004; Cole, Bergin & Whittaker, 2008; Hulleman, Durik, Schweigert, & Harackiewicz, 2008; Wigfield & Eccles, 2000).

Test stakes may affect students‟ achievement motivation because the more a test or assignment counts toward grades, promotion, graduation, or acceptance to a future program of study, the more value and importance the student is likely to assign to that task. At the most basic level of interpretation, increases in the perceived value of a task results in higher levels of motivation; however, ability beliefs and expectancy for success mediate the relationship between task value and motivation, and these relationships are further mediated by situation and context. The teacher comments analyzed in this study add to our understanding of these relationships, and add to the existing literature on this topic in the context of counting LSAs toward class grades.

Consistent with the literature on achievement motivation, the initial themes identified include task value, motivation, student characteristics and test characteristics. As coding of teacher comments progressed, I identified clusters of codes that conveyed the meaning of each theme. Based on expectancy-value theory, I posited that the student characteristics identified (e.g., ability and anxiety) would be predictors of task value. Using the same approach, I posited that test characteristics (e.g., question format and weighting of test toward class grades) would also predict task value because question format can determine the difficulty of a question and the weighting of the assessment can increase or decrease test-stakes for some students. Next, I hypothesized that task value, as identified by words like care, concern, and apathy, would predict motivation, and that could be measured by observing effort behaviours as described by teachers. Motivation is a variable that cannot be directly measured; however, motivation can be inferred from behaviours that demonstrate engagement and persistence in a task because students perceive some intrinsic or extrinsic value in the activity (Wigfield & Eccles, 2000). Figure 1

displays a conceptual model showing the code clusters that were determined to convey meaning for each theme and the proposed relationships between the themes student characteristics, test characteristics, task value, and student effort behaviours as a measure of motivation.

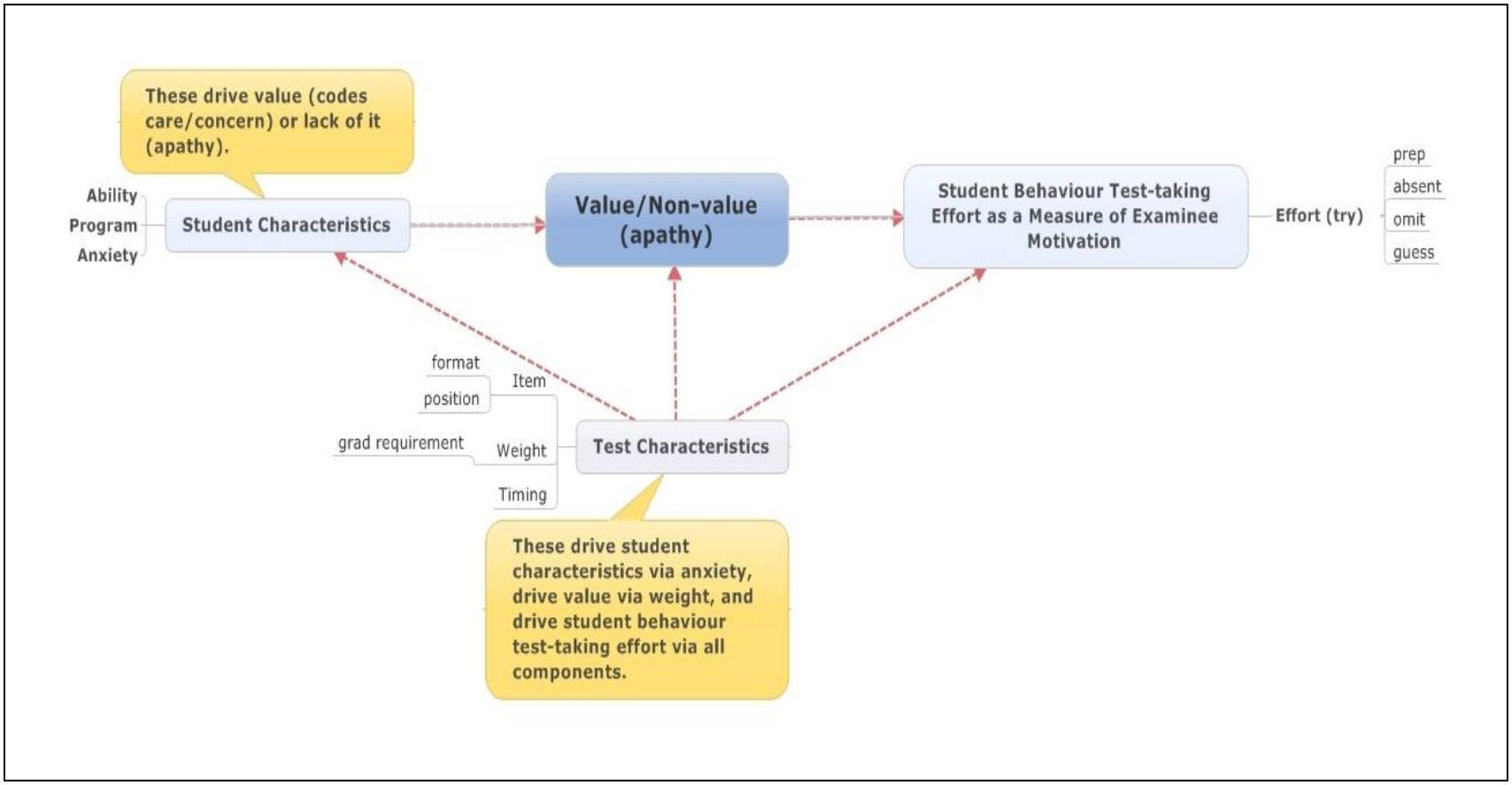


Figure 1. Conceptual model of test-taking motivation: Themes and codes.

The second theoretical framework that guides this research is Messick‟s (1989) unitary view of validity. Messick‟s work is seminal to our current view and understanding of the process of validating test content and design, justifying the proposed use(s) of a test, and identifying the consequences resulting from the proposed uses. Over the past 20 years, Messick‟s contribution to validity theory has had a major influence on educational measurement. Messick‟s theory introduces the responsibility to ensure that the validation of educational assessment includes, not only the effectiveness of the assessment to produce scores or grades that accurately reflect students‟ learning of the subject domain (content-related validity evidence), but also evidence to support the intended use(s) of scores or grades. This includes investigating potential consequences of all proposed uses for the various stakeholders involved.

More recent research on validating assessments for multiple-uses or purposes is indicating that the challenges and complexities of validating an assessment for each of its potential uses, intended and additional, may be much larger and more difficult than first realized. Proper validation of LSAs requires providing validity evidence for the assessment under the conditions of each proposed use. The validity evidence must be examined for each use in isolation and also for interactions when proposed uses are considered in combination. As Koch (2009) explains, “from a measurement perspective, multiple-use is problematic because intended and additional uses cannot be assumed to function in isolation” (p. 4). This requires the intended use(s) of the data be established prior to or during the test development process, and validity evidence for extended uses, post development, should be required.

Building on Messick‟s original theory, this growing research on the process of validating multiple uses suggests that it is not sufficient to provide validity evidence for each proposed use in isolation, which alone is a demanding task, but it is also necessary to study and understand the potential for multiple uses to interact with one another (Koch, 2009; Murphy, 2009; Wolming and Wikstrom, 2010). This study will focus specifically on the additional use of the EQAO Grade 9 Assessment of Mathematics as part of students‟ final grades and examine how that particular use impacts on student motivation and performance as a function of varying test stakes. This study provides an opportunity to examine potential interactions that can arise from multiple uses of LSAs and the implications for validity arguments for their use in calculating students‟ final grades with respect to potential uncontrolled sources of construct-irrelevant variance that may be introduced as an outcome of this practice.

1.3 Statement of the Problem

Student motivation has been identified as an important potential source of CIV in low-stakes test conditions (Haladyna & Downing, 2004; Wise & Bhola, 2006; Wise & DeMars, 2005). Because there may be little incentive to invest time and effort into preparation for and writing of low-stakes tests from the perspective of the student, it is important to understand the potential for underperformance by unmotivated students writing low-stakes LSAs. One of the strategies used to increase student motivation to write an LSA is to offer incentives, for example, counting the LSA toward students‟ class marks. This strategy may increase the stakes and thereby the value and importance of the assessment for some students, but not for all. The effectiveness of the incentive varies from student to student.

Student performance on an assessment may depend on a number of factors including test difficulty, student ability, student preparation, classroom instruction and classroom assessment strategies and their alignment with test content and format, and other contextual influences (Klinger, Rogers, Anderson, Poth, & Calman, 2006). For some students, there may be little motivation to take the assessment seriously even when it is counted toward class grade because they have limited expectations for success on the assessment. If class grades or other types of incentives are used with the expectation of increasing student motivation on LSAs, it is important to know which strategies contribute to improvement in student motivation, in what contexts such strategies can be successful and where they fail. The EQAO Grade 9 Assessment of Mathematics is a particularly useful assessment to study the impact of counting a large-scale assessment towards students' final grades because of the unique way in which this practice is implemented in the province of Ontario. Of the assessments administered by the EQAO in Ontario, the Grade 9 Assessment of Mathematics is the only assessment that the EQAO allows teachers to use in the calculation of students' final grades. Decisions regarding what parts of the LSA to mark and how much weight to give toward students' final grades are made at the individual teacher, school, or board level.

The conditions surrounding the administration of the EQAO Grade 9 Assessment of Mathematics, specifically the decisions that are being made with regard to counting and weighting the assessment, have the potential to impact significantly on student motivation to take the assessment more seriously (Cole, Bergin, & Whittaker, 2008; Sungur, 2007; Wise, Wise, & Bhola, 2006). Consequently, the conditions surrounding the administration and use of this assessment should be of interest for two reasons: the

potential for underestimation of student ability due to low student motivation and the threats to standardization and validity arguments for proposed intended uses of the test that may be introduced when counting and weighting is inconsistent from school to school and board to board (Koch, 2009; Office of the Auditor General of Ontario, 2009).

1.4 Purpose of the Study

The purpose of this study is to describe teacher opinion of counting all or part of the assessment toward class as a means to motivating students to take an LSA more seriously. Attention to the perspectives of teachers is an important focus of educational research (Pajares, 1992). Their beliefs are a valuable source of information that is necessary to describing students and student learning, and in this particular area of research, little is known of teacher opinion of the effectiveness of counting LSAs as a strategy to motivate students to put forth effort.

A triangulation mixed methods design was used to measure the same phenomenon with different yet complementary data (Creswell & Plano Clark, 2007; Greene, Winters, & Forester, 2004). The rationale for choosing this design was the source of some debate, and a basic interpretive qualitative study was considered (Merriam, 2002); however, the nature of the two questions posed, one being a predetermined closed-ended question and the other an open-ended question providing emerging data that allowed for the development of themes, led to the decision to label the design as a concurrent triangulation strategy as described by Creswell (2003). There are some concerns expressed in the literature on mixed-methods research about creating artificial binary divisions and whether or not paradigms should be mixed (Creswell, 2011), but the application of a methodological triangulation (Denzin, 1978) involving the use of more

than one method to gather data and using Creswell‟s more evidence the better argument for mixing seemed to provide the best design fit.

The coding terminology for the qualitative analysis was formulated based on achievement motivation theory. Content analysis of teacher opinion on the effectiveness of motivating students by counting the assessment toward class grades was completed for the constructed-responses provided by teachers who administered the 2010 assessment. It was hypothesized that there would be convergence between the quantitative data collected from the selected-response question and qualitative themes that emerged from the qualitative interpretation of the constructed-responses based on the tenets of achievement motivation theory.

This research used secondary data from the Teacher Questionnaire (TQ) that was collected as part of the 2009-2010 EQAO Grade 9 Assessment of Mathematics. Specifically, I analyzed teacher responses to a two-part question on the TQ. The first part of the question asked “In your opinion, does counting some or all components of the Grade 9 Assessment of Mathematics as part of class marks motivate students to take the assessment more seriously?” Teachers could select from three options, “Yes”, “No”, or “Undecided.” The second part of the question asked “Please comment” and there were four lines where teachers could construct their comments.

Responses to the constructed-response part of TQ22 were analysed using qualitative approaches. The first part of the question provided data on how many teachers agree, disagree or are undecided about this question, but it is the analysis of their constructed responses that contributed to understanding of the reasons why or why not

counting the assessment toward class grades may motivate students, under what circumstances and for which students.

1.5 The Research Questions

This study was designed to answer four specific research questions,

1. What percentage of teachers who administer the EQAO Grade 9 Assessment of Mathematics, 2010 administration, responded affirmatively that counting the assessment toward class grades motivates students to take the assessment more seriously? What percentage responded negatively and indicated that counting the assessment does not motivate students to take the assessment more seriously, and what percentage of the total number of teachers surveyed were undecided?
2. How did teachers' opinions vary by their response to the closed-ended question (i.e., yes, no or undecided), by major themes in the literature such as student characteristics (e.g., ability), student behaviours (e.g., guessing or absent), test characteristics (e.g., weight and timing), test item characteristics (e.g., format and language), and perceived value of the assessment for the student?
3. How do the opinions expressed by teachers on counting the assessment toward class grades relate to or inform expectancy-value theory of achievement motivation?
4. How do the opinions expressed by teachers on counting the assessment toward class grades relate to or inform validity theory?

1.6 The Importance of the Study

Answers to the research questions are important to describe how varying the test stakes by counting the assessment toward class grades may impact on student motivation and student performance on the assessment. Varying test stakes is a concern with regard to the degree of standardization of this assessment, and student motivation and subsequent effort on the assessment needs to be better understood in light of the role it may play as a source of construct-irrelevant variance (CIV) that may threaten validity arguments for all intended and any additional uses of this assessment. The findings of this research also contribute to the literature by adding to our understanding of student effort and performance as recommended by Sungur (2007) who advocated the need for integration of effort into the research design.

Understanding how the test is used and the potential consequences of its use is required for validity as described by Messick‟s (1989) unified theory of validity. Teacher opinion of the use of the test toward calculation of class grades bears directly on their opinion of the validity arguments, both evidential and consequential, for intended uses of the assessment. Teachers' opinions of the effectiveness of counting the test as a way to motivate students and any positive or negative effects noted by teachers will provide greater understanding of the potential consequences for students, teachers, and school planning practices.

