uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Hossam Meshref
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.A. (Education)
GRADE / DEGREE

Faculty of Education
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Linking Teachers' Classroom Assessment Practices and Students' Achievement on the Saip 2001 Mathematics Assessment

TITRE DE LA THÈSE / TITLE OF THESIS

Marielle Simon
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

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

David Trumpower

Christine Suurtamm

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

LINKING TEACHERS' CLASSROOM ASSESSMENT PRACTICES AND STUDENTS' ACHIEVEMENT ON THE SAIP 2001 MATHEMATICS ASSESSMENT

Hossam Meshref

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the degree of Masters of Arts in Education

Faculty of Education
University of Ottawa

Thesis Supervisor: Prof. Marielle Simon
Thesis Committee: Prof. Christine Suurtamm and Prof. David Trumpower

Your file Votre référence
ISBN: 978-0-494-61338-2
Our file Notre référence
ISBN: 978-0-494-61338-2

ABSTRACT

Linking Teachers' Classroom Assessment Practices and
Students' Achievement on the SAIP 2001 Mathematics Assessment

Hossam Meshref

The current study answers the question: What are the relationships between teachers' assessment practices and students' mathematics achievement in large-scale assessment tests? The investigated teachers' assessment practices and students' achievement were defined according to the School Achievement Indicator Program (SAIP) 2001 mathematics achievement test and its accompanying teachers' questionnaire. The study focused on the problem solving component of that test. The assessment practices were investigated within three dimensions: assessment tools (e.g. teacher made multiple-choice tests, projects, and portfolios), assessment purpose (e.g. feedback, and diagnosis), and homework purpose (e.g. feedback and grading).

Correlational analyses were employed to investigate whether there are relations between the assessment variables and students' mathematical achievements. Research findings proposed a few relations between students' achievement and the investigated assessment practices. The directions and the strengths of these relations varied based on a few factors such as the distribution of teachers' assessment practices, the amount of missing data, and the data deletions during analyses. The research findings could contribute to the literature on mathematical assessment, and also propose potential relations from which teachers, administrators, and policy makers may predict possible impacts of assessment practices on students' achievement.

ACKNOWLEDGMENTS

Nothing would have been done without the grace and power of God Almighty. I am most thankful to Him for supporting me throughout this program with His countless bounties.

I wish to express my deep appreciation to Professor Marielle Simon, my thesis supervisor. She patiently guided, supported, and taught me during my program of study. Her invaluable support seemed endless in both time and effort. As well, I am very grateful to my committee members Professor Christine Suurtamm and Professor David Trumpower for working with me at every stage of my research. I truly appreciate their insightful input and feedback.

I seize this opportunity to thank the Council of Ministries and Education Canada for sharing the SAIP data. A special thank-you is also extended to Ms. Marthe Craig and Mr. Herve Jodouin for assisting me in understanding the SAIP 2001 mathematics achievement data.

Last but not least, I am deeply grateful to my wife and kids for their beautiful spirit, love, and support. I particularly owe a great deal of gratitude to my wife, who took care of the family when I left to pursue my master's studies. My great appreciation and thanks go to my mother and father for their endless love and support throughout my life. They gave me all of the things that have gotten me here.

TABLE OF CONTENTS

LIST OF TABLES vi

LIST OF FIGURES vii

CHAPTER I: Introduction 1

Assessment Issues 2

The Focus of This Study 5

CHAPTER II: Literature Review 7

Using Assessment Information to Improve Students' Achievement 8

Using Homework to Improve Students' Achievement 11

Relating Teachers' Assessment Practices and Students' Achievement on PIRLS 13

Relating Teachers' Assessment Practices and Students' Achievement on TIMSS 16

Previous Research Done on SAIP 22

Summary of the Findings of the Literature Review 24

CHAPTER III: Methodology 27

Framework 27

Sampling 31

Instrumentation 31

Procedures 37

Data Analysis 37

CHAPTER IV: Results 40

Characteristics of the data bases 40

Description of Students' Achievement Dataset 46

Description of Teachers' Assessment Practices Dataset 47
Examining Correlations between Teachers' Assessment Practices and Students' Achievement 53
CHAPTER V: Discussion 57
Findings Related to Assessment Tools 57
Findings Related to Assessment Purpose and Homework Purpose 60
Limitations 62
Anticipated Contributions of the Current Study 65
Suggestions for Future Research 67
CHAPTER VI: Conclusion 69
REFERENCES 71

LIST OF TABLES

Table 1 SAIP Background Questionnaire Items Related to Teachers' Assessment Practices 36

Table 2 Percentage of Missing Data Level per Item within the Original Teacher Questionnaire Database 41

Table 3 Distribution of Class Size Found in the 266 Classes Remaining for Imputation 45

Table 4 Percentage of Teachers Giving Weight to Different Assessment Tools in the Practices-Achievement Dataset 48

Table 5 Percentage of Teachers Using Assessment Information for Various Purposes in the Practices-Achievement Dataset 49

Table 6 Percentage of Teachers Using Homework for Various Assessment Purposes in the Practices-Achievement Dataset 50

Table 7 Percentage of Teachers Giving Weight to the Various Assessment Tools in the Original Dataset 51

Table 8 Percentage of Teachers' Frequency of Use of Assessment Tools for Various Purposes in the Original Dataset 52

Table 9 Percentage of Teachers' Frequency of Use of Homework for Various Assessment Purposes in the Original Dataset 52

Table 10 Correlations between Teachers' Assessment Practices and Students' Achievement on the SAIP 2001 Mathematics Achievement Test (Problem Solving Component) 54

Table 11 Comparison of Significant Correlation Results of the Two Data Files (n>=10 / All) 56

LIST OF FIGURES

Figure 1. Relating teachers' assessment practices and students' achievement 30

Figure 2. Items from the SAIP 2001 mathematics content component 32

Figure 3. Items from the SAIP 2001 mathematics problem solving component 33

Figure 4. Class_Size distribution in the practices-achievement database after applying the criterion 43

Figure 5. Distribution of students' performance in the achievement database 47

CHAPTER I

Introduction

Student's success is the centre of attention of all educators, policy makers, and researchers. To determine the level of students' achievement, teachers have to assess their learning. During the classroom assessment process, teachers gather, and synthesize information pertaining to students' performance in a certain subject for a variety of reasons. One reason behind collecting assessment information could be for the purpose of instruction modification (McMillan, 1997). It could also be done for a diagnostic purpose at the beginning of a teaching/learning cycle to determine students' a priori knowledge and to prepare lesson plans. In addition, assessment can serve summative purposes at the end of learning where students are given a chance to demonstrate their achievement (Cizek, 1997; Popham, 1995). Moreover, assessment could be conducted as learning and teaching take place for the purpose of feedback (Brookhart, 2008; Earl, 2003; Chappuis & Stiggins, 2002). With this feedback students could answer questions such as "where am I right now?", and "how can I fill the gap to reach my goals?"

Currently, researchers suggest three purposes to classroom assessment: assessment *of* learning, assessment *for* learning, and assessment *as* learning (Chappuis & Stiggins, 2002; Earl, 2003; Stiggins & Chappuis, 2005). In assessment of learning, the teacher uses assessment information in order to report on students' achievement and grade it, whereas in assessment for learning, the teacher uses formative assessment as he gathers information and uses it to plan and improve instruction. On the other hand, in assessment as learning the student plays an important role as an assessor making sense of what is needed to master the targeted skills. Consequently, a

student can monitor his own learning, and perform the necessary steps to improve his understanding. A teacher can facilitate a combination of these three approaches in order to improve his students' learning and hence their achievement.

Assessment Issues

The previous conceptions about classroom assessment show a variety of purposes for which assessment could be used. The teacher has to choose the proper assessment tool that will fulfill the assessment purpose and eventually improve learning. This choice also differs from one teacher to another according to teachers' educational backgrounds and their students' learning styles. One of the issues regarding such variation is the extent to which it leads to improvement in students' achievement. Improvement is plausible when the teacher uses different assessments based on an informed decision. In that sense, the teacher's awareness of which assessment tool to use for which assessment purpose is anticipated to positively impact students' learning, and hence their understanding. Therefore, another related issue has to do with how teachers develop knowledge that enables them to reach such informed decision?

On the one hand, teachers could be informed by reading articles and stay updated with research, but that is a solitary effort that may or may not occur. As well, they may attend professional development workshops, which is more likely to occur as part of teachers' professional growth plans. However, one important source of opportunities for professional development is through policy, drawn to provide teachers with best practices to use within their classrooms. On the other hand, policy makers also need to be informed by best practices that could lead to better achievement. The intervention of research to help in this situation is vital to provide policy makers and teachers with the adequate knowledge that links practice and research.

Teachers as researchers in their classroom do not always have enough time or capacity to carry out a valid research experiment. Conversely, researchers do have time and aptitude to conduct research, but they do not always have access to the data. Therefore, researchers, as reform initiators, may intervene by collecting their data from the classrooms, or use data previously collected by other agencies, and then explore these data to propose useful findings that could help vested stakeholders. Motivated by such needs and data availability, my research efforts are devoted to the uncovering of possible links between teacher assessment practices and students' achievement. In addition, the study is expected to produce findings that contribute to the body of research on assessment as well as inform policy makers and educators about different assessment practices and their impact on students' achievement.

A few researchers have studied assessment practices and their relation with students' achievement through studies conducted in the actual classroom (e.g. McMillan, 2003; McMillan, Myran & Workman, 2002; Stiggins & Conklin, 1992). Other researchers chose to use secondary data analysis to investigate large scale assessment scores and questionnaire results to examine such relationships (e.g. Dudaite, 2006; Hastedt, 2004; Jones, 2004; Rodriguez, 1999). In my research, secondary data analysis is used to examine possible relations between teachers' assessment practices and students' achievement. Using secondary data analysis, for research purposes, has some advantages over other methods such as observation or surveys, because the data are already accessible (Barribeau et al, 2005). The previous paragraphs in this section addressed motivations behind conducting the current study to improve students' achievement in general. However, the next paragraphs deal with more specific motivations pertaining to improving achievement on mathematics.

As mentioned earlier, teachers' careful choices of proper assessment practices are important to the teaching and learning process. A suitable choice of assessment tools enables teachers to provide constructive feedback to their students on what they understand and also on how to improve their understanding (Callingham, 2008, p. 19). For example, when assessing achievement in mathematics, Barlow and Drake (2008) argue that classroom assessment practices are supposed to measure the level of students' understanding (p. 236). But again, math teachers need to be informed about how to use their assessment practices to adequately measure such understanding.

Given that educators are held accountable by stakeholders for their students' understanding, they are expected, collectively, to employ suitable instruction and assessment practices that should, in principle, improve students' achievement on large-scale assessment tests. Failing to do so may have a negative impact on their students' understanding of mathematical concepts, particularly in areas where higher order thinking is needed (e.g. in problem solving). For example, if classroom math teachers focus on a specific assessment tool such as multiple-choice questions tests, it may not adequately reflect the level of their students' understanding (Nelson & Sassi, 2007). On the other hand, if teachers use open-ended questions asking students to explain their reasoning during problem solving, students' answers could provide teachers with evidences about students' understanding. However, these hypotheses need empirical evidence, and that is one of the main objectives of this study.

Lastly, the need for measuring students' understanding urged policy makers in Canada to call for deeper and comprehensive analysis of large-scale achievement math tests and survey data. For instance, the Council of Measurement in Education of Canada (CMEC) is urging measurement specialists to focus their efforts on investigating possible relations between

contextual and educational factors and students' achievement in mathematics (CMEC, 2003). My research's secondary analyses are anticipated to provide insight into these potential relationships. Those relations, or lack thereof, could provide a reference point from which teachers can anticipate in advance the impact of their assessment practices on their students' achievement in mathematics large-scale assessment test.

The Focus of This Study

In this study, teachers' assessment practices are investigated as defined by SAIP 2001 math teachers' questionnaire. The purpose of this investigation is to examine possible relations between teachers' assessment practices and students' achievement in the problem solving component of the SAIP 2001 mathematics achievement test. Researchers who did research on SAIP argue that the use of secondary data analysis of SAIP has "merit" and they acknowledged that it can contribute to Canadian education (e.g. Rogers, Anderson, Klinger, & Dawber, 2006, p. 759). Therefore, using a reliable source of information such as SAIP data provides a solid ground for this study to build upon. If the proposition that certain assessment practices in math class impact students' achievement is proven to be true, the current research findings could confirm relationships between teacher assessment practices and students' achievement in mathematics. As well, since policy makers have the resources to modify assessment standards and to provide professional development accordingly, it would be possible at that point to draw policies that would enhance teachers' knowledge about assessment practices.

The next chapter investigates the current literature on studies that examined classroom assessment practices in general, studies that linked large-scale assessment results and classroom assessment practices, and those studies that specifically focused on SAIP. This is followed by the

methodology chapter, where the conceptual framework describes assessment practices that are expected to affect students' achievement. The data analysis results follow the methodology chapter as it highlights the current trends found in teacher assessment practices, and proposes correlations between teachers' assessment practices and students' achievement in SAIP 2001 mathematics achievement test. Finally, the study wraps up with the discussion of the findings and the overall conclusion of this research.

CHAPTER II

Literature Review

The literature review brings forth certain themes that are expected to have a bearing on the relation between teachers' assessment practices and students' achievement. The first section discusses research efforts reported in the literature that focused on using assessment information to improve students' achievement (e.g. Black & Wiliam, 1998; Brookhart, 2008; Chappuis & Stiggins, 2002; McMillan, 2003; Ciofalo & Wylie, 2006; Rodriguez, 1999). It was noticed that among classroom practices, homework was investigated in the literature as a tool that has a noticeable effect on students' achievement (e.g. Cooper, 1989; Cooper, Robinson & Pattal, 2006; Rodriguez, 1999). Therefore, the second section sheds some light on homework as an assessment practice that supports learning and improves students' achievement.