The subtleties found in the complex interactions that exist between student ability, task value, effort, motivation and test stakes remains far from being completely understood. Advocates of high-stakes testing believe that students require meaningful incentives such as requirements for promotion or graduation before they will be

motivated to put forth maximum effort. Those against high-stakes testing argue that these tests put low achieving students at greater risk of failure and lead to further disengagement from learning for groups of students that may already be marginalized because of ability or social contextual factors (Kajander, Zuke, & Walton, 2008; Roderick & Engel, 2001).

There will always be students for whom extrinsic motivation is not prerequisite to effort, but those are not the students that need concern us in standardized testing scenarios because those students often work to the best of their ability with less need for external motivators such as high test stakes. What we need to understand is what motivates the students who could go either way depending on the stakes, and what, if anything can motivate the apathetic student to take a LSA seriously. If the results of research indicates that the practice of counting the test toward students‟ final grades does influence student motivation and performance on the assessment by increasing the perceived value and importance of the assessment for students, additional research is needed to determine if counting the assessment toward class grades is the most effective strategy for improving student motivation. It is possible that alternative methods that are less likely to introduce additional threats to validity for the intended uses of the assessment can be identified and used to improve student motivation, effort and performance so that test scores accurately measure students‟ ability.

1.7 The Scope of the Study

This study uses data obtained from the province wide administration of the EQAO Grade 9 Assessment of Mathematics. Teachers responsible for the administration of this assessment are required to complete the Teacher Questionnaire. The focus of this study is

teacher response to question 22 on the Teacher Questionnaire of the EQAO Grade 9 Assessment of Mathematics 2009-2010.

1.8 Limitations of the Study

This study seeks to explore elements that contribute to student motivation to take the assessment seriously from the perspective of classroom teachers. The limitations of this study are:

1. Student motivation can be significantly influenced by factors outside of the school context. The interaction between student ability, student motivation, task value, and test stakes is complex, as are the learning environments and students‟ individual backgrounds and circumstances. Lack of information with regard to students and school for each teacher comment analyzed limits the overall generalizability of the results.
2. The limitations of using secondary data analysis may include missing data due to participant non-response, incomplete or incorrect documentation, and the inability to member check as a method of ensuring internal validity (Merriam, 2002; Rogers, Anderson, Klinger & Dawber, 2006).
3. The qualitative component of mixed-methods research is exploratory and descriptive, and the researcher makes no claims with regard to cause and effect.
4. Students enrolled in locally developed courses or math recovery courses do not write the EQAO, so teacher opinion with regard to student motivation for these students is not addressed by the data used in this research.

1.9 Definitions of Key Terms

Achievement Motivation - refers to an individual's drive or desire to succeed or accomplish a task.

Construct-Irrelevant Variance (CIV) is the introduction of extraneous, uncontrolled variables that result in systematic (non-random), reproduceable variance in test scores as a result of measurement of variables not intended to be measured by the test. The meaningfulness and accuracy of the scores may be adversely affected and the validity is reduced.

Constructed-Response Question - a question or stimulus that requires the response to be constructed by the respondent, e.g., completing a sentence or writing short answer response or. In contrast to multiple-choice questions, constructed-response items require the test taker to construct an answer rather than just recognize one.

Large-Scale Assessment (LSA) – standardized tests in which large numbers of students are assessed for achievement in specified domain(s) of learning. Results are usually used to compare groups of students in districts, regions, and nationally, often for the purposes of public accountability.

Selected-Response Question – a form of questioning that requires the respondent choose a response from a list of options. The multiple-choice question is the most common instance of this strategy.

Triangulation – a method of data analysis that combines independent but complimentary research methods (i.e., quantitative and qualitative methods) to test the consistency of findings obtained through different instruments. This is a common strategy of analysis in mixed methods designs when quantitative and qualitative methods are

combined to provide a more complete set of findings than could be arrived at through the use of either method alone.

Validity – is broadly defined as the extent to which a test measures what it was designed to measure.

Variance – in statistical analysis, the variance is used as a measure of how far a set of numbers are spread out from each other. This spread can provide useful information about the variables of interest providing all other extraneous variables that might interact with the variables of interest are held constant. It is one of several descriptors of a probability distribution that describes how far the numbers lie from the mean (expected value).

Chapter 2

Review of the Literature

2.1 Validation of Standardized Assessments

Gathering evidence for content validity on tests that measure academic achievement relies on arguments that support the relevance and representativeness of test tasks (Kane, 2006). Gathering construct-related evidence and evidence on consequences required to support proposed multiple uses of an assessment can be a challenging task when evidence is required for each component of validity as delineated by Messick‟s theory (Kane, 2008). Koch (2009) provides an excellent discussion of the problems inherent in validating multiple-uses of an assessment because "intended and additional uses cannot be assumed to function in isolation" (p. 4). Kane (2006) expressed these concerns in his comments on validity fallacies:

> The begging-the-question fallacy fits with the confirmationist tendency in many validation studies. For example, the proponents of authentic assessment have tended to emphasize the extent to which the performance observed in testing match the target performance while taking the generalization of observed scores over tasks, occasions, and conditions of observation for granted, even though empirical research consistently indicates that generalizability over performance tasks cannot be taken for granted. Similarly, developers of objective tests have tended to give a lot of attention to content representativeness and generalizability over items, while taking extrapolation to the target performance for granted. (p. 57)

Other researchers, critical of Messick‟s unified theory of validity, have attempted to narrow the objectives of construct validity by proposing that the terminology that has developed over the past 50 years be revised so that content validity is the primary source of evidence for validity and other factors that concern relationships to other measures be addressed under a new category called utility of test (Lissitz & Samuelsen, 2007). This would make the task of gathering validity evidence for an assessment much easier, but then there is the concern that many important aspects of validity, as it is presently understood, would go largely ignored. Kane (2008) also wrote about these concerns:

> Although I have some reservations about Messick‟s (1989) formulation of validity theory, I do not think that Lissitz and Samuelsen‟s proposals would improve the situation if they were implemented. They would either reduce validity to a very narrow concern about the representativeness of the test‟s content (along with some attention to reliability and scaling models) or, more likely, take us back to the situation in the 1970s and 1980s, when we had a profusion of specialized validation methods (shortcuts), each designed for a specific kind of application. (p. 77)

Kane questions what advantage there would be to changing current terminology, and he points out that “assuming that the issues of utility and consequences that would be removed to the external aspect are not going to be ignored, the need to evaluate relationships with external variables and to evaluate consequence will remain on the table” (p. 78). He suggests that, instead of adopting the changes suggested by Lissitz and Samuelsen (2007), we build on the work of validity theories that have come out of the past 50 to 100 years of research. He also makes it clear that, in his belief and that of

several other prominent measurement researchers (Gorin, 2007; Mislevy, 2007), a broader and more inclusive definition of validity than that proposed by Lissitz and Samuelsen is required.

Ungerleider (2006) echoes these concerns and supports Messick‟s (1998) position that stressed validation of all proposed uses of an assessment. “The validity of any assessment result is conditional on the fit between the purpose for which the assessment was designed and the use of the results” (p. 875). For this reason, it is important to ensure that evidence for validity has been established for all of the proposed uses, intended and additional, of large-scale assessments.

The intended purpose of the EQAO and the intended uses of the data collected from the assessments it administers are stated in its 2009 – 2010 Agency report (EQAO. 2010a):

> EQAO acts as a catalyst for increasing the success of Ontario students by measuring their achievement in reading, writing and mathematics in relation to Ontario Curriculum expectations. The resulting data provide a gauge of quality and accountability in Ontario‟s publicly funded education system. The objective and reliable assessment results are evidence that adds to the current knowledge about student learning and serves as an important tool for improvement at the individual, school, school board and provincial levels. EQAO helps build capacity for the appropriate use of data by providing resources that educators, parents, policy-makers and others in the education community can use to improve learning and teaching. (p. 24)

To be used effectively for these purposes, the assessment instruments must provide scores that accurately reflect student achievement and also meet validity arguments for multiple intended uses.

The goal of providing data that provides a gauge of accountability is broad, so it is important to define the criteria for evaluating the standards set for accountability. What information will be used to measure accountability? How does that information inform accountability, for whom, and for what purpose? Different decision makers will put more or less emphasis on specific criteria, and these differences in emphasis reflect potentially important differences in the criteria that each stakeholder values; however, to validate the use(s) of any test, it is important to assess these values. Murphy (2009) summarized this concept and draws attention to the importance of values and perspective. He stated:

> First, it is impossible to avoid the question of values. The choice of a particular criterion *is* a statement of values. That is, when researchers assume that the purpose of a test is to predict future performance and they choose operational measures of that criterion as a basis for validating tests, they are implicitly stating that other perspectives, preferences, and values are *not* part of the purpose of testing. At a minimum, validity researchers should try and find out how organizations actually use tests, what they hope to accomplish through testing, and how they make decisions about balancing the preferences and values of the various stakeholders in the testing process. (p. 428)

The data collected by the EQAO through assessments is currently used to provide feedback to students, parents, teachers and schools with regard to individual student achievement, for student learning improvement planning, for school improvement

planning, and in the development of intervention programs or "Turnaround Teams" for schools that lack the capacity to address needed improvements (Office of the Auditor General of Ontario, 2009; Ungerleider, 2006).

Several different groups of stakeholders are associated with the multiple intended uses of the assessment results, and each group of stakeholders will have different perspectives and opinions on what is valuable and important. Each stakeholder is subject to consequences as a result of the outcome of the assessments, and it is these consequences, in light of competing values, that present difficulty but must be considered in the validation process. Messick (1989) stated, "It must be recognized that the appropriateness, meaningfulness and usefulness of score-based inferences depend as well on the social consequences of the testing" (p. 19). Building on Messick‟s (1989) work, Crooks, Kane, and Cohen (1996) proposed the eight stage threats to validity model. In their model, the eighth link is defined as the impact of the assessment on students and other participants in the assessment process. Two threats to validity are associated with this link: positive consequences not achieved and serious negative impact outcomes. Positive consequences not achieved are often the result of sources of construct-irrelevant variance (Stobart, 2001).

Construct-irrelevant variance (CIV) in achievement testing has many sources that may originate in the test itself, in the administration of the test, or in the behaviours or characteristics of the students taking the assessment. Each may pose a threat to validity by introducing reliable and reproducible variance into test scores that may change the estimated true score by either underestimating or inflating the score (Haladyna & Downing, 2004). "CIV occurs when a test measures variables that were not intended to

be measured by the test designers. It is systematic error (rather than random error) introduced into the assessment data by variables unrelated to the construct being measured" (Downing & Haladyna, 2004, p. 38). The threat may or may not be serious for the score interpretation and use intended by the end user(s); however, it is important to consider the sources of CIV for any large-scale assessment and to identify those that may pose a serious threat to validity. Once identified and understood, it may be possible to make changes to the test or its administration so that the sources of CIV that pose a serious threat to validity can be reduced, or at least appropriately considered in the analysis of results.

Student motivation is a variable that researchers have identified as a potentially serious source of CIV in direct assessment contexts such as standardized achievement tests (Wise, Wise & Bhola, 2006). Motivation is a student-specific source of CIV, and compared with other sources, individual specific sources of CIV potentially contribute the most serious threat to validity (Haladyna & Downing, 2004). Within the body of research on the effects of student motivation on test performance, student motivation is correlated with student effort, and student effort is correlated with test consequences or stakes and test performance (Cole, Bergin, & Whittaker, 2008).

Research shows that the "reasons that students have for taking a test affect test performance, but nuances of how the reasons impact on test scores are only beginning to be investigated" (Cole, Bergin & Whittaker, 2008, p. 610). Student motivation to try hard on a test is important but not well understood. Motivation theories provide a place to start understanding what factors influence student motivation to put forth effort in testing situations. Specifically, expectancy-value theory of achievement motivation (Wigfield

and Eccles, 2000) provides a framework for understanding the relationship between test-taking motivation and test performance.

2.2 Achievement Motivation Theories

Research in achievement motivation provides a number of theoretical perspectives that can help us understand how motivation may influence student behaviour and performance. The one that is probably most germane to this research is the expectancy-value theory of achievement motivation because this theory describes how motivation and persistence on a task are determined by the value individuals allocate to the task, their ability beliefs, and their perceived likelihood of success. Wigfield and Eccles (2000) provide an expectancy-value model that describes why students may not be motivated to do their best work. Students‟ expectancy and value constructs directly influence what they choose to work on (achievement choices), their effort and persistence, and performance outcomes. Expectancies and values are determined by ability beliefs (can I be successful at this task?), perceived difficulty of the task, individual long term and short-term goals, self-schema and effective memories.

Ability beliefs are the perceptions people have about their competence to complete a task, which shape expectancies for success. If a student does not believe he/she has the ability to be successful at a task such as a test in mathematics, that student is unlikely to be motivated to put forth effort in the test situation (Roderick & Engel, 2001). Over years of schooling, students come to expect successes or failures in specific subject domains (Kajander, Zuke, & Walton, 2008; Wigfield and Eccles, 2000). Their positive and negative experiences shape future ability beliefs in any given domain and form the student‟s perception of competence in different activities and subject domains

over time. These perceptions become part of the individual‟s self-schema, which is more fixed and less amenable to change as students mature. Our self-schema or perception of self shapes our expectancy for success (Eccles & Wigfield, 2002; Pajares, 1996), and repeated experiences of failure contribute to the observed pattern of diminishing expectancy for success in specific domains as students move from the elementary to intermediate and high school grades. This effect has been found to be particularly strong for girls and their choice to pursue mathematics at higher grade levels (Watt, Eccles & Durik, 2006).

Expectancies for success and the value students place on a task or outcome are influenced by ability beliefs, and students will often place less value on activities on which they do not expect to be successful; however, ability beliefs alone do not determine the value students attach to a task or outcome. Students appear to begin rating the value of tasks early in elementary school. Very young children make task-value judgements based on complicated criteria that involve judgements of utility and intrinsic value (Wigfield & Eccles, 2000). They decide what activities to devote effort to based on what they think will be a skill or ability that may be useful in the future (extrinsic utility value) and those activities and tasks they think are fun or enjoyable (intrinsic value). Fortunately, many students will persist at a task when their ability is low if they can appreciate the need for or place importance in acquiring some level of proficiency at the task (Ryan & Deci, 2000). Understanding students‟ perceptions of task-specific abilities, expectancies for success, and subjective task value is critical to further analysis of motivation and the relationship of motivation on performance.

Students‟ commitment to academic achievement is not static, and there are many factors that influence and are constantly subject to change at any time (Levin, 1993); however, based on their expectancy-value model of achievement motivation, Wigfield and Eccles (2000) measured the effects of ability and task value on student motivation and test performance and found that students‟ ability beliefs were consistent, effective predictors of non-effort behaviours and low test scores. This is an important contribution to our understanding, yet there remains an intricate interplay of ability beliefs, expectancies for success, subjective task values, intrinsic and extrinsic motivation and interest that shapes students‟ performance and choice, and these factors are continuously subject to change as students‟ life experiences and circumstances change. It seems a daunting task to attempt to piece together how each variable influences student behaviour and performance; however, it is necessary that we continue to increase our understanding in order to improve assessment and best teaching practices.

In the development of the expectancy-value theory of achievement motivation, Wigfield & Eccles (2000) bring to the forefront the factors that contribute to student motivation, and by the understanding provided by motivational models of this nature, we are better able to devise methods for addressing some of the factors that may be influencing student performance on LSAs and suggest ways to control for the effects of student motivation as a source of CIV that threatens the valid use of test scores.

2.3 Student Motivation, Test Stakes and CIV

Students‟ perceptions of test stakes may have a significant and immediate effect on motivation, effort and persistence, so test stakes are an important consideration in the validation process. Test stakes may be high, low or somewhere in between, and

perceived stakes vary by stakeholder. A test may be considered high stakes for schools and boards because, for example, the outcome determines funding allocation; however, the same test may be considered low-stakes for the student because it is not counted toward students‟ final grades, nor is it a requirement for promotion or graduation (Rahn, Stecher, & Goodman, 1997; Cole, Bergin, & Whittaker, 2008; Sungur; 2007)

Based on expectancy-value theory, it is reasonable to predict that student motivation to put forth effort on an assessment will vary depending on perceived value of the assessment for the student, and the higher the test stakes for the student, the more motivated some students will be to put forth effort on the test. To test this hypothesis, Sungar (2007) examined the relationship between students‟ motivational beliefs and meta-cognitive strategies such as planning for and engaging in preparation prior to the test and their monitoring of their own progress while writing the test under consequential and non-consequential conditions.

In this study, students were randomly assigned to two testing conditions: the test counted toward students‟ final grades (i.e., consequential condition), or the test did not count toward their final grades (i.e., non-consequential condition). Sungar found that goal orientations mediated student performance, and students with mastery goals (i.e., students who desire to gain or improve knowledge) performed better under both consequential and non-consequential test conditions. Even when students are mastery goal oriented, there were subtle differences noted in their performance under consequential and non-consequential conditions. Mastery goal oriented students normally employed various test-taking strategies and metacognitive monitoring of their work but tended to do less of this under non-consequential test conditions. “Under non-consequential test conditions,

individual levels of student motivation became the main predictor of engagement with the task" (Sungar, 2007, p. 138). This study is important in that it brings attention to the complex interactions of students‟ perceptions of value and test-taking behaviours, and as Sungur points out, we lack understanding of how the perception of effort may differ among students. Sungar goes on to recommend that future research should include the integration of a measurement of effort in the research design as a variable, and measurement of effort should include qualitative components such as observations and interviews.

Similar findings were reported by Klinger and Luce-Kapler (2007) who studied student performance on a high-stakes LSA. They sampled Ontario high school students preparing to write the Ontario Secondary School Literacy Test (OSSLT) from two regions in Ontario. The OSSLT is a high-stakes literacy test and a graduation requirement in Ontario. They selected an equal number of male and female students from each region and classified them as likely or unlikely to be successful on the OSSLT based on information provided by the schools. Their objective was to gain a better understanding of the relationship of students‟ perceptions of the OSSLT, their views on the importance and value of the test, and their performance on the test compared to their performance histories in the subject domain. Klinger and Luce-Kapler (2007) asked the following research questions: "What specific preparation programs did the students complete in preparation for the OSSLT? What was the impact of the OSSLT with respect to school, future education, and career directions" (p. 32)?

The students did not place importance on the test in terms of having impact or implications for their learning in other subject areas. Interestingly, students had mixed

views on the value of the test, even though passing it was a requirement for graduation, and in 2007, almost one third of students enrolled in applied English programs failed the OSSLT. Klinger and Luce-Kapler (2007) were able to identify differences between the successful and unsuccessful students that were separate from task value. Based on interview comments from the students, most saw little value in the test with regard to their learning, so there were other factors mediating the behaviours of students likely to succeed and those categorized as less likely to succeed on the test.

The students whose prior history placed them in the likely to succeed category, while not necessarily valuing the test itself, seemed to appreciate the need to jump through the hoop in order to complete their high school studies and self-prepared for the test. They were more inclined to use a variety of test taking strategies (e.g., consciously choosing which questions to attempt first/last) that improved their chances of success. The successful students had a better understanding of the test, including the structure and expectations of the test, so although their comments indicated that they did not perceive any value in the test itself with regard to their learning, they did respond to the stakes and put forth effort to prepare for the assessment to meet graduation requirements.

In total, Klinger and Luce-Kapler (2007) obtained data from 42 students, 22 classified as successful or likely to be successful and 20 classified as unsuccessful or unlikely to be successful. They state, “Most surprisingly, perhaps, those students at the greatest risk for failure were also the least likely to report they engaged in self-preparatory literacy activities. Only three of the unlikely to be successful students we interviewed noted independent preparation for the test…” (p. 40). They reported students in the unlikely to succeed category to be in a form of denial, expressing expectancies for

success that were not congruent with their past performance in the subject domain. These students expressed the belief that they would pass, even though they had not put forth any effort to self-prepare, and during test taking, they worked from front to back in the booklet and complained of not having enough time to complete the test. Like their successful counterparts, they did not perceive any intrinsic value in the test, and their persistent belief in success suggests they were aware of the stakes; however, unlike their successful peers, these students appeared unable or unwilling to apply the necessary strategies to succeed.

Roderick and Engel (2001) reported that approximately one third of students in their study showed little work effort, even when promotion was dependent upon passing the test. These students had "significantly larger skill gaps and barriers to learning" (p. 219) both within and outside the school. Unlike the students who demonstrated good ability in the subject being tested, the low-achieving students were less likely to be motivated to take tests seriously in response to increasing test stakes, less likely to put effort into preparation prior to the test, or persist in their effort during the test. Once again, it is useful to refer back to Sungar (2007) and make note of the fact that there is a lack of research in the literature that would enable us to better understand whether low-achieving students knowingly put forth insufficient effort or if they believe they put forth sufficient effort.

Research on goal theories (Ames, 1992; Dweck, 1999; Pintrich 2000) and attribution theory (Pintrich & Schunk, 2002; Weiner, 1974) has contributed to our understanding of the factors that determine how students derive value in a task and how that leads to individual differences in behaviours, such as engagement, self-preparation,

and perseverance. Goal theory posits that students can have either mastery goals or performance goals. Mastery goal oriented students want to learn and increase proficiency. They tend to attempt more difficult tasks and persist at difficult problems. They attribute their successes to stable, internal factors (e.g., ability) and their failures to external factors (e.g., preparation) over which they have control (Pintrich, 2000).

Students exhibiting performance goal orientation work to attain positive judgements and/or avoid negative judgements. Performance goal orientations are associated with maladaptive behaviours because the "performance goal focuses the individual on judgements of ability" (Dweck & Leggett, 1988, p. 262). The mastery goal oriented student sees self-preparation and effort as consistent with task requirements and is able to focus his or her attention fully on the task that serves the goal. In contrast, the performance goal oriented student, when faced with a challenging task, may lose confidence in his/her ability because attention is divided between anxiety over outcome and strategy formulation and execution of the task. This results in a loss of belief in the efficacy of effort. If one is not able, effort will do nothing more than confirm lack of ability, and performance avoidance behaviours ensue. Interestingly, Dweck and Leggett (1988) found that mastery and goal oriented students were often equal in ability; however, the goal orientations appeared to lead to quite different behaviour patterns, and only those students having mastery goal orientation demonstrated the behaviours that ultimately led to success on challenging tasks.

Similarly, Skaalvik (1997) studied goal orientation for mathematics and found work avoidance strongly and negatively correlated with intrinsic motivation. Students with performance goal orientations are extrinsically motivated because their motivation

to perform is derived from outside sources of approval. As Dweck & Leggett (1988) found in their study, students who are performance goal oriented do not cope well with failure when compared with their mastery goal oriented peers who are more intrinsically motivated. The work that has come out of the studies on goal orientations and attribution helps to explain why some students may not put forth effort or be able to perform to the best of their ability on a task, regardless of their ability in the subject domain or their recognition of value in the task. If the performance goal oriented student perceives the task as too difficult and something they are not likely to succeed at, they may begin to exhibit an increasing number of performance-avoidance behaviours that often result in unsuccessful completion of the task (Eccles and Wigfield, 2002).

The work of Dweck and Leggett (1988) supports Wigfield and Eccles (2000) in the assertion that students‟ negative ability beliefs and decreasing expectancies for success compound over time and repeated failures in specific subject domains. Bong, (2004) reported, “the magnitude of the observed correlation among the subject-specific task-value beliefs was somewhat stronger than anticipated” (p. 295), and in the same study Bong found that, although students may express similar confidence in their abilities across domains, they do not necessarily attribute success or failure to ability in the same way across domains. Attributing success to ability appears to be more stable in mastery goal oriented students and develops over time and with repeated experiences of success in specific subject domains.

It is also possible that some students switch back and forth between mastery and performance goal orientations. Mattern (2005) studied 143 college students to determine how some students benefit from multiple goal orientations. In her study, she addressed

recent literature on performance-approach goals that assert the performance-approach orientation may not always be associated with low marks and maladaptive behaviours. There may be conditions where it is beneficial for students to choose between mastery-goal orientation or performance-goal orientation, depending on the task and the subject. Mastery goal orientation is a predictor of interest and importance of the subject for the student (Pintrich, 2000); however, performance goal oriented students may also perceive value in a task regardless of whether or not they have interest in the task or subject domain. These individuals seek success to maintain a positive image of their ability, so compared to their mastery goal oriented peers, these students might be at an advantage when having to complete a task in which there is little interest but some utility. Mattern (2005) found that some successful students had performance goal orientations, at least for some tasks, but mastery goal orientation resulted in consistently higher performance.

Attaching marks and grades to assessments may be considered a form extrinsic motivation that Ryan and Deci (2000) would categorize as external, less autonomous, and less effective over time. External forms of extrinsic motivation are characterized by responses to reward or punishment or ego involvement (i.e., seeking self-approval or the approval of others), and in ways similar to performance goal orientation, externally regulated students showed less interest and more maladaptive behaviours such as blaming other people for their failures. The judgments of value and importance that students make about a task and their impact on motivation and performance are salient, but it would seem that all value judgments are not equal. Task value must be present to drive motivation (Wigfield & Eccles, 2000), and the factors that lead to the assignment of value are equally important. Externally imposed rewards or punishments may produce a

differential effect on student motivation and test performance, but it is not a positive one for all students (O‟Neil, Sugrue, & Baker, 1995, 1996; Wise & DeMars, 2005; Wolf & Smith, 1995).

Grades and test stakes are only a part of what goes into determining task-value, student motivation and performance. Regardless of test stakes, the behaviours of some students suggest academic apathy (e.g., not studying, guessing, absence); however, these behaviours may be more indicators of avoidance behaviours that are the result of negative expectancies for success, extrinsically grounded performance goal orientation, or negative attribution patterns that result in feelings of helplessness and maladaptive behaviours, even toward tasks to which the student attaches some value (Locke & Latham, 2004). For these students, raising test stakes will not improve motivation, effort or performance. In fact, introducing more externally imposed contingencies for success are likely to create the opposite effect in these students by further alienating them from the task. The outcome for these students is greater apathy and increased performance-avoidance behaviours (Dweck & Leggett, 1988; Pintrich, 2000) Understanding and identifying the cause of apathy in some students may be as important to the understanding of test stakes, task value and student performance as is the understanding of factors that motivate students to put forth effort on assessments.

There continues to be much debate and mixed results on the effectiveness of test stakes as a strategy to motivate students (Eklof, 2007; Wise & DeMars, 2010; Abdelfattah, 2010). The problem that remains is the strategy works for some students, but not for all, and it is important to determine if it is a poor strategy more often than it is an

effective one. Kiplinger & Linn (1996) state in their review of the NAEP motivation studies:

> The NAEP Motivation Studies described in this special section present a somewhat mixed picture of the interrelationships among motivational factors, test stakes, and student performance. Responses to the NAEP motivation questions in 1991 and 1992 clearly demonstrate that many examinees are not motivated to perform well and do not try very hard to excel on the NAEP tests. (p. 129)

Expectancy-value theory describes a relationship between task value, perceived importance, motivation, and performance, but understanding how these relationships interact and shift under different circumstances and with individual students is complex. Specifically, with regard to student performance on LSAs, scores are influenced by outside factors such as student demographics, family income, and peer influences (Greene, Winters, & Forster, 2004). Many factors appear to determine how students make task value judgements and the degree to which they choose to engage and persist in a task; test stakes is only one (Schweigert & Harackiewicz, 2008).

2.4 Implications for the EQAO Grade 9 Assessment of Mathematics

Almost all (95%) school improvement plans in Ontario referred to the provincial assessments as a measure of student achievement (van Barneveld, Stienstra, & Stewart, 2006). The Ministry of Education in Ontario may develop policy using EQAO data or identify specific schools and select them for participation in programs; for example, the Lighthouse Program in Ontario is designed to assist with improvement plans (Literacy and Numeracy Secretariat, 2006). “EQAO assessments are directly linked to curriculum expectations, reflect the work being done every day I classrooms across Ontario and

produce information that is useful for improvement planning at the individual, school, school-board and provincial levels" (EQAO, 20101, p.20).