Given that this study uses secondary analysis to explore the relationship between teachers' assessment practices and students' achievement, it was necessary to devote the third and the fourth sections of the literature review to studies of a similar nature (e.g. Dudaite, 2006; Hastedt, 2004; Jones, 2004; Rodriguez, 1999). The two sections review studies that relate other teacher assessment practices such as the use of multiple-choice questions and projects to students' achievement in large-scale assessment tests such as Progress in International Reading Literacy Study (PIRLS) and Trends in International Math and Science Study (TIMSS). Unfortunately, none of the research found in the literature relating teacher assessment practices to students' achievement investigated SAIP, which is the database under investigation in this study. It was then essential to devote the fifth section to review research that focused on SAIP to get information about the nature of its datasets, as well as the challenges which researchers have

encountered (e.g. Anderson et al, 2006; Angelis, 2003; Emenogu, 2006; Rogers, Anderson, Klinger, & Dawber, 2006). In the final section, a summary of the findings discussed in the literature review is provided.

Using Assessment Information to Improve Students' Achievement

A few research studies, done on classroom assessment, focused on a variety of assessments practices and their best use in order to improve students' achievement (e.g Black & William, 1998; McMillan, 2003; Rodriguez, 1999). On the other hand, some theoretical work proposed different useful uses of assessment practices that are also anticipated to improve students' achievements (e.g. Chappuis & Stiggins, 2002; Ciofalo & Wylie, 2006; McMillan & Workman, 1998; Stiggins, 1994; Stiggins & Chappuis, 2005).

In their survey of 160 journals, Black and Wiliam (1998) found that the use of formative assessment to provide accurate and frequent feedback to students yields significant learning gains (p. 7). They argued that most of the teachers' practices that focused on merely assessment of learning encouraged superficial and rote learning. In particular, they argued that most teachers used assessment information for assigning grades rather than enhancing learning. Black and Wiliam believe that the previous trend in teachers' classroom practices reduces the level of students' understanding, and dramatically influence their performance, especially in situations where higher order thinking is invoked (e.g. in mathematics problem solving).They assert that "For assessment to be formative the feedback information has to be used" (p.16). In other words, the collected assessment information should be used to provide feedback rather than merely providing grades. According to Black and Wiliam's perception, a good classroom assessment practice that could develop students' understanding would be that teachers collect learning

evidences as they assess students' progress. Teachers then use this information to provide an ongoing feedback and adjust instruction accordingly, thereby improving students' understanding and hence their achievement.

These conclusions were echoed by McMillan (2003) who surveyed 79 Grade five teachers about their teaching and assessment practices. His goal was to relate their practices (e.g. the use of assessment tools such as multiple-choice tests or the use of assessment information for feedback) to students' performance in the Standards of Learning (SOL) large-scale assessment in math and reading/language arts test scores. McMillan asked teachers as to what extent do they use assessment to provide immediate feedback to students using a five-point Likert scale ranging from *not at all* to *extensively*. His analyses showed a small positive relationship between the use of formative assessment for feedback and students' achievement in reading/language arts ($r=0.23$). However, the relation with students' achievement in mathematics did not approach significance ($r=0.02$, *ns*). He ascribed the small correlations to having high mean scores and a small standard deviation, which restricted the ranges and eventually affected significance. In addition to this limitation, he also argued that the small sample size hindered the detection of different relationships between the variables of the study.

The aforementioned positive relationship between the use of formative assessment information for providing feedback and students' achievement attracted attention to other purposes for using assessment information. In their article, Chappuis and Stiggins (2002) suggested that teachers should conduct pretests before units for diagnostic purposes. They argued that such collected information helps teachers to adjust and plan their instruction according to their students' needs. In addition, diagnostic assessment is expected to help teachers find out what their students' know and more importantly, what they do not know at the beginning of a

learning cycle. Ciofalo and Wylie (2006) asserted that when teachers neglect using diagnostic assessment this may result in holding back some students' learning. They argued that if a student has a misconception about a central concept in mathematics, for example, learning new related concepts will be stalled until that misconception is removed. The work by Chappuis and Stiggins (2002) and Ciofalo and Wylie (2006); however, is theoretical and does not provide empirical evidence relating the use of assessment information for diagnostic purposes to students' achievement.

On the other hand, Rodriguez (1999) proposed empirical evidence relating the use of assessment information for diagnostic purposes and students' achievement in TIMSS 1995 mathematics achievement test. He used students' test scores as a measure of students' achievement. In addition, he used TIMSS teachers' questionnaire responses as a measure of assessment practices (e.g. their use of assessment tools such as short-answer tests, essays, or the use of assessment information for feedback or diagnostic purposes). TIMSS asked teachers about how often they use the mathematics assessment information to provide feedback to students and diagnosing students' learning problems. A four-point Likert scale was used ranging from *none* to *a great deal*. Rodriguez's analyses revealed that the frequency of use of assessment information for diagnostic purposes had a significant negative, but weak correlation with students' achievement ($r=-0.04$). He argued that according to his Hierarchical Linear Model (HLM), it was unlikely that the frequency of use of assessment information for diagnostic purposes would impact the residual variance in classroom achievement levels (p. 121). The proposed weak correlation, however, warrants further investigation to validate this relationship.

In addition to using assessment activities during class time for feedback and diagnostic purposes, teachers also assigned homework after school hours for different purposes. The

literature had many studies that confirmed the existence of a relation between the use of homework for different assessment purposes and students' achievement. The following section shed some light on different studies that focused on that topic.

Using Homework to Improve Students' Achievement

Homework can be assigned for a variety of purposes such as practice, feedback, grading, or disciplinary purposes. In general, the amount and time allocated for homework were found, in the literature, to be related to students' achievement (Cooper, 1989; Cooper, Robinson & Pattal, 2006). Cooper (1989) conducted a meta-analysis of 120 empirical studies done on homework and synthesized reported correlations between time spent on homework and students' achievement on standardized tests and grades. He found that studies done on mathematics reported the highest average correlations (r=0.22) between the two variables.

Recently, Cooper et al (2006) reviewed more than 60 research studies done on homework between 1987 and 2003, and concluded that the amount of homework is positively related to achievement as measured by standardized tests, grades, or composite achievement scores. The reported regression coefficients ranged from 0.09 to 0.16 for Math; 0.12 to 0.28 for Reading; 0.09 to 0.23 for Science; and 0.11 to 0.18 for Social Studies. Although these two reviews are considered the most comprehensive studies done on homework covering research done between 1930 and 2003, they did not focus on using homework as an assessment tool for feedback or for grading purposes. Instead, homework purposes were divided into instructional purposes (e.g. practicing materials presented in the class) and non-instructional purposes (e.g. establishing communication between parents and child).

On that last topic, two theoretical articles asserted that homework should not be used solely for grading purposes but also for improving learning and providing feedback (Tuttle, 2008; Wormeli, 2006). Tuttle suggested using homework as a formative assessment tool to improve student's proficiency in different strands of mathematics. He argued that by discussing homework in the classroom and providing feedback, students can underline parts that need improvement and write comments on how to enhance it. He asserted that "this self correction starts them on a formative path" (p.1). This type of feedback is expected to help them overcome difficulties they may encounter in problem solving such as misreading the question, getting confused during the process, or misunderstanding the mathematical concept behind the question (p.1). On the other hand, Wormeli (2008) argued that homework should not be used only for grading. He asserted that students' grades reflect, to a certain extent, their mastery of a certain concept. Based on that, he argued that students should not be graded while they are in the process of developing concept understanding (e.g. using problem solving to understand Place Value). However, these arguments remain theoretical and stronger evidence is needed to relate using homework as an assessment tool to students' achievement.

In his study of TIMSS 1995 mathematics achievement data, Rodriguez (1999) investigated the effect of discussing homework in the classroom and providing feedback to the students on their achievement. He found that students' achievement has a positive relation with the frequency of use of homework for the purpose of feedback ($r=0.12$). This finding complements McMillan's (2003) study's findings pertaining to the use of formative assessment for feedback. These studies, together with Tuttle's (2008) assertion, propose that using homework as a formative assessment tool has a positive relation with students' achievement in

mathematics. Lastly, regarding using homework for grading purposes, Rodriguez found that it has a negative but weak correlation with students' achievement (r=-0.06).

The previous propositions supported the argument that the use of homework assessment information for feedback purposes could positively impact students' achievement in mathematics (e.g. McMillan, 2003; Rodriguez, 1999; Tuttle's, 2008). On the other hand, the use of homework assessment information for grading purposes was argued to negatively impact students' achievement in mathematics (e.g. Rodriguez, 1999; Wormeli, 2008). Tuttle and Wormeli's assertions; however, were theoretical, while Rodriguez and McMillan's correlations were low. This study is expected to complement these efforts by conducting further investigation to validate the aforementioned relations.

Relating Teachers' Assessment Practices and Students' Achievement on PIRLS

While reviewing the literature, it was found that two studies examined the PIRLS 2001 database to understand the relationship between teachers' assessment practices and students' achievement in reading (e.g. Hao, 2005; Hastedt, 2004). Hao used students' achievement scores and their teachers' questionnaire information, collected by PIRLS, to investigate the relation between teacher assessment practices and the achievement of Grade four students in reading literacy. The practices included using multiple-choice, short-answer, and essay questions on materials read, group projects, tests, quizzes, as well as interviews about what the students read. In the teachers' questionnaire, PIRLS asked teachers about how often they ask their students to do a group project, or take a quiz or test about what they have read. A 4-point Likert scale was used ranging from *everyday or almost everyday* to *never or almost never*. Hao's results revealed a slight negative correlation between the frequency of use of multiple-choice, short-answer

questions, essays, and interviews and student achievement. However, the relation between the frequency of use of group projects, quizzes and tests and students' achievement was inconclusive.

Although Hao's study is not on mathematics, his analyses of the use of assessment tools (e.g. short-answer questions and essays) could inform this study. Santa Cruz (2009) argued that students who face challenges in reading and writing are likely to continue facing these challenges in their math classes. For example, those students may face difficulties while trying to understand the language of a math problem and at the same time find it difficult to express their reasoning as they solve. Santa Cruz anticipated that math teachers may have to perform double the work with those students as they plan their instruction on language objectives and mathematics objectives.

Based on the above argument, assessment practices that are related to students' achievement in reading and writing may provide some insight into potential relations between assessment practices and students' achievement in mathematics. Therefore, it is believed that Hao's (2005) findings could be useful to the current study. However, the slight negative correlation that Hoa found between the frequency of use of short-answer and essay questions and students' achievement seem to be counter intuitive to some extent. It is anticipated that if a student is frequently asked to explain his reasoning and gets the proper feedback in the classroom context, he is more likely to perform well on similar items on the standardized test. Such practices have been strongly promoted in current educational reforms which stress greater accountability. Therefore, Hao's findings regarding short-answer and essay questions necessitate further investigation to confirm or invalidate the reported relationship.

In another study on PIRLS, Hastedt (2004) focused on students' scores only. The purpose of his study was to evaluate the influence of item types used in PIRLS 2001 on students' scores. In particular, he investigated students' performance differences on multiple-choice and short-answer items used in PIRLS 2001. Despite taking guessing into account, his analyses confirmed that the short-answer items are harder for students than multiple-choice items (p. 2). He argued that the hardship ascribes to the additional information required in the case of short-answer items as an evidence of comprehension. Although Hastedt's study does not link teachers' assessment practices to students' achievement directly, it sheds some light on the effect of using different assessment tools on students' scores.

As teachers assess their students' understanding, they need concrete evidence about the level of that understanding before making a decision based on that assessment information. For example, having the right answer to a problem solving question item, posed in multiple-choice format, may be due to guessing. However, it is hard for the teacher to prove that guessing occurred without using other assessment tools (e.g. interviews). The teacher can also ask the student to answer the question again using short-answer format to evaluate his rationale. Having either of these follow up assessment steps done by the teacher after using multiple-choice item brings the question as to why did the teacher not use short-answer questions in the first place if assessing students' understanding were the goal? Therefore, it could make sense in mathematics' assessment that a few teachers may favor tools such as short-answer response questions over multiple-choice questions to tap into students' understanding.

In the previous discussion, the studies examined the relationship between teachers' assessment practices and students' achievement in reading literacy. It was argued that there are anticipated similar impacts on mathematics achievement when using assessment tools such as

short-answer questions, and essays. However, these relations were indirectly related to mathematics, and thus the findings need further investigation before also being considered valid within the mathematical context. The following section is devoted to studies that investigated the relation between teachers' assessment practices and students' achievement in mathematics (e.g. Dudaite, 2006; Jones, 2004; Rodriguez, 1999).

Relating Teachers' Assessment Practices and Students' Achievement on TIMSS

TIMSS provides data on the mathematics and science achievement of the 4th- and 8th-grade students around the world. In addition, it facilitates the comparison of students' achievements of the participating countries. TIMSS data have been collected in a cycle of four years: 1995, 1999, 2003, and 2007. To benefit from that large achievement database, the province of Ontario commissioned a research project that could be used to formulate a few recommendations to enhance the mathematics and science programs in Ontario (Jones, 2004). The study focused on different aspects such as curriculum, teacher education, achievement correlates (e.g. demographic variables, attitudes, assessment tools) and students' results.

Jones (2004) found that the higher frequency of use of some assessment tools such as quizzes, tests, and projects is associated with lower achievement (p. 4). He also mentioned that teachers in higher performance jurisdictions (e.g. Singapore, Japan, Alberta, and British Colombia) differently emphasized multiple-choice testing as an assessment tool. However, it was not clear if the word 'different' implied low or high frequency of use. Therefore, the strength of Jones' aforementioned associations as well as the effect of frequency of use of the aforementioned assessment tools on student's achievement need to be explored further to better understand the nature of such links.