In addition to drawing on EQAO data for policy making, individual results are reported to the student, and school and board results are reported to the public. Two validity issues immediately arise when one considers the proposed multiple uses of the test scores. First, it must be established that the test scores accurately represent student achievement and ability in specific domains of learning identified and that the arguments for validity support the proposed multiple uses of the assessment.

Low student motivation to take the test seriously is a concern because it may be a source of CIV that results in the underestimation of student achievement or positive consequences not achieved (Stobart, 2001). Currently, schools and boards in Ontario take different approaches to motivating students to take the provincial assessments seriously and some may not make any attempt at all to motivate students. As such, not all students are subject to consistent motivational strategies or incentives, so differences in student scores among schools or boards might not be indicative of real differences in student ability or performance, but instead the score differences may be measuring the effects of differential manipulation of test stakes or other motivational techniques.

Identifying student motivation as a potential threat to validity on the EQAO Grade 9 Assessment of Mathematics is an important first step; however, attempts to correct this problem by altering test stakes may not be the most effective strategy, especially if the stakes are varied inconsistently from school to school. Implementing different motivational strategies or varying test stakes inconsistently from school to school may undermine arguments for validity by creating different testing conditions for students, as

well as corrupting the standardization of the test. These decisions must be given careful consideration when determining assessment methods, and the juggling that must be done is well described by Rahn, Stecher, & Goodman (1997). They state:

> The quality of an assessment depends upon its reliability, validity and fairness. The feasibility of a test is judged by the factors of cost and time for administration, the complexity of the assessment, and the acceptability of the assessment. Quality and feasibility are both important but are sometimes at odds with each other, and schools can be forced to make trade-offs between them when making assessment decisions. Trade-offs occur between single and multiple purposes, high and low stakes, embedded and standalone tasks, voluntary and mandatory participation, and standardization and flexibility. (p. 85)

The differential manipulation of test stakes that currently exists in the administration of the EQAO Grade 9 Assessment of Mathematics leads us to ask important questions. Is it necessary to manipulate student motivation to take the test seriously by counting the test toward class grades and thereby increasing test stakes? For those students identified as low-achieving and/or apathetic toward academic achievement, altering the test stakes may do little or nothing to improve their level of motivation and may even have a negative effect on test performance for these students by increasing other sources of CIV such as test anxiety.

Even if it is found that counting the test toward students‟ grades is an effective motivator that results in some students taking the test more seriously, this technique should be uniformly applied to ensure that differential motivation is not a source of CIV. The importance of identifying and eliminating or reducing sources of CIV in any large

scale standardized assessment is critical to the process of appropriately gathering validity evidence that supports the intended uses of the assessment (Downing & Haladyna, 2004).

Chapter 3

Research Methods

3.1 Research Design

Mixed methods studies have been defined as studies that combine qualitative and quantitative approaches into the research methodology of a single study (Tashakkori & Teddlie, 1998). Quantitative and qualitative data collection, analysis and the mixing of quantitative and qualitative approaches occur within a single study, with data integrated at some stage (Creswell & Plano Clark, 2007). Mixed-method researchers believe that "the use of quantitative and qualitative approaches in combination provides a better understanding of research problems than either approach alone" (Creswell & Plano Clark, p. 5).

Quantitative data is not sufficient to understand the research questions in this study. While the quantitative analysis of the selected-response question can tell us the percentage of teachers who agree that counting the assessment toward class grade helps motivate students to take the assessment more seriously, how many do not, and how many are undecided, that data alone cannot provide information on teachers‟ opinions or concerns about how, why or why not counting the assessment motivates students. The beliefs and attitudes of teachers toward counting the assessment toward class grades is better captured by the constructed responses they provide.

Qualitative analysis of the comments teachers make can provide description of teachers‟ reasoning and a deeper understanding of the issue as perceived in its real-world context, which is the strength of qualitative methods. When used in combination, the quantitative and qualitative data sets collected in this study complement each other

(Tashakkori & Teddlie, 2003), resulting in a stronger research design and more meaningful analysis of data.

The instrument for data collection is the EQAO Grade 9 Assessment of Mathematics Teacher Questionnaire, and the question of interest, (Q22), collects both qualitative and quantitative data concurrently. A concurrent triangulation design was used to validate the quantitative data by analyzing the extent to which constructed-response themes support the survey results (Cresswell & Plano Clark, 2007). This model is used regularly when it is desirable to compare results or to validate, confirm, or corroborate quantitative results with qualitative findings (Creswell & Plano Clark, 2007, p. 64).

While designing a mixed methods study, the researcher needs to consider the timing of the data collection, the weighting or priority given to qualitative and quantitative data, and how and when the data will be mixed (Creswell, Plano Clark, Guttman, & Hanson, 2003). Priority refers to which method, either quantitative or qualitative, is given more, less or equal emphasis in the study. In a study using a triangulation design and convergence model, analyses of quantitative and qualitative data are done separately. The data is mixed in the integration phase where neither takes priority. Instead, the researcher converges the quantitative survey results with the qualitative findings with the intent of answering such questions as “how many” and “why” in the same study (Creswell & Plano Clark, 2007; Cooper, Porter, & Endacott, 2010).

Creswell (2003) states that there are three considerations in matching a research design to a problem: the audience, the problem, and the personal experience of the researcher. The primary audience for this study includes researchers in education,

administrators at the EQAO and Ontario‟s Ministry of Education, board and school levels, and practicing educators. The final reporting provides rich description of teacher opinions on student motivation and counting the assessment toward class marks as communicated in teacher comments. Reading, interpreting, coding and identifying emerging themes is an experience that evolves as the researcher becomes immersed in the data, and a large part of the interpretation found in the reporting of results occurs simultaneously with data analysis (Marshall & Rossman, 2006). Finally, practice in the real-world setting needs to be related to theory, and in the case of this research, my interpretation of teacher opinion of real-world practice will be described as it relates to expectancy-value theory of achievement motivation (Wigfield & Eccles, 2000), current research on test stakes and motivation, and relevant research in educational measurement on validating large-scale assessments.

3.2 Data Collection

Secondary data provided by the EQAO was used in this study. The EQAO Grade 9 Assessment of Mathematics is administered two times in the school year: testing of first-semester academic and applied students is conducted in January, and testing of second-semester and full-year students is conducted in June. The assessment is mandatory for all students enrolled in Grade 9 mathematics, academic or applied courses, in Ontario‟s publically funded schools. After the 2006/2007 administration of the assessment, exemptions have not been permitted (EQAO, 2010b), and accommodations and special provisions are made for ESL students and students on IEPs to allow them to participate wherever possible. Exemptions are only permitted for students who are temporarily absent due to illness or other reasons, and some of these students can defer

writing to a later date. The other group of students not participating in the assessment includes students who are not receiving the Grade 9 curriculum, applied or academic. These are students who have significant skill gaps and learning challenges, and they may be working toward a certificate of achievement, rather than a high school graduation diploma. They receive special programming, sometimes referred to as essentials courses or locally developed courses. Because these courses do not cover the curriculum tested by the EQAO Grade 9 Assessment of Mathematics, students in these classes do not write the assessment.

The data for this study was collected by the EQAO during the 2010 administration of the assessment. A major advantage to using the data available from the EQAO is the breadth of the data accessible to the researcher because the questionnaire is potentially completed by every teacher in the province who administered the test. A common disadvantage to using secondary data often cited is that the data is not collected to answer the specific research question (Boslaugh, 2007); however, in the case of this study, a collaborative team worked together to develop relevant questions that could be incorporated into the Student Questionnaire and Teacher Questionnaire of the 2010 assessment. Following a pilot test, three new selected-response questions were added to the Student Questionnaire. Seven new questions were added to the Teacher Questionnaire, including Q22, which is the focus of this study (van Barneveld, King, Simon & Nadon, 2010).

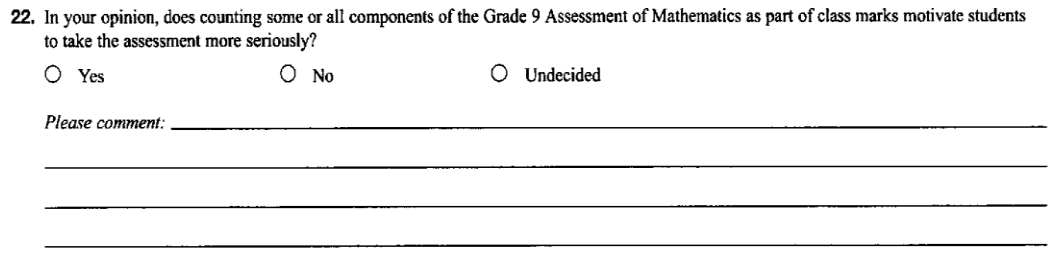
3.3 Participants

As part of its assessment program, the EQAO collects data about students‟ learning environment through the student, teacher and principal questionnaires, which are

completed at the time of the assessment. The teacher questionnaire is designed to collect contextual data, a part of which is information on the use of EQAO data and resources. The teacher questionnaire was completed by teachers in the province of Ontario who participated (n=4853) in the January and June 2010 English language administration of the EQAO Grade 9 Assessment of Mathematics for both applied and academic Grade 9 Mathematics courses (EQAO, 2010b). Completion of the teacher questionnaire is strongly urged by the EQAO and school principals, so a high degree of compliance can be expected. In total, 4,519 non-blank responses were analyzed. The difference (n=334) is believed to be due to a small number (7%) of teacher questionnaires that were not included because they were received by the EQAO after the data had been sent out for this research. The identity of the teachers who completed each form and personal information about them (e.g., age, gender, location) is not provided by the EQAO. The responses are anonymous to the researcher.

3.4 Instrument

The English data was provided by the EQAO in a PDF file that contained images of one page from the 2010 administration of the Grade 9 Assessment of Mathematics Teacher Questionnaire that showed teacher responses to Q22. Teachers were asked to provide a selected response to the question posed and were offered the opportunity to construct a response. Figure 2 shows the question as it appears in the Teacher Questionnaire for the 2010 administration of the EQAO Grade 9 Assessment of Mathematics.


22. In your opinion, does counting some or all components of the Grade 9 Assessment of Mathematics as part of class marks motivate students to take the assessment more seriously?

○ Yes ○ No ○ Undecided

Please comment: _____

Figure 2. Question 22 from the EQAO 2010 Teacher Questionnaire.

It is not possible to systematically distinguish between applied and academic programs from the images in the PDF file; although, some teachers made comments that referred specifically to applied or academic programs or students.

3.5 Data Analysis

The frequency of response for each option (i.e., yes, no, undecided) from the selected-response part of the question provided basic descriptive statistical data that contributed to the quantitative component of data analysis. Both the selected-response part and the constructed-response part of Q22 were coded using ATLAS.ti version 6, which includes full native Adobe PDF support. This qualitative data analysis software was used to arrange, reassemble, and manage the data in a systematic way to explore the relationship between teachers‟ responses (i.e., yes, no, undecided) and preliminary and emergent themes related to student characteristics, student behaviours, test characteristics and motivation as identified in teacher comments.

For the constructed-response part of Q22 (i.e., Please comment), both a deductive and inductive approach was used to identify the themes and main ideas reflected in the teacher responses. First, a preliminary list of themes was deduced from research

literature (e.g., DeMars, 2000; Stocking, Steffen & Eignor, 2001; Wise, 2006, Wolf & Smith, 1995; Wolf, Smith & Birnbaum, 1995).

Themes included:

1. Comments about student behaviors (e.g., student guessing, student self-preparation for the assessment) and student characteristics (e.g., ability) that are related to student motivation to take the assessment seriously.
2. Comments about the test characteristics (e.g., weighting) or test items (e.g., item format) that are related to student motivation to take the test seriously.
3. Comments about the value of the assessment for students that include indication of care, concern and importance placed on writing of the assessment. Conversely, comments that reflect a lack of value for the student include apathy, lack of care, concern or importance placed on writing the assessment by the students.

As coding progressed, an inductive approach was used to identify other themes or clusters of codes that conveyed meaning that emerged from the data (See Appendix 1 for the list of final themes). For each theme, the related codes, definitions, and examples were documented in a coding manual. Two coders, the primary researcher and thesis supervisor, independently coded a random sample of 113 teacher comments. We conducted a formative check of the inter-coder reliability for the constructed-response part of Question 22. The number of codes per teacher comment ranged from 2-5 codes. Agreement was defined as two coders independently applying the identical codes to the teacher comment. For example, if a teacher‟s comment was linked to 3 codes by one coder, then the other coder needed to assign the exact same 3 codes to be counted as

agreement. If there was one mismatch or more, then it was considered to be disagreement. The result of this check of inter-coder agreement was 36% agreement.

Next, we took a sample of 54 non-blank comments and assessed the code-by-code agreement. Code-by-code agreement used the code as the unit of analysis (not the teacher comment) and agreement was defined as two coders independently applying the same code to the same teacher comment. For example, if the two coders assigned the same 2 codes to a teacher‟s comment but mismatched on a third code, then the code-by-code agreement for that teacher‟s comment was 2 out of 3. The result of the code-by-code agreement was 74% agreement. Coders discussed the mismatches, which tended to be a result of minor differences in the operative definition of some codes. Where required, the codes, definitions and examples in the coding manual were clarified. Please see Appendix 1 for the list of codes, definitions and examples.

After coding was completed, relationships were examined between teachers‟ responses to the selected-response part of Q22 (i.e., yes, no, undecided) and the codes grouped by theme. The frequency of codes occurring within each of the teacher response options was considered; however, frequency alone was not used to determine importance. Clusters of codes that form themes that converged with the tenets of expectancy-value theory played an important role in understanding teacher comments in relation to the theoretical framework of the study.

3.6 Ethical Considerations

Because the secondary data being used comes from anonymous sources, the only ethical considerations that must be observed are to follow the guidelines set out by Lakehead University with regard to graduate research protocols and to ensure that access

to the data was restricted and privacy issues observed as consistent with EQAO data access agreements.