In another study, Dudaite (2006) examined students' scores on TIMSS mathematics achievement test to find out how Lithuanian Grade 8 students' test results in TIMSS changed from 1995 to 2003, and the possible explanations for that change. He observed that the students' achievement in mathematics was low in TIMSS 1995 and he ascribed that to the advent of multiple-choice and open-ended response test items, which were new to Lithuanian students. He supported his argument with evidence that the educational reform that took place on assessment practices in Lithuania geared classroom assessment practices more toward mathematical literacy than toward academic style mathematics. Coping with assessment reform efforts, teachers started assessing their students using a variety of assessment tools such as multiple-choice tests and open-ended response tests. That, in turn, resulted in the improvement of students' understanding of mathematical concepts, thereby they were able to better demonstrate their knowledge when higher order thinking was invoked (e.g. in problem solving questions). Thus, Lithuanian students' achievement in mathematics in TIMSS 2003 improved as a result of the emphasis on the use of the aforementioned assessment tools in the math classroom.

Dudaite's findings proposed evidence that the use of assessment tools, such as multiple-choice tests and open-ended response tests could impact students' achievement in mathematics. However, his finding did not relate the frequency of use of these assessment tools to students' achievement. Consequently, there were no correlations describing the strength of the relation between the use of these assessment tools in the math classroom and students' achievement at large-scale assessment tests. On the other hand, Rodriguez's (1999) study, described earlier, provided more reliable evidence relating the aforementioned assessment tools and students' mathematics achievement on TIMSS 1995.

Rodriguez investigated correlations between teachers' use of different assessment tools such as teacher-made multiple-choice items, short-answer tests, essays, and students' achievement. He found that the frequency of use of teacher-made multiple-choice and short-answer tests was negatively correlated with students' achievement ($r=-0.16$). Although Rodriguez's study provided correlations, his findings contradict those found by Dudaite regarding the impact of use of multiple-choice items on students' achievement. The aforementioned contradiction necessitates more investigation to further validate the relation between such assessment tools and students' achievement in mathematics large-scale assessment.

Despite having the aforementioned contradiction, the existence of a negative relation appears to have a better rationale. Cizek (2009) asserted that large-scale assessment reports evidences about students' achievement as an "overall judgment", while classroom assessment is more likely to produce "high quality information" (p. 68). However, he argued that quality information is not always guaranteed. The current study echoes the abovementioned standpoint as it asserts that developing teacher made multiple-choice tests that accurately diagnose students' understanding is a critical task, and it is not always guaranteed to happen at the classroom level. Perhaps students' achievement in large-scale tests is likely to improve if the quality of information produced by multiple-choice test items is guaranteed at both contexts (e.g. classroom and large-scale assessment levels).

Based on the previous arguments, giving weight to teacher-made multiple-choice tests at the classroom level does not necessarily lead to improvement in large-scale assessment tests. Unfortunately, the information provided in TIMSS documents as mentioned by Rodriguez, did not offer any information on the quality of the teacher made multiple-choice tests. Therefore, it

is believed that the quality of teacher made multiple-choice tests, may have contributed, as a confounding factor, to the aforementioned negative relation between giving weight to that assessment tool and students' achievement.

It is worth mentioning that most of the discussed studies that focused on PIRLS and TIMSS (e.g. Dudaite, 2006; Hao, 2005; Hastedt, 2004; Jones, 2004) examined the effect of teachers' use of assessment tools and their impact on students' mathematics achievement on large-scale assessment. Rodriguez (1999), on the other hand, not only examined the relationship between students' achievement and teachers' frequency of use of different assessment tools, but also looked at the purpose behind using these tools (e.g. for feedback or diagnostic purposes). His goal was to relate teachers' assessment practices to students' achievement in large-scale assessment tests, while studying students' self efficacy and effort as a mediating role within this relation. He used teachers' questionnaire, students' questionnaire, and students' scores on the TIMSS 1995 mathematics achievement test to examine the aforementioned relations. In his conceptual framework, he divided assessment tools into two facets: *homework*, and *others*. The homework facet had two dimensions: types of tasks and frequency of use for different purposes such as feedback or grading. Similarly, the others facet had two dimensions: types (e.g. teacher-made multiple-choice and short-answer tests and projects) and the frequency of use for different purposes (e.g. for feedback or diagnosis).

As Rodriguez studied correlations between the purpose behind assigning homework and student's achievement, he found that the frequency of use of homework for feedback was positively correlated with students' achievement ($r=0.12$), while its use for grades had near Zero correlation ($r=-0.06$). Regarding the second facet, other assessment tools, Rodriguez found that the frequency of use of teacher made multiple-choice and short-answer tests are negatively

correlated with students' achievement (r=-0.16). Rodriguez also found that the frequency of use of teacher-made short-answer tests, projects and observations had near Zero correlation with students' achievement (r=0.04, r=-0.01, and r=-0.06 respectively).

Again, as mentioned before, Rodriguez's analysis of the relation between the purpose of use of assessment tools and students' achievement distinguishes his research from other correlational studies that looked only at the frequency of use of assessment tools. According to his findings, he found that the frequency of use of assessment information for feedback and diagnostic purposes had near Zero correlations with students' achievement (r=-0.07 and r=-0.04 respectively). He considered these correlations to be "virtually no different than zero" and concluded that it is unlikely to impact students' achievement (p. 121). These results contradict the findings mentioned earlier asserting the positive impact of feedback on students' achievement (e.g. Black & William, 1998; McMillan, 2003). Thus the need for further replication studies on the topic.

The multifaceted explorations that Rodriguez employed to understand the relations between teachers' assessment practices and students' achievement provide a valuable conceptual basis for future studies in this area. For example, there are similar items of interest across the teachers' questionnaire in TIMSS, PIRLS and SAIP. However, the magnitudes of most of the correlations among variables calculated by Rodriguez are not large enough to indicate strong associations (e.g. r=-0.01 for projects, and r=-0.06 for observations). The reasons behind having such low correlation may lie with his study conditions such as missing data and its impact on different levels on his HLM model.

In summary, the studies discussed in the third and the fourth sections of this chapter highlighted several possible relations between teacher assessment practices and students'

achievement (e.g. Dudaite, 2006; Hao, 2005; Hastedt, 2004; Jones, 2004; Rodreiguez, 1999). For studies that were focused on PIRLS (e.g. Hao, 2005; Hastedt, 2004), although the focus was on reading, the research methodologies used gave an idea about different assessment tools that characterize teachers' classroom assessment practice (e.g. multiple-choice, essays, and interviews). However, despite the fact that these studies used secondary analysis, there is a need for extra efforts to justify the strength of the relation between assessment tools and students' achievement.

For the studies that focused on TIMSS (e.g. Dudaite, 2006; Jones, 2004; Rodriguez, 1999), the focus was on achievement in mathematics and thus more directly related to the proposed research. Again, the strength of relation between frequency of assessment tools used and students' achievement varied in such a way that necessitate further analysis. For example, the higher frequency of use of some assessment tools such as quizzes tests and projects was associated with lower achievement (Jones, 2004). Rodriguez (1999) investigated the value of these correlations, but most of them were small ranging from -0.16 to 0.04. In addition, there seems to be a contradiction between Rodriguez's findings implying a negative relation between the use of multiple-choice tests and students' mathematics achievement in TIMSS 1995 and Dudaite's findings regarding the positive impact of the same tool on students' mathematics achievement in TIMSS 1995. Perhaps this contradiction may ascribe to the varying definitions of assessment tools and their uses across different educational settings (e.g. the math class in the US and in Lithuania). That in turn may have influenced teachers' perceptions of these assessment tools and their uses as well as students' perceptions of the assessment environment. Thus, teachers' assessment practices and students' achievement varied accordingly in a way that may have caused contradiction. The previous contradiction causes confusion about the nature of

aforementioned relation and calls for further investigation to validate the type and strength of that association using similar or new data sets.

In the following section, studies that focused specifically on SAIP are examined for possible findings or methodological information that could benefit the proposed research. It is important to see the nature of the students' achievement and teachers' responses databases, and the expected level of missing data as well as methods of treatment.

Previous Research Done on SAIP

SAIP is a cyclical program of assessments of student achievement in mathematics, reading and writing. SAIP was conducted between 1993 and 2004 by CMEC. It was then replaced by the Pan Canadian Assessment Program (PCAP) in 2007. Because of such short interval of application, not many studies have investigated the SAIP data. A few studies, however, focused on SAIP achievement scores and on data from school, teacher, and student questionnaires (Anderson et al, 2006; Angelis, 2003; Emenogu, 2006; Rogers, Anderson, Klinger, & Dawber, 2006). Some of these studies were exploratory in nature (e.g. Angelis, 2003; Emenogu, 2006). For example, Angelis conducted a study to propose a framework for students' attitudes towards Science based on reviewing the literature. Her second goal was to investigate the framework's application to the data obtained from the SAIP 1999 Science students' questionnaire. Emenogu's goal was to examine the effect of the method of handling missing data on the Mantel Haenszel approach that detects differentially functioning items. The studies by Angelis and by Emenogu however, did not investigate students' achievement in mathematics. Anderson et al (2006) and Rogers et al (2006); on the other hand, applied statistical modeling to the SAIP data sets to determine possible relationships between the SAIP 2001 mathematics

achievement scores and student, school, and home variables. Although these studies did not specifically relate students' achievement to teachers' questionnaire responses, they provided insight into the nature of SAIP data, and the recommended approaches to deal with it when conducting secondary analyses.

The primary investigation of the SAIP data by Anderson et al (2006) and Rogers et al (2006) revealed two important issues pertaining to missing data which are worth considering when dealing with SAIP databases. First, they did not find the data loss to be systematic because it was spread in the two age groups: the 13 and 16 years old. In addition, they argue that the level of loss was low and, as a result, it supported the feasibility of secondary analysis (Anderson et al, p. 710). However, missing data caused problems when correlations were calculated, because of the considerable deletion of incomplete students' records in their studies. To deal with that problem, Anderson et al and Rogers et al used Expectation Maximization (EM) estimation procedures for imputation.

Second, to minimize the effect of missing data, they dealt with students' questionnaire items that had an acceptable response rate. In their case, that acceptable response rate was 10% or less missing data per each investigated item (p. 711). Nevertheless, in this study, it is premature to consider the missing data a serious issue until the actual analysis of the SAIP 2001mathematics achievement database takes place. In general, the previously discussed points give an insight into the expected level of missing data in the students' achievement database and its suitability for secondary analysis. They also give a hint on how to choose items in the teacher background questionnaire to guarantee minimal effect of missing data by setting the 10% threshold.

Summary of the Findings of the Literature Review

The discussed literature revealed various findings on relations between teachers' assessment practices and students' achievement. The studies shared a common theme reflecting the positive impact of the use of assessment information for feedback and diagnostic purposes on students' achievement (e.g. Black & William, 1998; McMillan, 2003; Rodriguez, 1999). However, the values of these reported correlations were between -0.07 and 0.02, which are extremely low.

The literature also covered the use of homework for different assessment purposes. As mentioned earlier, most of the studies focused on the time spent and on the amount of homework (e.g. Cooper, 1989; Cooper, Robinson & Pattal, 2006). Other studies focused on the use of homework for the purpose of feedback and grading (e.g. Rodriguez, 1999; Tuttle, 2008; Wormeli, 2006). The empirical evidence provided by Rodriguez showed a relationship between students' achievement in large-scale achievement tests and the use of homework for feedback and grading. However, the correlations that he provided were also low (0.12 and -0.06 respectively).

Finally, the use of different assessment tools and their relation to students' achievement in large-scale tests were the focus of many studies in the literature (e.g. Hao, 2005; Hastedt, 2004; Jones 2004; Rodriguez, 1999). Most of these studies focused on PIRLS and TIMSS databases such as students' scores, and teachers' questionnaire response databases. Their main efforts were to explore possible relations between different assessment tools used in the math classroom (e.g. multiple-choice and short-answer tests) and students' achievement in large-scale assessment tests (e.g. TIMSS). Some of the studies reported negative correlations between students' achievement and the frequency of use of teacher-made multiple-choice tests (e.g. Hao,

2005, Rodriguez, 1999). Other studies contradicted the previous findings as they reported positive correlations between the frequency of use of multiple-choice tests and students' achievement (e.g. Dudaite, 2006; Jones, 2004).

In their efforts to uncover the aforementioned relations, it was noticed that most of the studies used the HLM model in their analyses. It is believed that it is more suitable for the nature of large-scale achievement databases (e.g. students nested within classrooms nested within schools). However, the discussion of missing data levels and corresponding treatment needed to be more evident in most of these studies. Without having such information, the validity of the reported findings could be questionable. The studies by Anderson et al (2006) and Rogers et al (2006), although not focused on the relations between assessment practices and students' achievement, gave some insight that could help in crafting a few plausible explanations. In their studies, the volume of the missing data at different levels in the HLM analyses greatly influenced the strengths of the resulted correlations. Moreover, the restricted scaling, whether for achievement scores or for the questionnaire responses, restricted variability. That restriction, in turn, hindered the detection of the variation in students' achievement in response to the variation of teachers' assessment practices. The previous explanations could be considered as speculations about the possible reasons that may lead to having such low correlations or insignificant relations reported in different studies (e.g. Hao, 2005; Hastedt, 2004; Jones 2004; Rodriguez, 1999).

Additional rigorous examinations of the aforementioned relationships in same, or in new datasets, are expected to enhance the validity of any anticipated findings. These should involve thorough documentation of data characteristics, treatment conditions and proper interpretation of findings. In general, the literature review serves as a basis to formulate initial hypotheses that will lead to the sound articulation and exploration of a number of potential relations between

teachers' assessment practices and students' achievement in mathematics, and more specifically when dealing with problem solving skills, considered to be at the forefront of recent educational reforms and the accountability movement. More discussion about these issues is provided in the next chapters.

CHAPTER III

Methodology

For the purpose of this study, the focus is on math teachers' assessment practices and their students' achievement scores as captured by SAIP 2001. This research answers the question: What are the relationships between teachers' assessment practices and students' achievement? The research orientation is quantitative in nature as it involved correlations. It explores various relationships between teachers' assessment practices and students' achievement in mathematics.

This chapter starts with a demonstration of the study's conceptual framework. Also in that section, different hypotheses are proposed to describe the anticipated relations between teachers' assessment practices and students' achievement. A few sections focus on different study conditions such as number of participants, sampling methods, as well as different instruments used in the SAIP 2001 mathematics achievement test. Finally, the chapter concludes with a description of the data analysis procedures used to uncover the aforementioned relations.

Framework

The conceptual framework builds on Brookhart's theory (Brookhart, 1997). Her theory combines classroom assessment environment literature, and social cognitive theories of learning and motivation, and "is amenable to empirical testing" (p. 161-162). According to Brookhart's theory:

> In any particular class, the classroom environment is played out in repeated classroom assessment events, activity segments with associated expectations and assessments. Within a classroom assessment event, the teacher communicates to students through assignments,

activities, and feedback on performance, and students respond according to their perception of these learning opportunities... The classroom assessment event concept offers a mechanism for how curriculum, instructional activities, and assessments impact student effort and achievement (p.161).

Brookhart regards classroom assessment environment as the context where teachers and students engage in a classroom assessment event. According to her theory, such an event describes "discrete sets of objectives and assessments of whether and to what degree the objectives are met." (p. 166). During the assessment process, the teacher's choices of assessment practices are expected to affect the learning environment and hence their students' achievement. In Brookharts' theory, these teachers' choices could be about the assessment tool, frequency of use, or intended feedback (Brookhart, 1997, p.165-166). For example, the weight that the teachers give to various assessment tools as they assess their students' work may have an effect on achievement as would frequency of use of certain assessment tools for a specific purpose.

Brookhart argued that her theory is expected to provide a plausible explanation to the relation between assessment and achievement. It is believed that the current study could evaluate the credibility of that provision. Theoretically, in a math classroom assessment event, teachers may engage with their students through assigning activities such as problem solving activities, and then assess their understanding using different tools (e.g. projects, or interviews). Teachers, then, use the collected assessment information for feedback purposes to improve students' understanding and hence their achievement. According to Broohart (1997), the aforementioned assessment event should offer a mechanism that could impacts students' achievement.

In the framework of this study, teachers' classroom assessment practices are divided into three dimensions: the types of assessment tools, the purpose of their use, and the purpose of

homework (see Figure 1). The first dimension, types of assessment tools, includes standardized test, teacher made multiple-choice tests, teacher made short-answer tests, projects, portfolios, observations and interviews. In the SAIP teachers' questionnaire, teachers were asked about the weight they assign to those assessment tools as they assess their students' math work. The second dimension includes the purposes for collecting assessment information such as feedback and diagnosis. In the teachers' questionnaire items pertaining to this dimension, teachers were asked about their frequency of use of assessment data for such purposes.

The third dimension is the purpose of homework, which includes use of homework for the purposes of feedback and grading. Similarly, regarding this dimension, teachers were asked about their frequency of use of homework for feedback or for grading purposes. In this study, it is believed that the choice of aforementioned dimensions was affected by the items from the SAIP's teacher background questionnaire. For instance, SAIP only questioned teachers about their frequency of use of assessment information for feedback or diagnostic purposes, but not for summative purposes. Therefore, using assessment information for summative purposes, although may be related to students' achievement, was not included in the framework for lack of related items and corresponding data.

It is the interest of the current study to see if there is a relationship between weight assigned to various assessment tools (e.g. teacher made multiple-choice, and teacher-made short answer tests) and students' mathematical achievement in SAIP. The interest also extends to investigate the relationship between the frequency of use of assessment information for different purposes (e.g. feedback, or diagnostic) and students' achievement. As well, the study examines the relationship between the frequency of use of homework for feedback or grading purposes and students' achievement. The literature provided several evidences to formulate preliminary

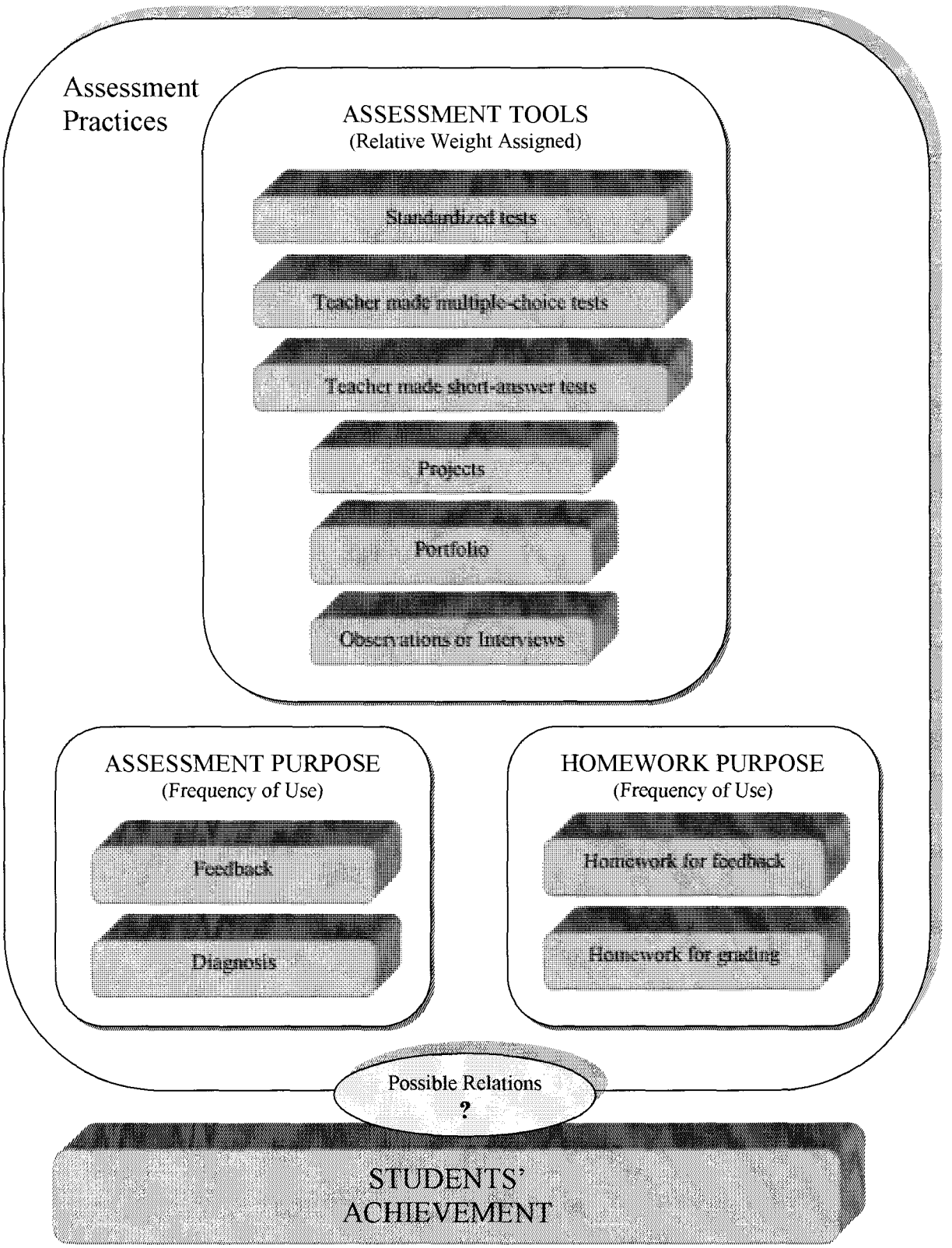

Figure 1. Relating teachers' assessment practices and students' achievement

hypotheses for some components of the proposed framework. Some of the assessment practices could be hypothesized to have a positive relation with students' achievement (e.g. giving weight to teacher made short-answer tests, or the frequency of use of assessment information for feedback). On the other hand, a few assessment practices could be hypothesized to have a negative relation with students' achievement (e.g. teacher made multiple-choice tests, or the frequency of use of homework for grading).

Sampling

The subjects in this study included students and their corresponding mathematics teachers who participated in SAIP 2001 mathematics assessment. The SAIP database included 13-years-old and 16-years-old students. According to CMEC (2003) documents, the total number of participating students was approximately 41,000 students: 24,000 13-years-old students and 17,000 16-years-old students.

Both the achievement test and questionnaires were administered to random samples of classes of students drawn from 18 different populations. These populations represented all of the provinces and territories in Canada and subpopulations in the two official languages. The sample sizes used by SAIP were proportional to the sizes of each population except for some subpopulations that needed to be oversampled. The sample size was decided to guarantee margins of errors of 3% or less at a confidence interval of 95% (CMEC, 2003).

Instrumentation

The mathematics assessment instrument of SAIP 2001was administered to collect information about the achievement levels of the 13- and 16-years-old students in mathematics

content and problem solving components. The mathematics content assessment component consisted of 125 questions. It was designed to evaluate students' achievement levels accomplished at different strands (e.g. numbers and operations, algebra and functions). Figure 2 illustrates question items from the mathematics content component (exemplars with responses were adopted from CMEC 2002).

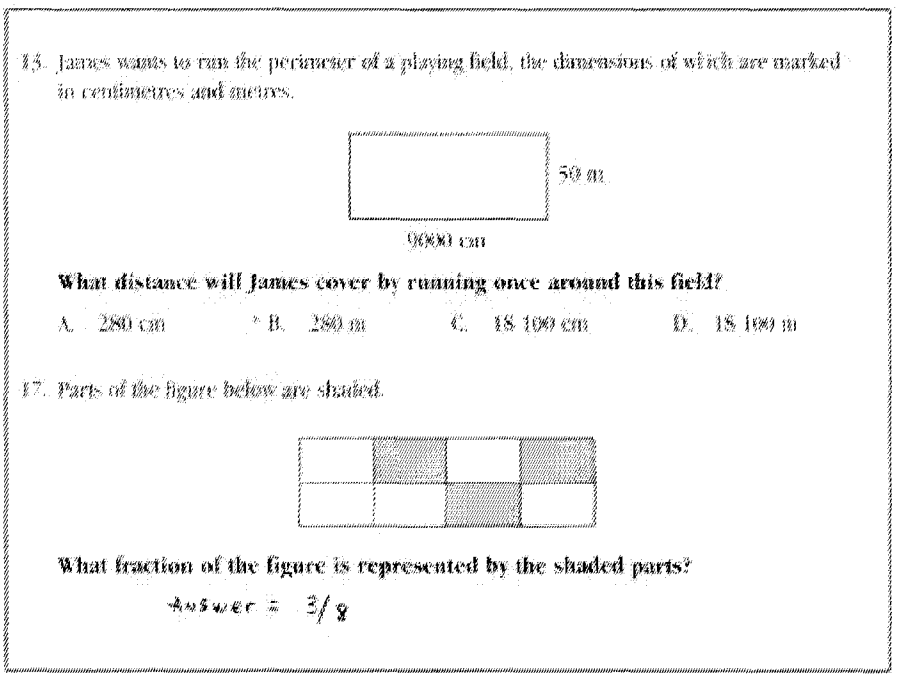

Figure 2. Items from the SAIP 2001 mathematics content component

On the other hand, the mathematics problem solving assessment component consisted of six problems, and each problem had sub problems within (e.g. A, B, C, D, and E). It was designed to evaluate students' achievement attained on other skills (e.g. the ability to reason and to construct proof, making inference, and use of communication skills). Figure 3 illustrates question items from the mathematics problem solving component (exemplars with responses were adopted from CMEC 2002).