Chapter 4

Results

4.1 Descriptive Statistics

The EQAO provided 6070 documents for analysis. Of the 6070 documents, 1536 were blank and two pages were incorrectly scanned. Of the remaining n=4532 documents, not included in the quantitative analysis were responses where multiple options were selected (n=5) and documents that included a comment but lacked a constructed response (n=8). The remaining n=4519 documents were analyzed according to constructed response and teacher comment where completed. Teachers provided a constructed response to the question "In your opinion, does counting some or all components of the Grade 9 Assessment of Mathematics as part of class marks motivate students to take the assessment more seriously?" Response options were Yes, No and Undecided. The EQAO documentation of the 2009-2010 Teacher Questionnaire results indicates a total of n=4853 respondents (EQAO, 2010b). The discrepancy in number of respondents was noted; however, it was found to be inconsequential with regard to the analysis of the constructed response component of question 22 because the quantitative results reported by the EQAO are consistent with the frequencies for Yes, No and Undecided reported here (EQAO, 2010b, p. 17).

Table 1 contains the frequency distribution of teacher responses for the winter and spring administration of the assessment. Please see Appendix 2 for a list of the codes and their frequency of occurrence.

Table 1. Frequencies for the selected-response and constructed-response parts of Question 22.

In your opinion, does counting some or all components of the Grade 9 Assessment of Mathematics as part of class marks motivate students to take the assessment more seriously?	Number of teachers who responded to the selected-response part of Question 22 (% of valid n)			Number of teachers who also supplied a written comment (% of valid n)		
	Winter	Spring	Total	Winter	Spring	Total
Yes	1810 (86)	2082 (86)	3892 (86)	444 (77)	515 (82)	959 (80)
Undecided	199 (9)	206 (9)	405 (9)	76 (13)	64 (10)	140 (11)
No	109 (5)	113 (5)	222 (5)	54 (10)	50 (8)	104 (9)
Total valid n	2118	2406	4519	574*	629**	1203

**Please note that there were 6 teachers who supplied a written comment but did not respond to the selected-response part of the question. These 6 teachers are not included in this number but their comments are included in the qualitative data analysis below.*
***Please note that there were 2 teachers who supplied a written comment but did not respond to the selected-response part of the question. These 2 teachers are not included in this number but their comments are included in the qualitative data analysis below.*

4.2 Formulated Meanings for Teacher Responses to the Constructed-Response Part of Question 22

To illustrate the approach to formulating meaning from teacher comments, teacher comments that were rich in information and informative for this study, as guided by the theoretical framework, were selected as examples to illustrate formulated meanings. The selected examples of significant comments and the related formulated meanings are found in Table 2.

Table 2. Selected examples of significant comments by teachers and related formulated meanings.

Significant Comment	Formulated meaning
"If you don"t count it, students either skip or do not take the time to answer the questions."	Student effort behaviors vary by test stakes.
"They do not take it seriously if they don"t feel they are „getting something" for it."	Students assess the value of engaging in the assessment.
"Last year it was mandated that the EQAO would count for 15% in lieu of a final exam. The results improved dramatically."	The value students attach to the assessment varies on characteristics of the test (e.g., weighting). Demonstrates a relationship between performance and test stakes.
"For those that are motivated by marks, they have the belief that these „marks" WILL BE VIEWED by others, thus they care. For those that have no interest in grades, it has no effect."	The effectiveness of using the assessment for class marks as a strategy to motivate students depends on the individual student"s self-schema, interests, and values.
"It does help, but by the end of the semester so many applied students have stopped working."	Student effort behaviours vary by characteristics of the test (e.g., timing) and characteristics of the student (program).
"Some students know we only count the multiple choice and take the open questions less seriously."	Students will select which portions of the test to apply their efforts based on their perception of value for items that count or do not count for marks. Student effort behaviour is related to their perceived ability and expectancies of success.
"They figure why bother trying if they won"t do well anyways."	
"Many students have commented „what is the point" of EQAO? With all of their courses getting busy with final assignments, they are not concerned about doing well on something that does not count as marks."	Demonstrates subjective task values made by students when they weigh the value of the assessment against their other school assessments.

These formulated meanings were arranged in clusters based on the teachers‟ response to Q22 (i.e., yes, undecided or no) and the themes (i.e., student characteristics student effort behaviours, test characteristics, and value). Table 3 shows the frequencies of the twelve clusters (3 response options times 4 themes of formulated meaning). The frequencies were calculated by counting the number of times a code pertaining to one of the identified themes occurred within a teacher comment for each of the three selected responses. These values are a percentage of the total number of codes recorded for the specified constructed response.

Table 3. Frequency of Theme Clusters as a Percentage of Total for Each Selected Response.

Total selected response with comment (n=1203)	Student characteristics	Student effort behaviours demonstrating motivation	Test characteristics	Value/non-value (apathy)	Comment student characteristics/ behaviour other
Yes (n=959)	7%	34%	14%	33% (value) 2% (non-value)	10%
Undecided (n=140)	16%	19%	19%	16% (value) 14% (non-value)	16%
No (n=104)	26%	20%	9%	13% value 21% (non-value)	11%

4.3 Coding and Qualitative Analysis of Teacher Comments

Table 4 consists of twelve clusters of formulated meanings that were based on the teachers‟ comments. These results are integrated into an in-depth description of teacher comments in the next section.

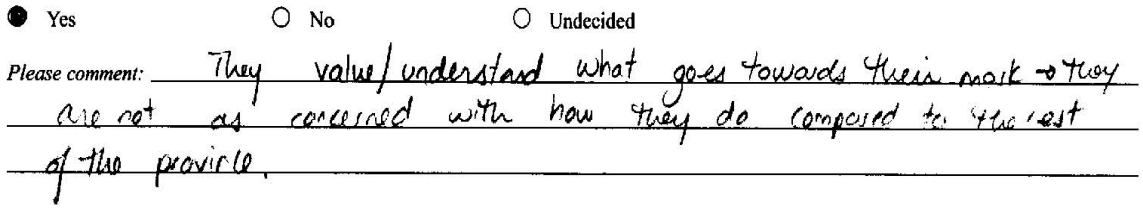
Table 4. Examples of theme clusters with their formulated meaning by teacher responses.

Constructed Response	Student Characteristics	Student Effort Behaviours and Motivation	Test or test items	Value
Yes	• However, I do think that it causes a lot of additional stress for students to have 2 extra math exams at the start of their high school career and further causes them to fear mathematics. • Otherwise they would not care, and would not try at applied level. • It seems to greatly motivate some but not others. • With a few exceptions, most of my applied students took the assessment seriously.	• Students are more motivated to prepare when the assessment counts. • Many students only do work when they are being marked. If the test is not worth anything, they will not try during the test or during the preparation. • Before the EQAO math assessment counted towards the marks on their report cards I saw a much higher degree of skipping and putting their heads down and dozing.	• Given the busy time of the year when the test is administered, students focus their efforts on assessments that count for marks. • Students focus their effort on only the test items that count for marks. • The weight of the assessment for class marks must be high enough to motivate students.	• They value/understand what goes towards their mark. They are not as concerned with how they do compared to the rest of the province. • Anything that does not count is unimportant and therefore no effort is generally put into it, to a grade 9 student. • Most common student questions are "Does it count"? and "For how much"?.
Undecided	• Counting the assessment for marks is a more effective motivation strategy for students in the Academic program. • The test is too difficult for students with lower math abilities. • Some students will not be motivated by a standardized test environment, but they may feel intimidated or their lack of ability unnecessarily revealed or exposed.	• Only for some students. Other students are satisfied with completing the assignment by guessing. • It provides a more concrete, immediate reason for being present and making a serious attempt. However, due to the anxiety caused by the unfamiliar format (despite having gone through the practice booklets), results do not usually reflect how a student is doing overall. • They barely prepare for final exam, so then EQAO priority is proportionately lower.	• Some students have difficulty with multiple-choice test items, multi-step test items, or an unfamiliar item/test format. • The timing of the test administration impacts the results. • Special Ed students have difficulty with multi-step questions that don't spell the steps out.	• The value that students assign to the assessment depends on student motivation and test anxiety. • The weighting is too low to have a significant impact. • Some students are indifferent towards assessments.

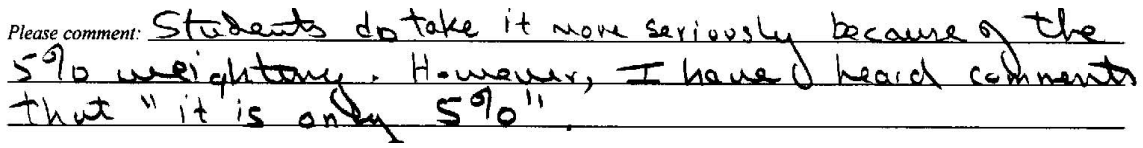
	• Some students don"t care either way; generally these are students who have already given up or students who have been unmotivated all semester.			• If teachers value the test, and communicate its importance to students, students will also value the test.
No	• Some Applied students care about passing the course but do not care about their final grade. • Some Applied students do not care about the assessment at all. • Some students experience substantial stress during the assessment even if it does not count for class marks. • Some students" approach to the test depends on their perceived math ability.	• Some students do not adequately prepare for the assessment even if it counts for their class mark. • Due to being applied students, they would work during class, but very few did any work outside the class. • They work hard no matter what the assessment is. • I told them that they should do the best they can regardless and they seemed to agree. However, it"s possible that they would not review as much if it did not count.	• For academic students, the 5% motivates them, but for applied students most just want to get the credit. Also, the EQAO doesn"t reflect their abilities since it is very literacy based. These kids don"t like to read. They find the test too "wordy" and then do not try.	• The students view EQAO testing as a waste of time. They do not get timely or informative results. • They do not take it seriously. They already have a terrible idea of what the test is about after writing so many standardized tests in elementary school. Also, it doesn"t count as much as the OSSLT so they see no consequences for doing badly.

4.3.1 Yes – Teacher response to Q22 with comment

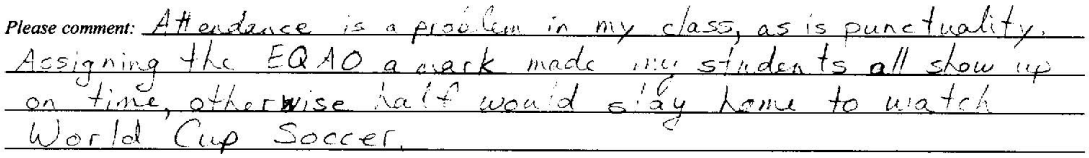
For those teachers that responded *Yes (*86% of respondents*)* and supplied a written comment (25% of total Yes respondents), comments were primarily focused on the themes of *Value of the Assessment* and *Student Effort Behaviors and Motivation*. Approximately one-third of teachers who responded *Yes* and supplied a comment specifically refer to the word value or terms that identify task value such as care, concern, worth, and counts. For example,



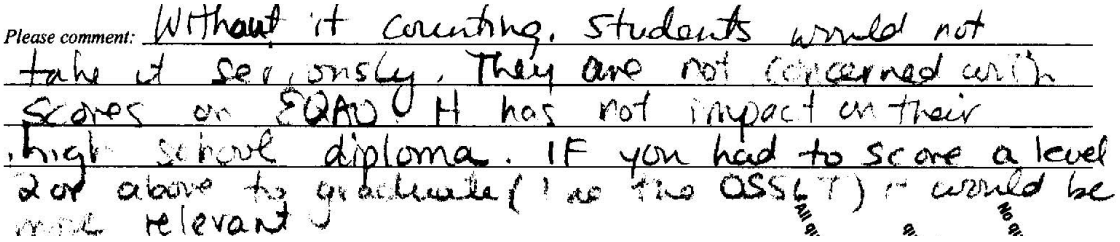
Teachers acknowledged that students better prepared for the test and persevered with difficult questions when the assessment counted; however, the weighting of the assessment played an important role in determining the extent to which students value the assessment. Some teachers felt that 5%, as indicated in the comment below, was too low to motivate students. For example,



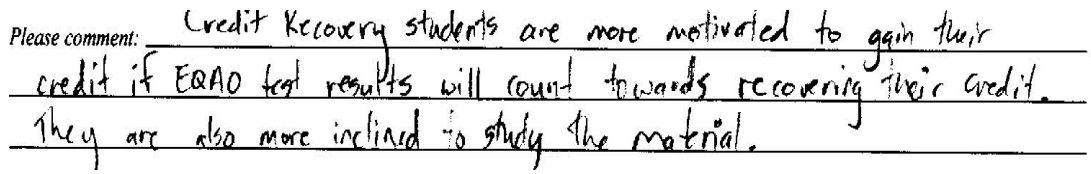
The majority of teachers agreed that counting the assessment, even for a small percentage, was necessary to ensure that some students attend the day of the test and try to complete it. For example,



Overall, the teachers commented that counting the assessment was an effective motivator for students. Some teachers suggested that the EQAO assessment be counted in lieu of a final exam, while others recommended that it be a graduation requirement. For example,



In addition to the theme of value, comments for those teachers that responded *Yes* also frequently referenced the theme of *Student Effort Behaviour and Motivation.* 34% of comments made by these teachers described a relationship between counting the assessment for class marks and motivation to take the test seriously or specific student effort behaviors such as more focused preparation for the assessment, greater attendance on assessment, and trying harder on the test. For example,



Some teachers also indicated that counting the assessment for class marks was a salient motivator for students that were borderline passing/failing, since the results may

determine a passing grade. Also, counting the test better motivated the students who cared about their marks throughout the term. For example,

Please comment: Some students who need to "boost" their mark at the end of the semester put a little more effort into the EQAO review + test, others "do not care" about it.

Please comment: It only motivates the students who care about their marks, for the others, it doesn't make a difference

For students who did not care about how high their mark was or for those that had no chance of passing the course, teachers generally felt that counting the assessment did not motivate the student to take it seriously. Some teachers acknowledged that some students did not care about the assessment regardless of the stakes.