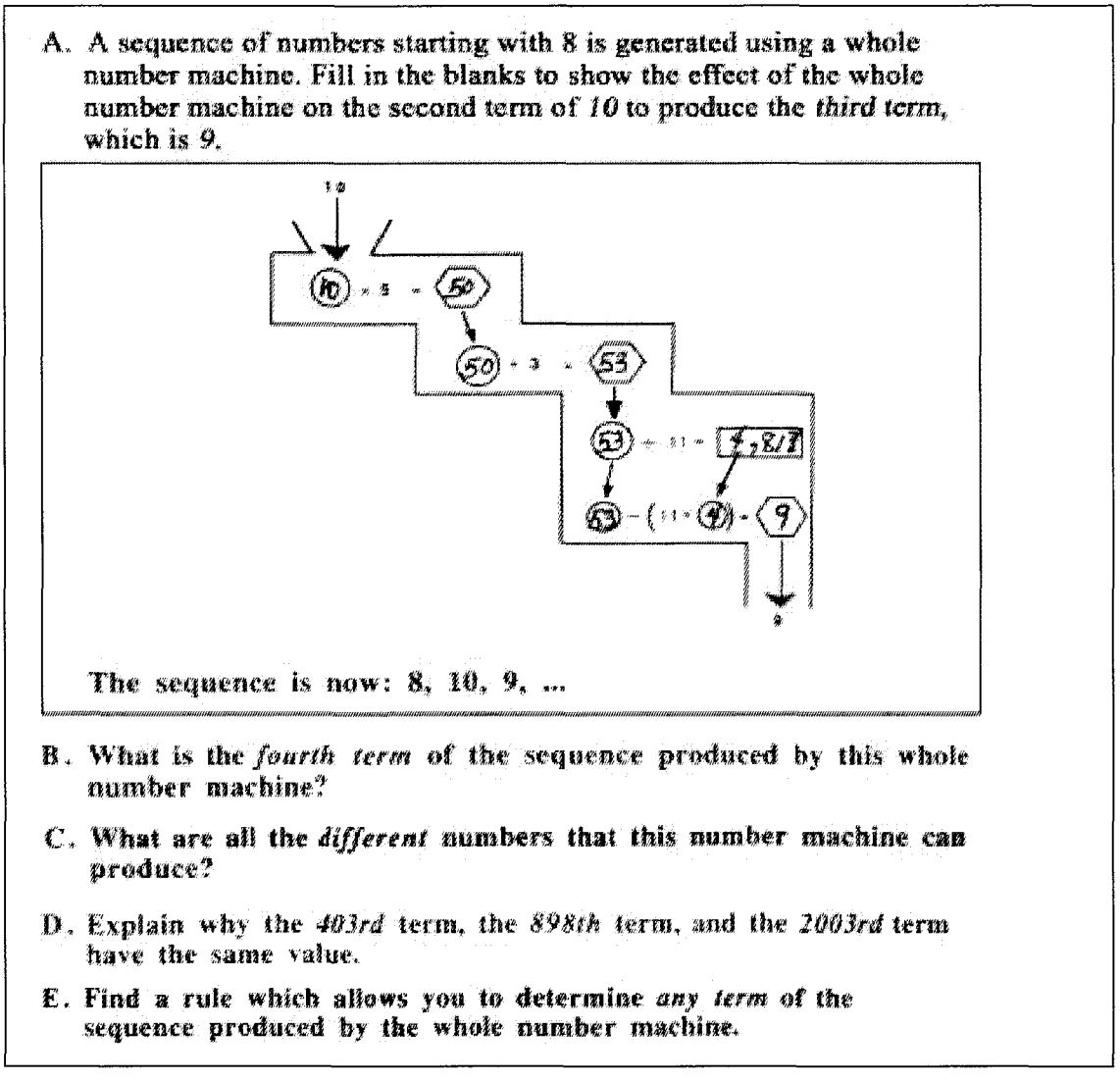

Figure 3. Items from the SAIP 2001 mathematics problem solving component

By investigating the nature of questions asked in the mathematics content component (e.g. item 16 and item 17), it was found that SAIP was examining students' knowledge of how to calculate dimensions, perimeters and areas of plane figures (CMEC, 2002). On the other hand, in questions asked in the mathematics problem solving component (e.g. questions A, B, C, and D) SAIP was examining students' ability to establish proof, make a choice of algorithms to find the solution to multi-step problems, and to use mathematical common vocabulary (CMEC, 2002).

As mentioned earlier, the current study is focused on the assessment practices that are related to improving students' understanding of mathematics and eventually improve achievement. The students' abilities investigated by SAIP in the problem solving component are closely related to the expected problem solving skills (e.g. understanding the problem, creating sub problems if needed and using data correctly) set, for example, by the Ontario Ministry of Education (OME, 2005) in the mathematics curriculum. Therefore, it is believed that focusing on the achievement of the mathematics problem solving component rather than the content component of SAIP 2001 could provide more appropriate results.

Student's achievement was described by an individual score of six levels: level-1 to level-5 with a possibility of zero. Level-1 reflects knowledge of an early stage of a typical elementary education. Level-5, on the other hand, describes the knowledge of a student, who has a wide range of mathematical knowledge typically expected at the end of secondary school education (CMEC, 2003). The mathematics assessment instrument described above was accompanied with questionnaires for schools, teachers, and students. In this research, the main interest is the teachers' questionnaire data.

The teachers' background questionnaire contained 29 items that covered the teachers' professional background and expertise, as well as their assessment practices. Each teacher was asked to choose the option that best reflected his or her frequency of use of the collected assessment for different purposes. In addition, teachers were asked about the relative weights they assign to various given assessment tools. Table 1 illustrates the questions found in the teachers' background questionnaire that pertain to assessment practices.

Four sets of variables were isolated in this study. For the first set of variables titled "assessment tools", teacher responses of interest included those related to items on relative

weights assigned to standardized tests, teacher made multiple-choice and short-answer tests, projects, portfolios, and observations or interviews (Items 23-A, 23-B, 23-C, 23-E, 23-F, and 23-G). The second set of variables included "assessment purpose" and grouped those items on frequency of feedback, and diagnosis (Items 14-H and 14-J). The third set of variables titled "homework purpose" included items on frequency of homework for classroom feedback, and for grading (Items 22-D and 22-G). The fourth set contained only one variable, which is the student's score in the problem solving component of the SAIP 2001 mathematics achievement test. This variable is considered as a response variable, and is expected to be associated with the explanatory variables from the other three sets of variables: "assessment tools", "assessment purpose", and "homework purpose".

All these variables are measured using ordinal scales. For example, the students' achievement scores is an ordinal variable with 6 levels: *Zero, level-1, level-2, level-3, level-4, and level-5.* Items for the "assessment tools" set of variables use the following scale: *a great deal, quite a lot, a little, and none*. This scale measures the weight assigned by teachers to each assessment tool. On the other hand, items for the "assessment purpose" set of variables use the following levels: *rarely or never, a few times a month, a few times a week, and almost every class*. This scale; however, measures the "frequency of use" of assessment information for a specific purpose. Regarding items for the "homework purpose" set of variables, the following levels are used: *rarely or never, a few times a month, a few times a week, and almost every class.* This scale also measures the "frequency of use" of homework for feedback and for grading.

Table 1

SAIP Background Questionnaire Items Related to Teachers' Assessment Practices

Questionnaire Items	Item #
How often do the following things happen in your mathematics classes?	
• *I give feedback to the class on assignments, tests or other evaluations*	*14-H*
• *I attempt to diagnose and address individual student problems or needs in learning*	*14-J*
(1-Rarely or never 4-Almost every class)	
If you assign written homework, how often do you do the following?	
• *Give feedback on homework to whole class*	*22-D*
• *Use homework to contribute towards students' grades or marks*	*22-G*
(1-Rarely or Never 4-Almost every class)	
In assessing the work of students in your mathematics courses, how much weight do you give each of the following?	
• *Standardized tests produced outside the school*	*23-A*
• *Teacher made short-answer or essay tests that require students to explain their reasoning*	*23-B*
• *Teacher made multiple-choice, true-false, or matching tests*	*23-C*
• *Projects*	*23-E*
• *Portfolios of student work*	*23-F*
• *Observations or interviews of students*	*23-G*
(1-None 4-A great deal)	
Total	10 items

Procedures

The first procedural step was getting permission to access the SAIP 2001 data from CMEC. Following that was preparing the data for analysis by a statistical analysis package. The intended statistical analysis package for this study was the Statistical Package for the Social Sciences (SPSS). Since the raw data were provided in a text format (_.dat), it was necessary to import it into the package and save it in SPSS format (_.sav). The previous process had to be done for both the students' achievement database and the teachers' responses database. The next step was to identify each teacher and his or her corresponding students' scores in the students' achievement database in order to link the two aforementioned databases. Before the linking process, databases were examined for missing data pertaining to both teachers and students.

Once the two databases were linked, the third step was to perform statistical analyses to investigate relationships among the variables of the study. It is worth mentioning that after processing the data to make it ready for correlational analysis, CMEC was contacted to get their feedback on the appropriateness of the methods used to deal with SAIP's data. The analytical procedures were confirmed to be proper, and as a result the final process was to investigate relationships among variables and then interpret and report the statistical results.

Data Analyses

According to the literature, the best missing data treatment for this study is employing the Expectation Maximization (EM) estimation procedures for imputation (Allison, 2002; Anderson et al, 2006; Croy & Novins, 2005). In addition, inferential data analysis methods were applied to investigate the relations among the variables of this study.

Given the 4-point Likert type rating scale used in teachers' questionnaires as well as the 6 achievement levels used to describe students' scores, according to the literature, the suitable analytical model to use was the Kendall's τ_b statistic (CMEC, 2003; Kendall, 1962; Goodman & Kruskal, 1954, 1979). To limit the likelihood of making *Type-I* error, an α level of 0.05 was used. Deciding upon the direction of association, whether it is positive or negative, was based on calculating the number of Concordant (C) and Discordant (D) pairs within the final linked database.

In case of positive association, the number of Concordant pairs are expected to be more than the Discordant pairs ($C > D$). Conversely, under negative association, the number of Concordant pairs are expected to be less than the Discordant pairs ($C < D$), but if $C \approx D$ then there is no association. The value of the τ_b coefficient ranges between -1.0 for a strict negative association, and 1.0 for a strict positive association. The statistical package, SPSS, was used to investigate the associations among variables.

While calculating the Kendall's τ_b correlations, SPSS provides a test of significance. The test provides confidence that the results at hand are unlikely to have occured by chance. However, statistical significance can not tell much about the results if used as a stand alone measure. As well, Maxwell and Delaney (2004) argued that as the sample size gets larger it increases the chances of having a significant test. In this study, for example, despite the data loss that took place during linking the databases, the remaining records (296 teachers and 3,696 students) could still lead most of the findings to be statistically significance. Thus, a search for a better interpretation method of the findings other than the test of significance was warranted.

Cohen (1988) provided a benchmark to help interpret the size of correlations obtained in social sciences research. He considered a correlation coefficient to be small if it ranges from 0.1

to 0.3, or from -0.3 to -0.1 for negative values. Similarly, he labeled values from 0.3 to 0.5 to be of medium size, while values from 0.5 to 1.0 to be large. However, Cohen warned researchers in social sciences from generalizing his benchmark or using it so strictly, because it should be interpreted differently according to the context of each research. For example, he observed that in studies where a good experimental control is present, such as in economics or in experimental psychology, the correlations are likely to be large (e.g. ranging from 0.5 to 1). On the other hand, he noticed that in other studies, related to education for example, the correlations are likely to be small (e.g. ranging from 0.1 to 0.3). Therefore, research findings should be interpreted within their own context and compared to other similar studies to avoid misleading the general audience.

The current study intends to use Cohen's benchmark for interpreting the findings. In that sense, this study's correlations that are between Zero and 0.1 imply a very weak or a nonexistent relation between the investigated variables. Findings that range from 0.1 to 0.3 or from -0.3 to -0.1 will be considered to have a small correlation, which implies the existence of a small relation. The findings will be explained using this rule of thumb. Nonetheless, the interpretation will occur within the educational field context, and relative to the findings of other studies that focused on relating teachers' assessment practices to students' achievement (e.g. Dudaite, 2006; Hao, 2005; Jones, 2004; Rodriguez, 1999).

CHAPTER IV

Results

Before introducing the results pertaining to the research question, a brief description of students' achievement on the SAIP 2001 mathematics tests and teacher assessment practices as reported on the SAIP 2001 background questionnaire is presented. All the students' scores reported in this study pertain to the problem solving component of the test. Throughout the chapter, the proportion of missing data is also presented to give an idea about the data loss during the analyses. The chapter also offers results in relation to the process of linking the students' achievement and the teachers' questionnaire databases. The main section of the chapter; however, presents the results of the correlations between teachers' assessment practices and their students' achievement.

Characteristics of the data bases

The missing data level was first investigated within the SAIP 2001 students' achievement scores dataset. A total of 19,266 students completed the problem solving part of the SAIP 2001 mathematics achievement test. Out of these students, 879 students had no score, resulting in a data loss ratio of 4.6%. An additional 2118 students had no corresponding teacher identification code. Therefore, those students, representing 11% of the database, had to be deleted from the students' achievement database before linking it with the teachers' questionnaire database. Missing data thus formed 15.6% of the students' achievement scores database.

As mentioned earlier in the methodology section, a set of 10 items on teachers' assessment practices were selected from the SAIP 2001 mathematics teachers' questionnaire to

represent teachers' assessment practices (See Table 2). Missing data per item ranged from 1.3% to 4.2%, which is acceptable and is believed to have minimal effect on the validity of the analyses according to Anderson et al (2006).

Table 2.

Percentage of Missing Data Level per Item within the Original Teacher Questionnaire Database

Item number	Question Item (paraphrased)	Missing Data % (N=5362)
23-A	Standardized tests produced outside the school	3.0
23-B	Teacher-made short-answer or essay tests that require students to explain their reasoning	2.6
23-C	Teacher-made multiple-choice, true-false, or matching tests	2.3
23-E	Projects	2.8
23-F	Portfolios of student work	3.3
23-G	Observations or interviews of students	3.0
14-H	Giving feedback to the class on assignments, tests or other evaluations.	1.3
14-J	Attempting to diagnose and address individual student problems or needs in learning	1.7
22-D	Giving feedback on homework to whole class	4.2
22-G	Using homework to contribute towards students' grades or marks	3.9

The teachers' questionnaire and the students' achievement databases were then linked to form a new database: practices-achievement database. The linked database included teachers, their assessment practices, and their students' performance. The linking process was based on using teacher ID as a common key to link teacher's assessment practices and their corresponding

classroom students' achievement. As a result of the data merge, a few data were rendered missing, when the teacher ID's did not match in both databases. Eventually, the practices-achievement database included approximately 85% of the teachers' records, which were originally found in the teachers' questionnaire database. In addition, the practices-achievement database included approximately 84% of the students' achievement scores, which were originally included in the students' achievement database.

As students were grouped by their classroom teachers in the practices-achievement database, it was observed that the number of students per class varied from 1 to 34 students. It was further noticed that a large number of classes had between 1 and 10 students. For instance, a large number of classes had a total number of 3 students. The achievements of those students were all above level-3, and as a result the variability within these classes was restricted. That eventually may reduce the chances of detecting differences in students' achievement. Besides, the original class size, although not provided by SAIP, is expected to be larger than 3. So, not only the variability is restricted, but also the validity of the overall class performance could be highly questionable. It is believed that the aforementioned problems related to the inclusion of these classes (size $n<10$) could impact the chances of detecting differences in achievement, and also could render the calculated correlations to be useless.