Although relatively few teachers who answered *Yes* commented on specific aspects related to test items or format, the ones who did provided informative feedback. The most frequent comment was that students were aware that the multiple-choice questions counted towards the final grade, so students focused on the multiple-choice items. For example,

Please comment: Since only the multiple choice section counts towards their mark, the open response section is not taken as seriously

Please comment: Absolutely. Students heard rumours that only mc were going to count and asked if they could just leave open response blank. (We later told them 1 open response question was going to count)

A few teachers admitted to telling their students that the entire assessment would count when only the multiple choice items were going to be marked so that students would complete the open response questions.

Some teachers expressed concern over the level of literacy required to successfully complete some items and commented on the extent to which the assessment tested literacy over numeracy. For example,

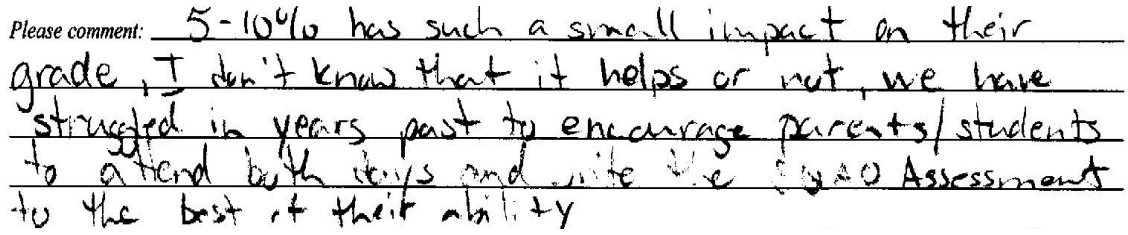
Please comment: Very limited → only if students are at risk of not receiving a credit; few want to maintain a good grade
Most - don't care
Test is tricky, literacy seems to override Math skills

The timing of the assessment was reported as problematic because the assessment is administered just prior to the students' regular exams. This is a busy time and students" performance on the assessment may not well reflect their best effort or ability because their time and attention may be divided between many competing tests and final assignments. Teachers also felt that student stress was compounded because of the demands being placed on them at the end of the term. For example,

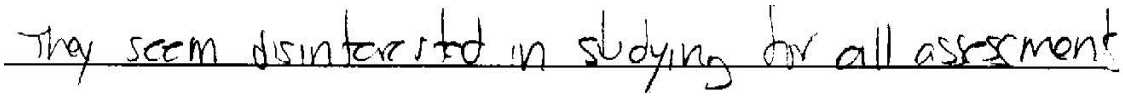
Please comment: Since the EQAO test is administered so close to the end of the semester, many students perform poorly because they are overwhelmed with summatives from other courses.

4.3.2 Undecided – Teacher response to Q22 with comment

For those teachers that responded *Undecided (*9% of respondents*)* and supplied a written comment (35% of Undecided respondents), comments focused on the themes of *Value of the Assessment* and *Students Characteristics*. Undecided teachers expressed concerns over the balance between weighting the test and student stress. They commented that giving the assessment some weight stressed some students, but weighting the assessment too low was insufficient to motivate others. For example,



The teachers who were undecided about counting the assessment toward class grades appeared to be unsure because they felt that some students care and some students do not care regardless of test stakes. Compared to the teachers that responded *Yes*, the undecided teachers had relatively greater number of comments regarding student apathy (14% vs. 2% for Yes respondents). For example,



The undecided teachers reported that the effectiveness of the strategy depended on the student. They listed student related variables such as ability (sometimes related to academic or applied courses), effort, stress, and engagement in preparation tasks as relevant factors in determining the effectiveness of the strategy. For example, the strategy was less effective for students of very high or very low ability or for those students who

experienced so much stress that it affected their performance on the test. The strategy was more effective if the student needed it to pass the course or valued high marks.

A few undecided teachers expressed concerns over item format and timing of the assessment and felt that some students struggled with multiple-choice, multi-step questions or the unfamiliar format of the test and therefore did not perform to the best of their abilities. For example,

Please comment: Special Ed Students have difficulty with multi step questions that don't spell the steps out.

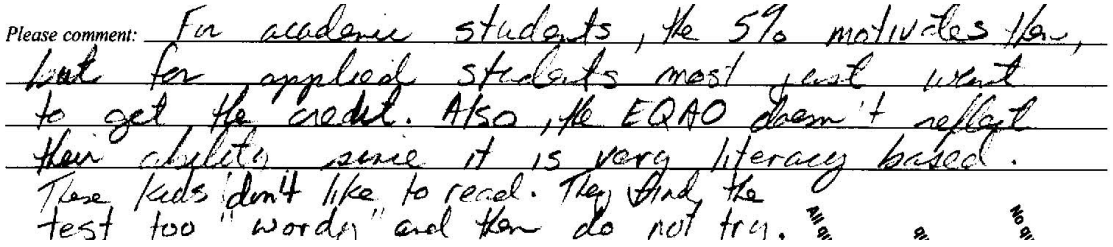
Please comment: It provides a more concrete, immediate reason for being present and making a serious attempt. However due to the anxiety caused by the unfamiliar format (despite having gone through practice booklets), results do not usually reflect how a student is doing overall.

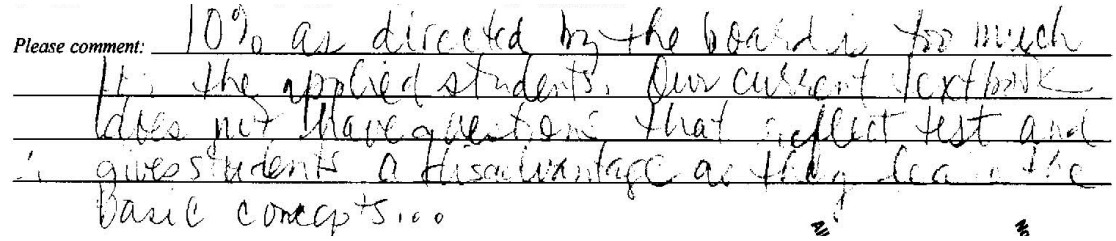
The issue of student stress was also mentioned with regard to the timing of the test because students had to prepare for two math exams at the same time and prepare for other final exams and assignments at the end of the term. For example,

Please comment: For some it is quite stressful to have two Math exams around the same time.

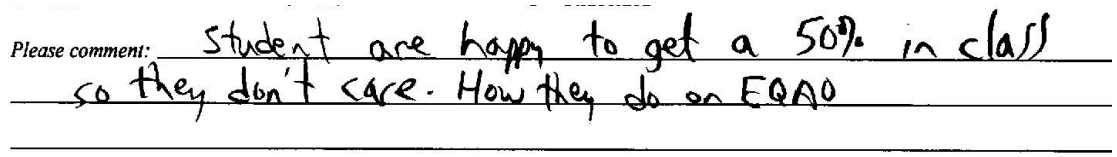
4.3.3 No – Teacher response to Q22 with comment

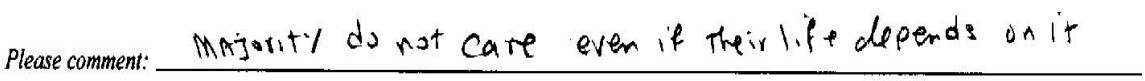
For those teachers that responded *No* (5% of respondents) and supplied a written comment (47% of No respondents), comments were spread out across the themes of *Student Characteristics* (26%), *Student Effort Behavior and Motivation* (20%), and *Value of the Assessment* (Apathy 21%). The comments provided by *No* respondents shared some similarities with those of the undecided teachers in that they acknowledged that the strategy might be effective for some students, but not for others. The main difference was the frequency with which they referred to the students" program and student apathy. Some teachers did not believe that counting the assessment toward class grades would motivate students in the applied mathematics program because they felt that the test was either too difficult, too different from what the students were familiar with, or too literacy based. Some students would have a difficult time completing the assessment; so counting the assessment for class marks was not an effective motivator. For example,



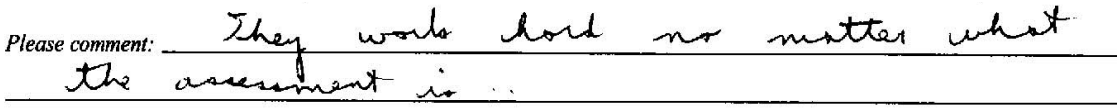


Student apathy was also a common concern among teachers who responded *No,* and many of their comments indicated that their students would not put forth effort on the assessment and did not care. For example,





Of the three selected response options, teachers who responded *No* tended to provide comments that were the most divergent in perspective. While many of these teachers" comments related to apathy and lack of concern on the part of students, there were also several comments that revealed a good deal of optimism and belief that students would try their best no matter what, so some teachers who responded *No* did so because they felt that counting the assessment toward class grades was unnecessary, since they believed their students would put forth effort under any circumstance. For example,



Chapter 5

Discussion

5.1 Frequencies of Responses

The first research question asked: What percentage of teachers who administer the EQAO Grade 9 Assessment of Mathematics, 2010 administration, responded *Yes*, *No*, or *Undecided* that counting the assessment toward class grade motivates students to take the assessment more seriously? One of the most powerful findings of the present study was the overwhelming number of *Yes* responses. 85% of teachers responded affirmatively when asked if they believed that counting some or all components of the assessment as part of class marks motivates students to take the EQAO Grade 9 Assessment of Mathematics more seriously. It is also noteworthy that the same percentage of teachers (85%) responded *Yes* for both the winter and spring administrations, and the consistency this demonstrates adds strength to the findings. Of the remaining responses, 9% of teachers responded were undecided, and 5% responded *No*.

These findings are not consistent with those of Koch (2009) who found that most teachers were not certain that counting the assessment resulted in students taking the test more seriously. Different methods for obtaining teacher opinion were used in the two studies, and although Koch (2009) carried out an extensive analysis of EQAO documents, the question analyzed in this study was not part of the Teacher Questionnaire at that time, so it was not available to include it as part of that study. It is possible that sample size and availability of data that specifically addresses the question used in this study contributed to the conflicting observations.

The results of Wolf and Smith's (1995) study showed that student motivation was higher when the assessment counted toward part of the course grade, and their findings support the assumptions of the majority of teachers in this study. Wolf and Smith (1995) also observed a significant relationship between anxiety, consequence and performance, which was recurrent in teacher comments analyzed in the current study.

Even though the frequency of *No* and *Undecided* selected responses was small compared with the *Yes* response, the comments provided by these teachers were informative in terms of the nature of the themes they drew on to support their positions. There was variation observed in reasoning provided in support of each of the three selected-response choices, and this data speaks to the second research question.

5.2 Variations in Opinion by Response

The second research question asked: How did their opinions vary by their response to the selected response question (e.g., yes, no or undecided), by math program (e.g., academic or applied), and by major themes in the literature (e.g., student qualities, test or test items, perceived value of the assessment)?

The three response choices shared two common themes: motivation and value; however, their interpretation varied depending on the response. Teachers who responded *Yes* tended to describe a strong relationship between value and test stakes, believing that the test needed to count toward student class grades for the students to take the test seriously. The *Yes* respondents often indicated that counting the test was important for students in the applied program because these students tended not to work unless there were marks attached to the task. Teachers who responded *No* often expressed the opposite sentiment, believing that counting the assessment would only motivate students

who already cared about grades, and since students in the applied program didn't care, counting the assessment would be of little utility in motivating those students. The *Undecided* teachers most frequently expressed concerns about the impact of student anxiety and stress as an outcome of counting the assessment toward class grades.

Other differences in the themes emerged while coding the open response-comments. The teachers who responded *No* were more inclined to comment on apathy (21%), stress (12%), academic or applied program (10%), and timing of the test (9%). There were teachers who commented that counting the test did not matter because they felt their students would put forth their best effort regardless; however, more often, the teachers who responded *No* supplied less optimistic written comments about the prospect of motivating their students. These teachers' comments listed apathy as a barrier to motivation under any circumstances and believed this was most especially true for students in the applied program.

The teachers who responded *No and Undecided* often drew a relationship between students in the applied program, ability, stress and/or apathy. Applied courses focus on the essential concepts of a subject, and develop students‟ knowledge and skills through practical applications and concrete examples. The documents that were coded for this study were not identified as coming from teachers of applied or academic courses, but some teachers self-identified the courses they taught. When applied students were mentioned in comments provided by *No* and *Undecided* respondents, they often indicated that their applied students were not well prepared for the test, may not do well with the format of the test or the wording of the questions. Although few teachers made specific reference to ESL students, some teachers did have concerns about the wording of the test

and that their students might find the test "tricky." Conversely, teachers who responded *Yes* tended to take the position that their applied students would not try at all if the test did not count for marks and that the test might help some of their students pass the course.

There are a number of accommodations and special provisions for students that have Individual Education Plans (IEPs), so it may be a matter of teachers making sure that all possible accommodations are implemented for their applied students. Ensuring that regular classroom lessons and activities expose students to the types of problem solving approaches found on the EQAO grade 9 assessment would also help to prepare students over the school term. Some applied students may need more practice with application of math concepts and more exposure to hands-on activities to acquire an understanding of the topics that is beyond procedural (Kajander, Zuke, & Walton, 2008).

5.3 Informing Expectancy-Value Theory of Motivation

The third research question asked how the results of the study inform expectancy-value theory of motivation, and the results do support the tenants of the theory. Teachers very clearly linked motivation with value and value with effort, whether or not they believed counting the assessment toward class grades was an effective motivational strategy. Expectancy-value models conceptualize students' ability beliefs, expectancies for success, and task value judgments as the driving forces behind motivation to engage in a task. If students are motivated to engage, one would expect to observe behaviours such as preparation and persistence on the test itself. These are the behaviours that teachers frequently identified as indicators of motivation by making specific reference to preparation (17%) and trying on the test (22%). Teachers also commented on student

characteristics, specifically ability and program, as predictors of motivation related to the value theme. These comments most frequently included the words value, care, concern, and worth, and conversely, don't care and apathy in relation to student program and anxiety.

Teacher comments that made reference to student apathy or lack of concern provide an opportunity to consider those students who did not appear to respond to test stakes (i.e., counting the assessment toward class grades). These comments sometimes included concerns about item format and excessive literacy demands of the test as being barriers for some students and thus lowering their expectancy of success. More often, teacher comments implied a complete disregard or concern for the assessment or marks. To understand these students, it may be useful to shift attention away from test stakes to goal-setting mechanisms.