Based on the previous argument, it was decided to look at a subset of the achievement-practices database that could have more variability in hopes of discovering some potential relations between teachers' assessment practices and students' achievement. In this study, the class size limit was set to 10 students per class to have a comparable number of students among classes. On applying that criterion, the final practices-achievement database had 296 math classroom teachers (their classes represented 7% of the total number of classes) and 3,696

students, which represented 22% of the students in the initial achievement databases (See Figure 4). Although the practices-achievement database was largely reduced, the variability in students' achievement may have been improved, which in turn could lead to the detection of some meaningful relations. The effect of the exclusion of classes that have less than 10 students on the findings of this study will be presented in later sections.

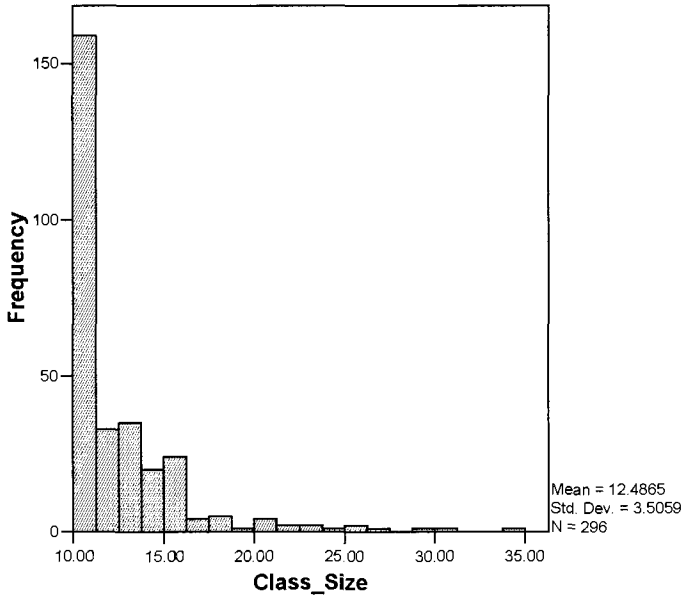

Figure 4. Class-size distribution in the practices-achievement database after applying the criterion

After merging the two databases and having that volume of data deletion it was then necessary to examine the impact of applying the criterion on the datasets. Three approaches were taken. First, the data distribution of students' achievement before and after the application of the class-size limit criterion was examined. Second, teachers' assessment practices distributions were also examined before and after applying the same criterion and deleting the cases. Third, the deletion of students' achievement for classes that had 10 students or less was investigated to

see the impact it had on the study's findings. In that particular analysis, all the achievements of the classes having one student to forty-three students, rather than those classes with size 10 and above only, were merged with the teachers' assessment practices. Then, possible relations between teachers' assessment practices and students' achievement were examined. Those three approaches are reported in the next three sections after the presentation of the applicability of a missing data imputation technique.

Having addressed the abovementioned inevitable data deletion, it was imperative to shed some light on the impact of applying a missing data imputation technique such as the Expectation Maximization method on the current analyses. This verification was essential before investigating the correlations between teachers' assessment practices and students' achievement. As mentioned earlier, the original students' achievement database had 879 students (4.6% of the total number of students) missing their achievement data. Out of the 879 students' records, only 395 students' records (45% of the total missing data) had their corresponding teacher identification codes. On the other hand, 484 students' records (55% of the missing data) had no corresponding teacher identification codes. Without these codes it is impossible to link the teacher questionnaire and students' achievement databases. Therefore, those 484 records had to be excluded from imputation because they cannot be used in the databases' merging process.

The only missing achievement data that could be imputed were those 395 students records with their corresponding teacher identification codes. By conducting further investigation on those 395 students' records, it was found that they include students from 266 classes (5% of the total number of classes). Table 3 represents the class size distribution of those classes. In that distribution, the number of students per class ranged from 1 to 9 students. However, the aforementioned inclusion criterion states that the smallest class size has to be

greater than or equal to 10 students. This means that even if the 395 students' achievement records were imputed, they were going to be excluded during the analyses, because their class sizes are less than 10. Therefore, applying the EM imputation technique to the records that have missing data proved insignificant, because of the missing teachers' identification codes and the small class size problems.

Table 3

Distribution of Class Size Found in the 266 Classes Remaining for Imputation

Class_Size (Number of Students per Class)	Frequency	Percent
1	196	73.7
2	43	16.2
3	15	5.6
4	4	1.5
5	3	1.1
6	2	0.8
8	2	0.8
9	1	0.4
Total	266	100
Mean	1.4850	
Standard Deviation	1.12676	

Description of Students' Achievement Dataset

As mentioned earlier, students' achievement had to be grouped by their teachers before linking the students' achievement and the teachers' questionnaire databases. This step deemed necessary in order to investigate the relation between teachers' assessment practices and their students' achievement. It was noticed that the variability within the students' performance improved after the linking process. Before the linking process, the primary investigation of the original students' achievement scores database showed that approximately 60% of the students had performance level-2 and level-3. On the other hand, around 30% of the students had performance level-1 and below (See Figure 5). Having most of the students' scores within these adjacent score levels constrained variability within students' scores (S^2=1.34).

In addition to the limited variability problem, having an ordinal scale to measure students' performance posed a challenge in the data aggregation process. The challenge had to do with aggregating the ordinal data scores within each classroom to form a single score. In this study, classroom achievement was aggregated by computing the percentage of students scoring at or above Level-3 at the SAIP 2001 mathematical problem solving component (CMEC, 2003). Thus, in the new linked practices-achievement database, the students' achievement scale was changed from an ordinal to a continuous scale ranging from 0 to 100. The analyses showed that having this new scale changed the variability within students' performance from 1.34 to 107.58.

As mentioned earlier, in the full data analyses, there were 296 classrooms. Using the new continuous scale, the average overall class performance, in the problem solving component, was 6.78 with a standard deviation of 10.37. The minimum overall classroom score was Zero, while the maximum was 46.15 with a positive skewness coefficient of 1.624. Having a Zero overall classroom score indicated that no students in that particular math classroom scored at or above

level-3 at the SAIP test. On the other hand, having a maximum score of 46.15 showed that almost half of the students in that particular classroom scored at or above level-3.

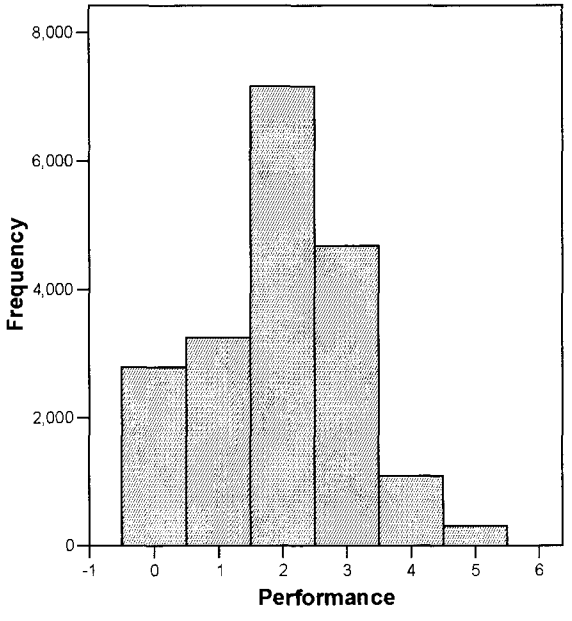

Figure 5. Distribution of students' performance in the achievement database

Description of Teachers' Assessment Practices Dataset

Table 4 describes the *assessment tools* that teachers used within their mathematics classroom. The data showed that over half of the responding teachers did not assign any weight to standardized tests produced outside the school or to portfolios of students' work. On the other hand, approximately half of the participating teachers assigned a little weight to projects, teacher made multiple-choice, true-false, and matching tests. Likewise, half of the respondents gave weight to assessment tools such as teacher made short-answer or essay tests.

Table 4.

Percentage of Teachers Giving Weight to Different Assessment Tools in the Practices-Achievement Dataset

Assessment Tools	A great deal %	Quite a lot %	A little %	None %
Giving weight to standardized tests produced outside the school	1.7	13.1	23.0	54.3
Giving weight to teacher made short-answer or essay tests that require students to explain their reasoning	10.0	39.2	36.4	7.6
Giving weight to teacher made multiple-choice, true-false, or matching tests	1.4	17.5	50.9	23.4
Giving weight to projects	1.4	14.1	43.0	34.0
Giving weight to portfolios of student work	1.0	8.2	27.8	54.6
Giving weight to observations or interviews of students	1.7	14.1	41.2	34.4

For the *assessment purpose* dimension (See Table 5), teachers' attitudes towards a few purposes appear to be similar. For example, approximately one third of the teachers use assessment information for feedback and diagnostic purposes a few times a month. In addition, almost 40% of the teachers use assessment information a few times a week for the same purpose.

Table 5.

Percentage of Teachers Using Assessment Information for Various Purposes in the Practices-Achievement Dataset

Assessment Purposes	Almost every class %	A few times a week %	A few times a month %	Rarely or never %
Giving feedback to the class on assignments, tests or other evaluations	27.1	36.8	27.8	1.0
Attempting to diagnose and address individual student problems or needs in learning	15.1	37.1	27.1	12.7

For the *homework purpose* dimension (Table 6), the data showed that 50% of the teachers use homework assessment information a little to give feedback to the whole class. On the other hand, 40% of the teachers were equally divided between using homework assessment information a great deal and using it quite a lot for the purpose of grading. It was also observed that among all assessment purposes, 20% of the teachers refrain from using assessment information for the purpose of grading.

Table 6.

Percentage of Teachers Using Homework for Various Assessment Purposes in the Practices-Achievement Dataset

Homework purposes	A great Deal %	Quit a lot %	A little %	None %
Using homework to contribute towards students' grades or marks	21.3	21.3	26.8	21.3
Giving feedback on homework to whole Class	7.2	28.2	49.5	8.2

Given the amount of missing data in the linked practices-achievement database, it was imperative to cross validate teachers' assessment practices before and after the linking process. This step was expected to examine the integrity of the information presented in the linked data base prior to proceeding with secondary analyses. In other words, this step ensures that the patterns observed within teachers' assessment practices are comparable before and after the linking process despite the volume of data deletion. Regarding the *assessment tools* dimension, the percentages of teachers who did not give weights to standardized tests produced outside the school were almost identical before and after the linking process (see Tables 4 and 7). Similarly, the percentages of teachers, who give a little weight to teacher-made multiple–choice, true-false, or matching tests, and observations/interviews were nearly identical.

Table 7.

Percentage of Teachers Giving Weight to the Various Assessment Tools in the Original Dataset

Weights assigned to assessment tools	A great deal %	Quite a lot %	A little %	None %
Standardized tests produced outside the school	2.5	13.3	30.5	53.7
Teacher-made short-answer or essay tests that require students to explain their reasoning	12.7	40.0	37.4	10.0
Teacher-made multiple-choice, true-false, or matching tests	3.6	20.6	50.3	25.6
Projects	1.3	13.4	49.2	36.1
Portfolios of student work	1.9	9.5	28.1	60.5
Observations or interviews of students	2.5	14.7	40.3	42.5

On the other hand, regarding the *assessment purpose* dimension, there were a few noticeable changes in the frequency of use of assessment information for diagnostic purposes (see Tables 5 and 8) as well as the frequency of use of homework for feedback purposes (see Tables 6 and 9). For example, the frequency of use of assessment information almost every class for diagnostic purposes changed from 35% to 15%. In addition, the frequency of use of homework a great deal for feedback purposes changed from 45% to 7%. The impact of these differences on the results of this study will be discussed in the next chapter. In general, except for these two abovementioned practices, the trends in the two assessment practices databases were almost similar.

Table 8.

Percentage of Teachers' Frequency of Use of Assessment Tools for Various Purposes in the Original Dataset

Assessment Purposes	Almost every class %	A few times a week %	A few times a month %	Rarely or never %
Giving feedback to the class on assignments, tests or other evaluations	31.6	39.6	28.3	0.6
Attempting to diagnose and address individual student problems or needs in learning	34.6	36.7	25.7	3.0

Table 9.

Percentage of Teachers' Frequency of Use of Homework for Various Assessment Purposes in the Original Dataset

Homework purpose	A great Deal %	Quit a Lot %	A little %	None %
Using homework to contribute towards students' grades or marks	22.4	23.7	33.8	20.1
Giving feedback on homework to whole class	45.7	31.9	18.4	4.1

Examining Correlations between Teachers' Assessment Practices and Students' Achievement

To evaluate the effect of teachers' use of the aforementioned assessment practices on students' achievement, correlations were computed (see Table 10). The values of these correlations ranged from r=-0.241 to r= 0.204, and they were all statistically significant. As mentioned earlier, the assessment practices had three dimensions: *assessment tools*, *assessment purposes*, and *homework purposes*. Regarding the first dimension, a small positive correlation was found between assigning weight to projects and students' achievement in the SAIP 2001 mathematics large scale assessment test (r=0.204). This means that an increase in giving weight to projects could be related to an increase in students' mathematical achievement in problem solving.

On the other hand, it was found that giving weight to portfolios and observations/interviews had small but negative correlations with students' mathematical achievement (r=-0.136, and r=-0.241 respectively). These correlations imply that the increase in giving weight to portfolios and observation/interviews could be related to a decrease in students' mathematical achievement. However, the rest of the assessment tools in this dimension (e.g. giving weight to teacher-made short-answer tests) had very weak relations with students' achievement.

By examining the second dimension, assessment purposes, it was found that the frequency of use of assessment information for the purpose of feedback had a small positive correlation with students' achievement (r=0.119). This suggests that the reported increase in the frequency of use of assessment information for feedback purposes could be related to an increase in students' mathematical achievement. This finding supports the preliminary hypotheses assumed in this study, as well as the research covered in the literature. However, the use of

Table 10.