Goal orientation, as a sub-construct of task value, also determines the degree to which motivation results in effort as a primary predictor of performance. Young children, assuming that they are healthy and able, start out in school wanting to learn. They have not yet internalized the concept of success or failure; they are oriented toward play, which is a pathway to learning that involves exploration and discovery without external performance outcomes and consequences for performance not achieved. Once in school, they are confronted with expectations of performance that are set by other people. In other words, their behaviour or reason to act is perceived as less self-determined. Ryan and Deci (2000) refer to this as locus of causality that swings between impersonal, external and internal, with intrinsic motivation being associated with an internal locus of causality or self-determined behaviour. All children experience this shift from self-

determined choice of tasks to engage in to tasks that are assigned as a part of the school experience; however, for reasons not fully understood, children adopt different mechanisms for responding to this experience.

Some children adapt to the external expectations by seeking out new knowledge. They seem to be oriented to accept tasks and challenges for the sake of mastery and acquiring new skills and understanding. In a sense, they continue to approach assigned tasks as they did self-determined play and seem less concerned about the fact that they did not determine the task presented. These children tend to be confident in their ability and undaunted by failure, which they generally attribute to lack of effort on their part or some other external factor over which they have control. Other children exhibit a performance goal orientation, and these children engage in tasks for positive reward or feedback or to demonstrate their superiority over other students. When they fail, their coping strategies are maladaptive. They tend to attribute failure to their own internal, stable traits such as ability. They become frustrated easily, and over time and repeated failure, they begin to adopt task avoidance behaviours to escape negative feedback and appearing inferior or stupid in comparison to others (Kajander, Zuke, & Walton, 2008; Pintrich, 2000).

For the students in this study who were reported to be apathetic or unconcerned with the test or marks, repeated failure in mathematics through school up to Grade 9 has likely led to decreased value in succeeding in mathematics and possibly other domains of learning as well. Low perceived competence results in non-relevance and a host of maladaptive behaviours such as lack of preparation, guessing on the test, or absence on the day of the test. For these students, failure with the least amount of damage to self-

esteem is desired and accomplished by attributing failure to lack of interest and effort in the domain of learning. In essence, this strategy helps them to compartmentalize and contain failure, so they can maintain a positive view of self by focusing on tasks they can be successful at and dismissing those for which there is little or no expectancy for success. For these students, counting the test or increasing test stakes as a strategy to motivate is ineffective.

Teacher comments indicated that counting the test did motivate some students to prepare for and put forth effort on the test. Although teachers indicated that counting the test may effectively motivate some students to take the assessment more seriously, that does not necessarily mean that counting the test or increasing test stakes is the most effective strategy for optimizing student effort and performance. In LSA scenarios, even students who normally exhibit mastery learning orientations may switch over to performance or ego-involved goals, and under these circumstances metacognitive strategies are used less, and effort becomes the primary predictor of performance (Sungur, 2007). The teacher comments recognize this as well and note that students in the academic program put forth effort because they care about their marks. High achieving students may approach regular classroom learning with mastery goal orientation because deeper learning and understanding is their objective in the classroom; however, the same students may not preserve this strategy when they write an LSA.

5.4 Informing Validity Theory in the Context of Large-Scale Assessments

The fourth research question asked how the opinions expressed by teachers on counting the assessment toward class marks relate to or inform validity theory. The validity of the assessment was a concern of this research for two main reasons. First, it

was recognized that student motivation could be a source of construct-irrelevant variance on this assessment because of the effect of test stakes on value. If the test stakes are low, task value is reduced for some students, and this was acknowledged by teachers who commented that the weighting of the assessment is too low or indicated that it should be a graduation requirement. If students are not motivated to prepare, put forth effort, and complete as much of the test as they are able, then performance may be underestimated and the scores will not be an accurate representation of ability due to CIV introduced by low student motivation.

The literature on motivation and the findings of this research confirm that determining test stakes is a critical component of test design because it is an important contributing factor to task value for both low and high achieving students. Understanding the impact of test stakes on performance for the EQAO Grade 9 Assessment of Mathematics is further complicated by the fact that test stakes vary from school to school and board to board and can range anywhere between zero and 30% of the final grade, but the majority of schools weight it between 5% and 15% of the final grade (EQAO, 2010b). This means that student motivation can vary from school to school and board to board as the stakes vary. This should be of concern for several reasons including validity and fairness as the basis of quality assessment as described by Rahn, Stecher, and Goodman (1997) who acknowledge that compromises are sometimes required as test designers strive to find an acceptable balance between quality and feasibility. However, in the case of this assessment, the combination of threats to validity and the loss of standardization suggest a need to examine the consequences of counting the assessment toward student final grades in the manner that it is currently undertaken.

5.5 Convergence of Data

The decision to use a mixed methods design is based, in large part, on the belief by the researcher that both sets of data will support and enhance the other. When a concurrent triangulation design is used in the research, it is important to report the extent to which the quantitative and qualitative data converge, and in the case of this study, the convergence is high. The very high percentage of *Yes* responses to the selected response component of Q22 indicates that teachers do believe that counting the test or test stakes is connected in some way to student motivation. Teacher comments were very direct on this point, and the most common theme observed can be distilled down to one concise comment, "If the test doesn't count, the students wont' try." Even when teachers responded *No* to the selected response item, some of these teachers qualified that response by acknowledging that counting the test might have some utility in motivating students, but the stakes might need to be as high as graduation requirement to have any effect on their students. Conversely, the teachers also made note that any gains made in motivating some students by counting the assessment may be lost in the case of other students whose performance may suffer due to increased anxiety caused by the potential impact of the test on their final grades.

The reported results provide evidence of convergence between the quantitative and qualitative data collected in this study. The results of the quantitative data analysis were clear and the teacher comments so definitely and pointedly expressed, I was quite confident in my ability to correctly interpret the qualitative component and to use the quantitative data to compliment the qualitative interpretation and provide additional evidence of confirmation of findings.

Chapter 6

Summary and Conclusions

In this study, a large majority of teachers reported that counting some or all components of the Grade 9 Assessment of Mathematics as part of class marks motivated students to take the assessment more seriously. Counting the assessment encouraged students to prepare, attend the assessment days and make an effort during the assessment. The weight of the assessment towards class marks was important in that it needed to be high enough to motivate the students to make an effort, but not so high that it caused them to perform poorly due to stress. Most teachers felt that assigning any weight, however, was helpful. Further, some teachers commented that students would only make an effort on those test items that they knew counted, for example, only on the multiple-choice portion of the test.

When teachers were undecided or were negative about the strategy of counting the assessment to motivate students to take the assessment seriously, their comments suggested that the success of the strategy depended on student-related variables such as ability, test anxiety, and program (academic or applied). It was not a successful strategy for some students who were anxious or were not motivated by marks. These students were sometimes identified as being in the applied program, and the frequent concern expressed by teachers for students in the applied program indicates that research is required to determine what is needed to assist these students in meeting the provincial standards and how best to assess their achievement.

6.1 Recommendations for Further Study

Teacher comments frequently revealed concerns over the weighting of the assessment toward class grades. Many teachers indicated that the weight that was being used in their school was insufficient, and others suggested that passing the assessment should be a graduation requirement. As an outcome of this study, it is recommended that further research be conducted on the weighting of this assessment toward students‟ final grades. It may be useful to consider what parts of the test should be counted and for how much so that student motivation is, as much as is possible to determine, maximized by this practice. At the same time, care must be taken to prevent the introduction of other sources of CIV such as stress and test anxiety. Finally, validity arguments for the use of this assessment as a contributing component of students‟ grades must be well formulated.

The results of this study are consistent with other research on test consequences and student performance. It contributes new insights, however, into the detailed themes that arise when teachers are asked to comment on the topic. These themes can be used to flag issues and inform the design of future research studies on the topic. It is interesting to note that the code clusters and themes fit well with the expectancy-value theory of motivation and also led the researcher to look deeper into the connections between expectancy-value theory, goal orientation theories and attribution theory. This led to the conclusion that there is a need for a more unified approach that may involve bringing these theories together to provide a more precise framework with which to guide future research on test consequences and student performance.

References

Abdelfattah, F, (2010). The relationship between motivation and achievement in low-stakes examinations. *Social Behaviour and Personality, 38*(2), 159-168. doi: 10.2224/sbp.2010.38.2.159

Ames, C. (1992) Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*(3), 261-271. doi: 10.1037/0022-0663.84.3.261

Bong, M. (2004). Academic motivation in self-efficacy, task value, achievement goal orientations, and attributional beliefs. *Journal of Educational Research, 97*(6), 87-297. doi: 10.3200/JOER.97.6.287-298

Boslaugh, S. (2007). *Secondary data sources for public health: A practical guide.* New York, NY: Cambridge University Press.

Cheng, L., Klinger, D. A., & Zheng, Y. (2007). The challenges of the Ontario Secondary School Literacy Test for second language students. *Language Testing, 24*(2), 185-208. doi: 10.1177/0265532207076363

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*(4), 609-624. doi:10.1016/j.cedpsych.2007.10.002

Cooper, S., Porter, J., & Endacott, R. (2010). Mixed methods research: A design for emergency care research? *Emergency Medicine Journal.* doi:10.1136/emj.2010.096321

Creswell, J. W. (2003). *Research Design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W. (2011). Controversies in mixed methods. In N. K. Denzin and Y. S. Lincoln (Eds.) *SAGE handbook of qualitative research* (4th ed., pp. 269-283). Thousand Oaks, CA: Sage.

Creswell, J. W., Plano Clark, V. L., Guttman, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioural research* (pp. 209-240). Thousand Oaks, CA: Sage.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research.* Thousand Oaks, CA: Sage.

Crooks, T. J., Kane, M., & Cohen, A. S. (1996). Threats to the valid use of assessment. *Assessment in Education: Principles, Policy & Practice*, *3*(3), 265-286. doi: 10.1080/0969594960030302

DeMars, C. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55-77. doi:10.1207/s153248ame1301_3

Denzin, N. K. (1978). *Sociological methods.* New York: McGraw-Hill.

Downing, S. M., & Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education, 38* (3)*,* 327-333. doi: 10.1046/j.1365-2923.2004.01777.x

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality and development.* Philadelphia: Taylor and Francis/Psychology Press.

Dweck, C., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review 95*(2), 256-273. doi: 10.10370033-295X.95.2.256

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values and goals. *Annual Review of Psychology, 53,* 109-132. doi: 10.1146/annurev.psych.53.100901.135153

Eklof, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003 *International Journal of Testing, 7*(3), 311-326. doi: 10.1080/15305050701438074

Education Quality and Accountability Office. (2008). *EQAO's Technical Report for the 2006 – 2007 Assessment.* Toronto, Ontario, Canada: Queen‟s Printer for Ontario.

Education Quality and Accountability Office. (2010a). *Driving Purposeful Improvement: 2009-2010 Annual Report*. Toronto, Ontario, Canada: Queen‟s Printer for Ontario.

Education Quality and Accountability Office. (2010b). *EQAO's Provincial Secondary School Report on the Results of the Grade 9 Assessment of Mathematics and the Ontario Secondary School Literacy Test (OSSLT), 2009-2010 English-Language Students.* Toronto, Ontario, Canada: Queen‟s Printer for Ontario.

Education Quality and Accountability Office. (2011). *Grade 9 Assessment of Mathematics Administration Guide 2011.* Toronto, Ontario, Canada: Queen‟s Printer for Ontario.

Fairbairn, S. B., & Fox, J. (2009). Inclusive achievement testing for linguistically and culturally diverse test takers: Essential consideration for test developers and decision makers. *Educational Measurement Issues and Practice, 28*(1), 10-24. doi: 10.1111/j.1745-3992/2009/01133.x

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher, 36*(8)*,* 456-462. doi: 10.3102/0013189X07311607

Greene. J. P., Winters, M. A., Forster, G. (2004). Testing high stakes tests: can we believe the results of accountability tests? *Teachers College Record, 106*(6), 1124-1144.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27. Doi: 10.1111/j.1745-3992.2004.tb00149.x

Hulleman, C. S., Durik, A. M., Schweigert, S. A., & Harackiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology, 100*(2), 398-416. doi: 10.1037/0022-0663.100.2.398

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 29, 2010 from http://www.cehd.umn.edu/NCEO/onlinepubs/Techinical37.htm

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.

Kajander, A., Zuke, C. & Walton, G. (2008). Teaching unheard voices: Students at-risk in mathematics. *Canadian Journal of Education, 31*(4), 1039-1064. Doi: 10.3102/0013189X08315390

Kiplinger, V. L., & Linn, R. L. (1996). Raising the stakes of test administration: The impact on student performance on the National Assessment of Educational Progress. *Educational Assessment 3*(2): 111-133.

Klinger, D. A. & Luce-Kapler, R. (2007). Walking in their shoes: Students‟ perceptions of large-scale high-stakes testing. *Canadian Journal of Program Evaluation, 22*(3), 29-52.

Klinger, D. A., Rogers, W. T., Anderson, J. O., Poth, C, & Calman, R. (2006). Contextual and school factors associated with achievement on a high-stakes examination. *Canadian Journal of Educational Research, 29*(3), 771-797.

Koch, M. J. (2009, May). Validation in the context of multiple-use: Investigating the multiple use of a large-scale mathematics assessment. Paper presented at the meeting of the Canadian Society of Studies in Education, Ottawa, ON.

Lam, T. C. M., & Bordignon, C. (2001). An examination of English teachers‟ opinions about the Ontario grade 9 reading and writing test. *Interchange, 32*(2), 131-145. doi: 10.1023/A:1011938109935

Lissitz, R. W, & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 6*, 437-448. doi: 10.3102/0013189X07311286

Levin, B. (1993). Students and educational productivity. *Education and Policy Archives, 1*(5). Retrieved from http://epaa.asu.edu/epaa/v1n5.html

Literacy and Numeracy Secretariat. (2006). *Schools on the Move: Lighthouse program.* Retrieved July 16, 2011, from http://www.edu.gov.on.ca/literacynumeracy/onthemove.pdf

Locke, E. A., & Latham, G. P. (2004). What should we do about motivation theory? Six recommendations for the twenty-first century. *Academy of Management Review, 29*(3), 388-403.