Correlations between Teachers' Assessment Practices and Students' Achievement on the SAIP 2001 Mathematics Achievement Test (Problem Solving Component)

	Assessment Practices	Correlations With Students' Achievement*
Assessment Tools	Standardized Tests	-0.029
	Teacher-Made Short answer test	-0.074
	Teacher-Made Multiple Choice / True False test	0.094
	Projects	0.204
	Portfolios	-0.136
	Observation / Interviews	-0.241
Assessment Purposes	Assessment for Feedback purpose	0.119
	Assessment for Diagnostic purpose	-0.130
Homework Purposes	Homework for Grading purpose	0.067
	Homework for Feedback purpose	-0.024

*Correlations were calculated using the Kendal Tau-b statistic at the α=0.05 level & for Class_Size>=10

assessment information for diagnostic purposes was negatively correlated to students' achievement (r=-0.130). This means that, an increase in the use of assessment information for diagnostic purposes could be related to a decrease in students' mathematical achievement. This relationship is nevertheless considered to be small according to Cohen's rule of thumb.

Finally, the investigation of the third dimension, homework purposes, yielded very weak but statistically significant correlations. For instance, the frequency of use of homework for feedback purposes had a very weak negative correlation r=-0.024 with students' achievement,

whereas using it for grading had a very weak positive correlation r=0.067. Therefore, in this study, it is believed that the frequency of use of homework for feedback or grading purposes had very weak relations with students' achievement. These findings disagree with the preliminary hypotheses assumed in this study. More interpretations about the aforementioned correlations are discussed in the next chapter.

Next, the impact of deleting achievement of classes that have less than 10 students on this study's results was investigated. The records used in this analysis included achievement of all classes with sizes ranging from 1 to 34 students. Kendall's τ_b coefficients were recalculated for all the teachers' assessment practices that revealed small relations with students' achievement (e.g. giving weight to projects and portfolios). It was found that all of the correlations calculated using the achievement of all the classes (size 1 to 34) have lower values than correlations reported earlier for classes with more than 10 students (see Table 11). For example, the relation between giving weight to projects and students' achievement changed from a small positive relation to a very weak positive relation. However, the relation pertaining to giving weight to observation/interviews remained as a small negative relation. More about the effect of excluding classes with size $n<10$ is coming in the next chapter.

In summary, the findings of the current study proposed a few relations between teachers' assessment practices and students' achievement in the problem solving component of SAIP 2001 mathematics large-scale assessment test. Some positive relations were uncovered between giving weight to projects, as well as, the frequency of use of assessment information for feedback purposes and students' achievement. On the other hand, giving weight to portfolios and the frequency of use of assessment information for diagnostic purposes were found to be negatively related to achievement. As mentioned previously, the strengths of these positive and negative

relations had values that varied from r=-0.241 to r= 0.204. The reasons behind such variation as well as possible interpretations of the findings will be discussed in the next chapter.

Table 11.

Comparison of Significant Correlation Results of the Two Data Files (n>=10 / All)

Assessment Practices	Correlations With Students' Achievement n>=10* / All**
Projects	0.204 / 0.035
Portfolios	-0.136 / -0.093
Observation / Interviews	-0.241 / -0.105
Assessment for Feedback purpose	0.119 / -0.048
Assessment for Diagnostic purpose	-0.130 / -0.034

*Correlations were calculated using classes with Class_Size >=10

** Correlations were calculated using all classes (Class_Size ranging from 1 to 34)

CHAPTER V

Discussion

The purpose of this chapter is to interpret and evaluate the findings of this study. It includes a review of the hypotheses and results while discussing the findings in the context of the previously covered literature. In addition, the limitations of the study and its implications on drawing firm conclusions are also discussed. The chapter wraps up with a discussion about the possible contributions of the current study as well as suggestions for future research.

The analyses of this study revealed that most of the investigated teachers' assessment practices were somewhat related to the students' achievement in the problem solving component of the SAIP 2001 mathematics large-scale assessment test. The correlations varied in their strengths and directions. The values of these correlations ranged from r=0.204 (relating giving weight to projects to students' achievement) to r=-0.241 (relating giving weight to observation/interviews to students' achievement). However, interpreting these findings has to be dealt with carefully to avoid reaching wrong conclusions. As mentioned earlier at the methodology section, Cohen's (1988) benchmark was used as a rule of thumb to interpret the current study's findings.

Findings Related to Assessment Tools

As mentioned earlier, giving weight to projects had a small but positive relation with students' achievement. Giving weight to portfolios, and observation/interviews also had a small but negative relation with students' achievement. However, findings pertaining to giving weight

to standardized tests, teacher-made short answer and multiple-choice/true-false tests had very weak or almost nonexistent relations with students' achievement.

Compared to the findings found in the literature (e.g. Rodriguez, 1999), the current study's correlation coefficient pertaining to projects proposed a relatively better relation. For example, the correlation found by Rodriguez (1999) pertaining to the use of projects was r=-0.01 (almost nonexistent relation), as compared to 0.204 in the current study (small relation). Therefore, the current study proposes a relatively better relation, although small, between weight given to projects and students' achievement. However, Rodriguez focused on relating teachers' assessment practices to students' mathematics achievement in general, and not particularly on problem solving. As well, the volume of missing data at the different level of the HLM model employed in Rodriguez's analyses may have reduced the strengths of his correlations.

In this study, on the other hand, the choice of focusing on the problem solving component rather than the general math achievement may have led to the discovery of a relatively stronger relation. As mentioned earlier, the skills assessed in the problem solving component of the SAIP 2001 mathematics achievement test were more likely to provide evidence about students' understanding than about the problem content component. It is believed that focusing on the problem solving achievement in this research may have raised the chances of associating an increase in giving weight to projects with a corresponding increase in students' achievement in problem solving. However, despite having a relatively stronger relation than Rodriguez finding, the strength of that this study's relation is considered small according to Cohen's benchmark.

In general, using projects as an assessment tool could improve teachers' confidence about their students' mathematical understanding. As students engage in a math projects and practice their problem solving skills (e.g. understanding the problem, create sub problems if needed and

use data correctly), teachers observe, and give feedback accordingly. Using projects to improve problem solving skills facilitates different opportunities for teachers to assess concept understanding, students' ability to apply concepts and procedures, and eventually communicate these ideas. Therefore, teachers will be more confident that students' correct responses are based on proper concept understanding. This confidence will allow teachers to teach different problem solving strategies such as using logical reasoning, or guessing and checking. By doing that teachers may improve students' understanding and eventually may improve achievement (Blumenfeld et al, 1991; OME, 2005).

In addition to the current study's findings discussed so far, the study could contribute to the literature on large-scale assessment by adding new empirical relations. For instance, the analyses revealed that giving weight to portfolios and observation/interviews had small negative relations with students' achievement in SAIP 2001 problem solving items. It is believed that assessment practices such as portfolios and observation/interviews require much time, which is a rare commodity especially when class size increases. As a result, the weight given to such practices during class time may be reduced.

The previous argument was checked in the practices achievement database, where class size ranged from 10 to 34 students. Focusing on the teachers' responses distribution, it was noticed that almost 35% of the teachers gave no weight to observation/interviews, while 41% gave little weight (recall Table 4). Those teachers' trends towards observation/interviews suggest that it is unlikely for teachers to perform that practice as the class size gets larger. One speculation of the reason behind having such a negative relation could be that students had high achievement in those classes where teachers gave little or no weight to observation/interviews.

Following up on the previous argument, it was further noticed that deleting students' achievement of classes with size $n<10$ students resulted in having a correlation of $r=-0.241$, while including these classes yielded a correlation of $r=-0.105$. According to Cohen's benchmark, both correlations are considered small; however, the correlation in the case of excluding classes with size $n<10$ is slightly larger, in absolute value, than the correlation in the case of their inclusion. It is believed that the increase in value may support the previous argument that the exclusion of those classes could have improved the variability within students' achievement. That improvement, in turn, may have led to the detection of a few more differences in achievement, than in the case of inclusion, and eventually may have increased the value of the proposed correlations.

The previous findings discussed in this section covered the relation between giving weight to different assessment tools such as projects and portfolios and students' achievement in mathematics large-scale assessment tests. As mentioned earlier, the use of assessment information for different purposes also affected students' achievement. The following section discusses the use of assessment information for feedback and diagnostic purposes and its relation to students' achievement in mathematics. As well, it discusses the use of homework assessment information for feedback and grading purposes and its relation to students' achievement.

Findings Related to Assessment Purpose and Homework Purpose

This study's analyses revealed that the frequency of use of assessment information for feedback purposes had a small positive relation with student achievement, while using assessment information for diagnostic purposes had a small negative relation. However, using homework for feedback or grading purposes had very weak relations with students' achievement.

The literature showed that the purpose for using assessment information impacts students' achievement (e.g. Black & Wiliam, 1998; Chappuis & Stiggins, 2002; Rodriguez, 1999). It is worth mentioning that in this study and in most of the theoretical articles and empirical studies found in the literature (e.g. Black & Wiliam, 1998; Brookhart, 2008; McMillan, 2003; Rodriguez, 1999) there was an agreement that using assessment information for feedback purposes positively impacts students' achievement. Even at the classroom level, the teachers' responses to the teachers' questionnaire reflected almost a consensus among teachers on using assessment information for feedback purposes. In addition, the corresponding correlation coefficient value in this study, despite having a small value, proposed relatively stronger relation than the findings proposed by other empirical studies such as McMillan (2003) and Rodriguez (1999).

Besides using assessment information for feedback purposes, teachers also used assessment information for diagnostic purposes. This study revealed a negative but small relation between using assessment information for diagnostic purposes and students' achievement. That relation, however, contradicted assertions found in the literature (e.g. Chappuis & Stiggins, 2002; Ciofalo & Wylie, 2006) which proposed a positive relation.

It was found that the missing data led to a noticeable change in teacher frequency of use of assessment information for diagnostic purposes before and after the linking process. For example, the percentage of teachers who used that assessment almost every class decreases from 35% to 15%. In addition, the percentage of teachers who rarely or never use it increased from 3% to 13%. It is believed that the change in distribution and the amount of data deletion that took place during linking the databases may have contributed to the aforementioned negative relation ($r=-0.13$). It is possible that the missing data could have resulted in associating lower frequencies

of use with higher achievement. This speculation is consistent with findings by Rodriguez (1999) who reported the association of lower frequencies of use of assessment information for diagnostic purposes with students' achievement in mathematics.

Nevertheless, this negative correlation (r=-0.13) seem to be counterintuitive, because using assessment information for diagnostic purposes at the beginning of a teaching cycle is anticipated to improve teachers' understanding of students' capabilities, resulting, in principle, in better teaching, which in turn would yield higher achievement (Chappuis & Stiggins, 2002; Ciofalo & Wylie, 2006). Chappuis et al assert that teachers should use assessment information, when they "pretest before a unit of a study and adjust instruction for individuals or the entire group", (p. 40).

Particularly in problem solving, Ciofalo and Wylie (2006) urge teachers to use diagnostic assessment information to gather information about students' prior knowledge to positively impact students' learning (p. 1). They as well assert that ongoing diagnostic assessment should be utilized to elicit students' misconceptions before they engage in problem solving activities. Typically, having "ongoing" evidence entails having a high frequency of use of assessment information for diagnostic purposes to positively impact students' understanding, and hence their achievement. Therefore, the frequency of use of assessment information for diagnostic purposes could be positively related to students' achievement. The contradiction pertaining to this finding; however, calls for more investigation to justify that relation.

Limitations

This study investigated the possible relations between math teacher classroom assessment practices and students' achievement in the problem solving component of the SAIP 2001 large-

scale assessment test. Although the current study proposed a few insightful findings while exploring the aforementioned potential relations, there were a few limitations that restricted the generalization of the findings. First, the current study is non experimental and that makes it impossible to manipulate the conditions or the context of SAIP. Therefore, the assumption of causality is not present and the findings only describe potential relations between teachers' assessment practices and students' achievements. Second, the findings at hand apply within the SAIP 2001 context only and mainly to the mathematics problem solving component of that large-scale assessment test. In fact, given the large volume of data deletion that took place during the analyses process, it is hard to claim that the current findings generalize to the whole SAIP population.

The third limitation pertains to the fact that the current study focused on students of two age groups: 13- and 16-years-old. Having these two age groups assessed together may pose a few interesting questions. It would have been helpful to know, although it is impossible at this point, whether the variation in students' performance is due to the variation in assessment practices only, or due to other confounding factors? In other words, was there a chance that the 16 years old students may have developed better problem solving skills more than the 13 years old? Did this contribute to the variation in achievement, and if so, how? Investigating that possibility was not feasible in the current study because of the secondary nature of the analyses. Therefore, this study's findings may not be generalized to other age groups.

The SAIP sampling procedures also limited the generalization of the aforementioned findings to the population. According to SAIP, students were sampled from 18 different populations representing all of the provinces and territories in Canada. Teachers, on the other hand, were not sampled. Instead, they were derived from the students' sampling scheme (CMEC,

2003). The teacher questionnaire responses database included all the teachers who taught the students who participated in the SAIP 2001 mathematics achievement test. According to CMEC (2003), that sampling method resulted in a difficulty in determining if all possible teachers have been identified. Sampling students but not teachers is believed to have limited the generalizations of the relationships to the population. That limitation questions how representative were the data pertaining to teachers' assessment practices. However, again, the SAIP context and sampling were not under the control of this study as they were already administered by CMEC.