Marshall, C., & Rossman, G. B. (2006). *Designing Qualitative Research.* Thousand Oaks, CA: Sage.

Mattern, R. (2005). College sutdents‟ goal orientation and achievement. *International Journal of Teaching and Learning in Higher Education, 17*(1), 27-32.

Merriam, S. B. (2002). *Qualitative Research in Practice: Examples for discussion and analysis.* San Francisco: Jossey-Bass.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (*3rd ed., pp. 13-103). NewYork: Macmillan.

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

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*(8), 463-469.

Murphy, K. R. (2009). Validity, validation and values. *Academy of Management, 3*(1), 421.

Office of the Auditor General of Ontario. (2010). *2009 Annual Report.* Retrieved October 8, 2010 from http://69.164.72.173/en/reports_en09/304en09.pdf

Ontario Ministry of Education. (2005). *The Ontario Curriculum, Grades 9 and 10: Mathematics, 2005 (revised).* Available from http://www.edu.gov.on.ca/eng/curriculum/secondary/math.html

Ontario Ministry of Education. (2011). *Education Facts.* Available from Ontario Ministry of Education Web site, http://www.edu.gov.on.ca/educationFacts.html

O‟Neil, H. F., Sugrue, B., & Baker, E. L. (1995/1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment, 3*(2), 135-157.

Pajares, F. M. (1992). Teachers‟ beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62*(3), 307-332.

Pajares. F. M. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*(4), 543-578. doi: 10.3102/00346543066004543

Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology, 92*(3), 544-555.

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications.* Upper Saddle River, NJ: Merrill Prentice-Hall.

Rahn, M. L., Stecher, B. M., & Goodman, H. (1997). Making decisions on assessment methods: Weighing the tradeoffs. *Preventing School Failure,41*(2), 85-89.

Roderick, M., & Engel, M. (2001). The grasshopper and the ant: Motivational responses of low-achieving to high-stakes testing. *EducationalEvaluation and Policy Analysis, 23*(3), 197-227.

Rogers, W. T., Anderson, J. O., Klinger, D. A., & Dawber, T. (2006). Pitfalls and potential of secondary data analysis of the Council of Ministers of Education, Canada, National Assessment. *Canadian Journal of Education*, *29*(3), 757-770.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25,* 54-67.

Skaalvik, E. (1997). Issues in research on self-concept. In M. Maehr & P. R. Pintrich (Eds.), *Advances in Motivation and achievement* (Vol. 10, pp 51-97).

Stobart, G. (2001). The validity of the National Curriculum Assessment. *British Journal of Educational Studies, 49*(1), 26-39.

Stocking, M., Steffen, M., & Eignor, D. (2001). *A method for building a realistic model of test taker behavior for computerized adaptive testing* [Research report]. Princeton, NJ: Educational Testing Service.

Sungur, S. (2007). Contribution of motivational beliefs and metacognition to students‟ performance under consequential and nonconsequential test conditions. *Educational Research and Evaluation, 13*(2), 127-142.

Suurtamm, C., Lawson, A., & Koch, M. (2008). The challenge of maintaining the integrity of reform mathematics in large-scale assessment. *Studies in Educational Evaluation, 34*, 31-43.

Tashakkori, A., & Teddlie, C. (1998). *Mixed Methodology: Combining qualitative and quantitative approaches.* Thousand Oaks, CA: Sage.

Ungerleider, C. (2006). Reflections on the use of large-scale student assessment for Improving student success. *Canadian Journal of Education, 29*(3), 873-883.

van Barneveld, C., King, S., Simon, M., & Nadon, C. (2010). Final report on the analysis of teachers‟ responses to the constructed-response questions on the grade 9 assessment of mathematics teacher questionnaire, winter and spring 2010 administration. Ontario, Canada. Lakehead University, Faculty of Education.

van Barneveld, C., Stienstra, W., & Stewart, S. (2006). School board improvement plans in relation to the AIP model of educational accountability: A content analysis. *Canadian Journal of Education, 29*(3), 839-854.

Watt, H. M. G., Eccles, J. S., & Durik, A. M. (2006). The leaky mathematics pipeline for girls: A motivational analysis of high school enrolments in Australia and the USA. *Equal Opportunities International, 25*(8), 642-659.

Weiner, B. (1975). *Achievement motivation and attribution theory.* Morristown, New Jersey: General Learning Press

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68-81.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95-114.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-17.

Wise, S. L., & DeMars, C. E, (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*, 27-41.

Wise, S. L., Bhola, D. S., & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice, 25*(2), 21-30.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment, 11*(1). 65-83.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequences of performance, test motivation, and mentally taxing items. *Applied Measurement in Education,* 8, 341-351.

Wolf, L. F., & Smith, J. F. (1995). The consequence of consequence – motivation, anxiety, and test- performance. *Applied Measurement in Education, 8*, 227-242.

Wolfe, R., Childs, R. A., & Elgie, S. (2004, May). Final report of the external evaluation of EQAO‟s assessment processes. Toronto, ON: Education Quality and Accountability Office.

Wolming. S., & Wickstrom, C. (2010). The concept if validity in theory and practice. *Assessment in Education: Principles, Policy & Practice, 17*(2), 117- 132.

APPENDIX 1

Code List for Teacher Questionnaire Question 22

Theme: Response to selected-response portion of the question

Code	Definition and notes	Sample quote
Resp_yes	Respondent chose the yes option	**22.** In your opinion, does counting some or all components of the Grade 9 Assessment of Mathematics as part of class marks motivate students to take the assessment more seriously? ● Yes ○ No ○ Undecided
Resp_no	Respondent chose the no option	**22.** In your opinion, does counting some or all components of the Grade 9 Assessment of Mathematics as part of class marks motivate students to take the assessment more seriously? ○ Yes ● No ○ Undecided
Resp_undecided	Respondent chose the undecided option	**22.** In your opinion, does counting some or all components of the Grade 9 Assessment of Mathematics as part of class marks motivate students to take the assessment more seriously? ○ Yes ○ No ● Undecided
Resp_none	Respondent did not choose any option (blank)	**22.** In your opinion, does counting some or all components of the Grade 9 Assessment of Mathematics as part of class marks motivate students to take the assessment more seriously? ○ Yes ○ No ○ Undecided
Rest_multiple	Respondent chose more than one of the three possible options	**22.** In your opinion, does counting some or all components of the Grade 9 Assessment of Mathematics as part of class marks motivate students to take the assessment more seriously? ● Yes ● No ○ Undecided

Theme: Comments related to student effort behaviors and motivation

Code	Definition and notes	Sample quote
Comment_motivation	Text specifically refers to the word motivation or a derivative (e.g., motivate, motivator, motive).	Please comment: The opposite is more true. If they know it doesn't count for marks, there will be much less motivation for most students. Emphasis on seriousness on my part translates into motivation & seriousness on their part.
Comment_sb_effort	Text described student behaviors that related to trying harder or putting forth effort.	Please comment: The grade 9 applied students often try harder when a grade is assigned to a task.
Comment_sb_prep	Text described student behaviors related to studying or preparing for the assessment.	Please comment: The EQAO (Grade 9 math) should be treated like the OSSLT. It should count and be placed on the student's transcript. Will give teachers and students more incentive to prepare and do well. This will increase students scores. If one wants to take this assessment seriously it needs to count.
Comment_sb_absent	Text described student behaviors related to being absent for the test or skipping the entire test.	Please comment: If the assessment does not count towards their mark, many students will not write the assessment.

Comment_sb_guess	Text described student behaviors related to guessing on some items.	Please comment: Most applied students struggle with this assessment and they find it very difficult so they often guess; especially with multi step questions. There is too much content on these tests and they simply cannot remember the processes.
Comment_sb_omit	Text described student behaviors related to omitting some items on the test or lack of completion of assigned work.	Please comment: Before it counted Applied Math Students would give up on the question since they did not have to pass it.
Comment_sb_other	Text described student behaviors other than those listed above (e.g., being disruptive during the writing of the test).	Please comment: Otherwise they would just quickly guess at MC and leave open responses blank or not show up for school or be disruptive in class.

Theme: Comments related to student characteristics

Code	**Definition and notes**	**Sample quote**
Comment_sc_ability	Text described student ability.	*Please comment:* For my applied students, they don't (in general) care because they know that they are going to fail anyways (Applied avg. grade in province is ~45%). The test is way too difficult for Applied students and has been for years. Nothing has been able to change that.
Comment_stress	Text referred to stress or related concepts (e.g., student stress, teacher stress, feeling overwhelmed).	*Please comment:* No, the anxiety and stress levels are high to begin with. Students will do as well as they can no matter what.
Comment_sc_program	Text referred to program of study (applied or academic)	*Please comment:* 9 Applied typically have little concern about assessment
Comment_sc_language	Text described student characteristics related to language.	*Please comment:* Yes, but the students that I teach get frustrated sometimes because they don't understand what the question is asking because they are ESL students.

Comment_sc_other	Text described student characteristics other than those listed above or identifies a subgroup of students specifically or more generally.	*Please comment:* Only some students

Theme: Comments related to the value of the assessment

Code	**Definition and notes**	**Sample quote**
Comment_value	Text refers to value of test and/or test related tasks. The definition of value includes care, concern, importance, and interest.	*Please comment:* It only motivates the students who care about their marks, for the others, it doesn't make a difference
Comment_sc_apathy	Text described student characteristics related to apathy. This code is used when the comment indicates lack of care or indifference.	*Please comment:* Majority do not care even if their life depends on it

Theme: Comments related to the test and test items

Code	**Definition and notes**	**Sample quote**
Comment_test_item_position	Text described student behaviors related to the position of items on the test (e.g., towards the end of the test)	*Please comment:* We are using the multiple choice only, so my Students have completed those first. The test was too long and so a number of them have not completed the written answer component.
Comment_test_item_format	Text referred to the format of items on the test (e.g., multiple choice item, open response items, multi-step items)	*Please comment:* But only if you don't tell students which components will be marked so that they will work hard on all questions. Since we count multiple-choice questions, then field-test MC questions need to be consistent throughout all students booklets!
Comment_test_language	Text referred to wording or language in reference to questions on test.	*Please comment:* This is used only in communication to me that they are doing their best. In the end EQAO may count for less than 1% of their final mark. The wording of questions, especially if applied, coupled with their weak reading skills can be unfair to some. EQAO is not to be considered a punishment - rather as a precursor to the test that really counts - the final exam. Have they learned what has been taught in order to be successful in Grade 10? (Not always the same emphasis as eqao)

Comment_weight	Text included a statement about the weight (e.g., 5%) of the test towards classroom assessment.	Please comment: students realize that unlike a 3% [illegible] exam, the 5% EQAO weighting reduces the exam to only 25% and is therefore less stressful to the students
Comment_grad	Text included a statement about the test as a graduation requirement.	Please comment: Because the numeracy EQAO is not a requirement for the OSSD, many students deem the assessment as a joke.
Comment_timing	Text included a statement that described some aspect if time (e.g., time of year, time to complete).	Please comment: There are so many projects and summatives going on in other class while this assessment is given. Students are overwhelmed with the workload.

Theme: Other comments

Code	Definition and notes	Sample quote
Comment_reasons _other	Text described reasons to mark the EQAO assessment OTHER than motivating students for the EQAO test (e.g., as practice for final exam).	Please comment: It acts a preparatory activity for the final exam, so students take it seriously and prepare.
Comment_other	Text was anything other than the codes above.	Please comment: It may be an attempt by teachers to influence student attitudes about the Assessment, but I don't know if student attitudes are altered.
Comment_illegible	Text was illegible. Reasons may include illegibility due to poor penmanship or to poor scan quality.	Please comment: However, it is conducted at an extremely stressful time in the school year and covers most of the material on the exam itself.
Comment_none	Blank response	Please comment:
Comment_page-_blank	No responses recorded for any questions on the page	
Wrong_page	The wrong page was scanned.	

Appendix 2

Code Frequencies

	Frequency of the Code		
Theme and code descriptions	Winter	Spring	Total
Theme: Value of the assessment			
Text refers to value of test and/or test related tasks. The definition of value includes care, concern, usefulness, importance, and interest.	127	307	434
Text described student characteristics related to apathy. This code is used when the comment indicates student indifference, lack of care or concern for marks or assessment.	17	42	59
Theme: Student effort behaviours and motivation			
Text specifically refers to the word motivation or a derivative (e.g., motivate, motivator, motive).	50	100	150
Text described student behaviors that related to trying harder or putting forth effort.	38	112	150
Text described student behaviors related to studying or preparing to write the EQAO assessment.	17	70	87
Text described student behaviors other than those listed above (e.g., being disruptive during the writing of the test).	34	1	35
Text described student behaviors related to being absent for the test or skipping the entire test.	8	25	33
Text described student behaviors related to omitting some			

items on the test or lack of completion of assigned work.	6	18	24
Text described student behaviors related to guessing on some items.	2	3	5
Theme: Student Characteristics			
Text included reference to student program (i.e., applied or academic)	26	62	88
Text described student characteristics other than those listed above or identifies a subgroup of students specifically or more generally.	40	44	84
Text referred to stress or related concepts (e.g., student stress, teacher stress, feeling overwhelmed).	18	25	43
Text described student ability.	7	13	20
Text described student characteristics related to language fluency	4	0	4
Theme: Test or test items			
Text included a statement about the weight (e.g., 5%) of the test towards classroom assessment.	21	58	79
Text included a statement about the test as a graduation requirement.	14	27	41
Text included a statement that described some aspect if time (e.g., time of year, time to complete).	16	23	39
Text referred to the format of items on the test (e.g., multiple choice item, open response items, multi-step items)	7	29	36
Text referred to wording or language in reference to questions on test.	4	0	4

Responses not relevant to question asked, blank, or illegible			
Blank response	281	940	1221
No responses recorded for any questions on the page	277	938	1215
Text was anything other than the codes above.	41	104	145
Text described reasons to mark the EQAO assessment OTHER than motivating students for the EQAO test (e.g., as practice for final exam).	12	33	45
Text was illegible. Reasons may include illegibility due to poor penmanship or to poor scan quality.	10	9	19