In addition to the generalizability limitation, the missing data represented a challenge in exploring possible relations between teachers' assessment practices and students' achievement and posed the greatest threat to the credibility or trustworthiness of the results. Although all the findings of the study were statistically significant, the explored correlations ranged from -0.241 to 0.204. As mentioned earlier, there was a considerable amount of deletion during linking the teachers' responses database and the students' achievement database. It is possible that the type of practices that the missing teachers would likely be applying in their classes may have changed the strength of the relations.

Lastly, as issues of reliability and validity of questionnaires are considered crucial to ensure the credibility of the results, it was vital to shed some light on the SAIP 2001 teachers' questionnaire. Cizek (2009) argued that reliability of an instrument (e.g. teachers' questionnaire) is satisfied if the instrument yields same, or better, results (e.g. responses) if repeated under analogous conditions (p. 65). On the other hand, Colton and Covert (2007) asserted that instrument validity is established when the instrument measures what it intended to measure and produces results that allow intended inferences (p. 65). In their opinion, this is achievable when an instrument is pre-tested and evidences about validity are collected. On that last topic, CMEC

reported that it took care of reliability and validity issues pertaining to its teachers' questionnaire design (CMEC, 2003. p. 94). According to their report, the first draft of the teachers' questionnaire was pre-tested by approximately 20 teachers. That produced a second draft, which was reviewed by the developmental consortium, and then submitted to teachers through SAIP coordinators in each jurisdiction. In addition to pre-testing, CMEC sought the Canadian Teachers' Federation supportive feedback about the SAIP math and science teachers' questionnaires. That feedback, in turn, resulted in additional modifications to the questionnaires to assure validity (CMEC, 2003).

Overall, CMEC made considerable efforts to remove ambiguity that could be encountered when responding to the math teachers' questionnaire. However, in this study, it is believed that altering the wording of some question items in the math teachers' questionnaire could have made it even clearer. For example, teachers were asked about the "weight" given to different assessment tools in their assessment. It was not clear if the required weight should be relative or not (e.g. to weights given to other assessment tools). Perhaps, asking about "the frequency of use" of these assessment tools would have been better. As well, it is believed that the number of teachers involved in the questionnaire pre-testing should have been larger than 20 teachers. Given that for the students' questionnaire the pre-test was done using 535 students (1.3% of the original students' population), the teachers pre-test should have been done using around 80 teachers and not 20 (0.4% of the original teachers' population).

Anticipated Contributions of the Current Study

The current study is explanatory in nature, seeking answers to questions relating teachers' assessment practices to students' achievement. The findings of this study's correlational analyses

did not imply any causal inference. Instead, they proposed suggestions about potential relations between teachers' assessment practices and students' achievements in the problem solving component of the SAIP 2001 mathematics achievement tests. It is anticipated that the proposed findings may contribute to the empirical research and body of knowledge in the educational assessment field. In addition to the theoretical contribution, practical implications could also be anticipated.

As mentioned earlier, the conceptual framework of this study builds on Brookhart's theory (Brookhart, 1997). The current findings tried to support Brookhart's assertion by providing plausible explanations or predictions of the impact of math teachers' assessment practices on students' achievement in problem solving. However, given the findings at hand, that support was not strong. According to findings, in a classroom assessment event, different math classroom assessment practices such as giving weight to projects, or frequently use assessment information for feedback purposes, could have an impact on students' achievement. Some of the assessment practices had small positive relations with students' achievement (e.g. giving weight to projects and the frequency of use of assessment information for feedback), while others had small negative relations (e.g. giving weight to portfolios).

Lastly, although the findings of the current study did not provide strong correlations or causal inferences, it is comparable to findings proposed by other studies done within the same research context. It is believed that with further research on the current topic, evidences may become available that could better inform education policy leaders and educators. Those evidences could provide math educators with different potential relations through professional development initiatives so they may base their assessment practices on an informed decision. At that point, teachers may, to some extent, predict in advance the impact of their assessment

practices on their students' achievement. Eventually, that could improve students' learning and hence their achievement.

Suggestions for Future Research

In this study, it is believed that the use of Kendall's τ_b statistical procedures was appropriate, because of the ordinal nature of the investigated variables. However, other statistical procedures could also be used to investigate the relation between teachers' assessment practices and students' achievement in large-scale assessment tests. The nature of the datasets in such tests (e.g. SAIP, TIMSS) is students nested within classes, and the classes in turn are nested within schools. Given this nested structure, it was noticed that the HLM model is employed by different secondary analyses studies in the literature (Anderson at al, 2006; Rodriguez, 1999; Rogers, Anderson, Klinger, & Dawber, 2006). These studies claim that the use of HLM model will reduce the possibility of error in the findings of the analyses. However, the condition of having a large sample size at different levels is not always guaranteed with the possible volume of missing data in the analyses.

On the other hand, Rogers et al assert that the lowest sample unit of the analyses should be the class instead of the student. In doing this, they suggest that while sampling, classes within schools should be chosen and all the students with these classes are chosen as well (p. 767). The current study echoes that suggestion as it will allow for more control over data collection procedures and reduce the possibility of missing data. Perhaps, more empirical studies could use recent large scale assessment data (e.g. SAIP or TIMSS) to re-examine the aforementioned relations to provide better explanations. As well, perhaps if enough sample size is established at

different levels of an HLM model, its use could reduce the possibility of error in the findings of the secondary analyses, thereby having more confidence in the findings.

CHAPTER VI

Conclusion

Improving students' achievement has been the centre of attention of educators, policy leaders, and parents. That improvement comes into effect through continuous reform efforts done by all the vested stakeholders. Research has always been the main source of providing valid evidences to those stakeholders, enabling them to accomplish educational reforms. This study strived to provide researchers in the assessment field as well as different stake holders with evidence about the impact of assessment practices on students' achievement in mathematics.

The thesis proposed a few relations linking classroom assessment practices and students' achievements in large-scale assessment tests. These relations complemented other efforts that tried to evaluate assessment theories (e.g. Brookharts' theory). This evaluation is substantial as it provides education policy leaders with information, through which they can craft effective policy statements. These policy statements are expected to familiarize educators with the need for assessment practices' reform as well as different methodologies to implement it. By doing that, the gap between research and practice is expected to diminish gradually, and more communication channels will be facilitated.

Some of the findings of this study proposed the existence of a few relations between teachers' assessment practices in the math classroom and students' achievement in problem solving in large-scale assessment tests (e.g. SAIP 2001). These relations varied in their strengths and directions. Some of the findings revealed positive relations between teachers' assessment practices and students' achievement (e.g. giving weight to projects and the frequency of use of assessment information for feedback purposes). These positive relations mean that an increased

use of these assessment practices could be associated with higher levels of students' mathematics achievement in problem solving. On the other hand, some of the findings proposed negative relations between teachers' assessment practices and students' achievement (e.g. giving weight to portfolios and the frequency of use of assessment information for diagnostic purposes). These findings indicated that less use of these assessment practices could be associated with higher levels of students' mathematics achievement in problem solving.

Future research efforts trying to produce more evidences should consider a few important factors. First, the size of missing data could affect correlations among variables. Therefore, EM methods should be considered before starting data analysis. Second, using other analytical methods (e.g. HLM model) could be helpful in understanding the relations among variables. Third, all the secondary analysis studies, including this one, lack causal relationships among variables. It is anticipated the following these studies with experimental or quasi-experimental studies may confirm current relations and could propose cause and effect assertions. Finally, teachers should be questioned about their perception of what projects, portfolios, or observation/interviews mean. Having an understanding of the operational definitions pertaining to these practices could help in designing more valid questionnaire items, and eventually avoid awkward interpretations by teachers. The previous suggestions are only a short list of an array of possibilities that could help linking teachers' assessment practices and students' achievement in large-scale assessment, thereby leading to improving students' achievement.

REFERENCES

Allison, P. (2002). *Missing data.* Thousand Oaks, CA: Sage Publications.

Anderson, O. Rogers, T., Klinger A., Ungerlieder, C., Glickman, V. & Anderson, B. (2006). Student and school correlates of Mathematics achievement: model of school performance based on Pan-Canadian student assessment. *Canadian Journal of Education, 29(3),* 706-730.

Angelis, Z. (2003). *The structure of attitudes towards science*. M.A. dissertation, Toronto: University of Toronto. Retrieved from ProQuest Digital Dissertations.

Barlow, A. & Drake. J. (2008). Division by a fraction: assessing understanding through problem writing. *Mathematics Teaching in the Middle School, 13(6),* 326-332.

Barribeau, P., Butler, B., Corney, J., Doney, M., Gault, J., Gordon, J., Fetzer, R., Klein, A., Rogers, C., Stein, I., Steiner, C., Urschel, H., Waggoner, T. & Palmquist, M. (2005). *Survey Research.* Writing@CSU. Colorado State University Department of English. Retrieved June 10, 2008 from http://writing.colostate.edu/guides/research/survey/.

Black, P. & Wiliam D. (1998). Assessment and classroom learning, *assessment in education, 5(1),* 7-74.

Blumenfeld, P., Soloway, E., Marx, R., Krajcik, J. S., Guzdial, M. & Palincsar, A. (1991). Motivating project-based learning. *Educational Psychologist, 2*6(3 & 4), 369 - 398.

Brookhart, S. (1997). Theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education, 10(2),* 161-180.

Brookhart, S. (2008). Feedback that fits. *Educational Leadership, 65(4)*, 54-59.

Callingham, R. (2008). Dialogue & feedback: assessment in the primary mathematics classroom. *Preview By: Australian Primary Mathematics Classroom, 13(3)*, 18-21.

Chappuis, S. & Stiggins, R. (2002). Classroom assessment for learning. *Educational Leadership, 60(1)*, 40-43.

Ciofalo, J. F., Wylie, E. C. (2006). Using diagnostic classroom assessment: one item at a time. *Teachers College Record*. Retrieved March 15, 2008 from http://www.tcrecord.org/content.asp?contentid=12285.

Cizek, G. (1997). *Learning, achievement, and assessment: Construct at the crossroad.* Handbook of classroom assessment, NY: Acedemic Press.

Cizek, G. (2009). Reliability and validity of information about student achievement: comparing large-scale and classroom testing contexts. *Theory Into Practice, 48(1)*, 63-71.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Colton, D. & Covert, R. (2007). *Designing and constructing instruments for social research and evaluation.* San Francisco, CA: Jossey-Bass Publishing.

Cooper, H. (1989). *Homework*. White Plains, NY: Longman.

Cooper, H., Robinson, J. & Patall E. (2006). Does homework improve academic achievement? a synthesis of research, 1987-2003. *Review of Educational Research, 76(1)*, 1-62.

Council of Ministries of Education, Canada (2002). Report on mathematics assessment III: school achievement indicators program 2001.Retrieved May 15, 2008, from http://www.cmec.ca/Programs/assessment/pancan/saip2001/Documents/saip2001math.en.pdf.

Council of Ministries of Education, Canada (2003). Mathematics learning: the Canadian context, school achievement indicators program – mathematics III, 2001. Retrieved May 15, 2008, from http://www.cmec.ca/Programs/assessment/pancan/saip2001/Documents/context.en.pdf.

Croy, V. & Novins, D. (2005). Methods for addressing missing data in psychiatric and developmental research. *Journal of the American Academy of Child and Adolescent Psychiatry, 44(12),* 1230-1240.

Dudaitė, J. (2006). Change of mathematical achievement in the light of educational reform in Lithuania. Paper presented at the International Association for the Evaluation of Educational Achievement (IEA) Web site: http://www.iea.nl/fileadmin/user_upload/IRC2006/IEA_Program/TIMSS/Dudaite1.pdf.

Earl, L. (2003). *Assessment as learning: using classroom assessment to maximize student learning.* Thousand Oaks, CA, Corwin Press.

Emenogu, B. & Childs, R. (2003). Curriculum and translation differential item functioning: a comparison of two dif detection techniques: a comparison of dif detection techniques. Annual Meeting of National Council of Measurement in Education (Chicago, Il, April 22-24, 2003). Retrieved July 10, 2008, from http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/0b/d3.pdf.

Goodman, L. & Kruskal, W. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49(268),* 732- 764.

Goodman, L. & Kruskal, W. (1979). *Measures of association for cross classifications.* New York: Springer.

Hastedt, D. (2004). Differences between multiple-choice and constructed response items in PIRLS 2001. Paper presented at the International Association for the Evaluation of Educational Achievement (IEA) Web site: http://www.iea.nl/fileadmin/user_upload/IRC2004/Hastedt.pdf

Hao, S. (2005). Teachers' assessment practices and fourth graders' reading literacy achievements: An international study. Ph.D. dissertation. South Carolina: University of South Carolina. Retrieved from ProQuest Digital Dissertations.

Jones, R. (2004). Research on TIMSS data provides information for educational improvement in Ontario. Paper presented at the International Association for the Evaluation of Educational Achievement (IEA) Web site: http://www.iea.nl/fileadmin/user_upload/IRC2004/Jones.pdf.

Kendall, G. (1962). *Rank correlation methods.* New York: Hafner.

Maxwell, S. E. & Delaney, H. D. (2004). *Designing experiments and analyzing data: a model comparison perspective (2nd ed.).* Mahwah, NJ: Lawrence Erlbaum Associates.

McMillan, J. (1997). *Classroom assessment: principles and practice for effective instruction.* Boston: Allyn & Bacon.

McMillian, J. (2003). The relationship between instructional and classroom assessment practices of elementary teachers and student scores on high-stakes tests (Report No. TM034718). Virgina: ERIC Clearinghouse on Assessment and Evaluation. (ERIC Document Reproduction Service No. ED472164) Retrieved on August 2, 2008 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1a/b0/31.pdf.

McMillan, J., Myran, S. & Workman D. (2002). Elementary teachers' classroom assessment and grading practices. *Journal of Educational Research, 95(4),* 203-213.

McMillan, J. & Workman, D. (1998). *Classroom assessment and grading practices: A review of the literature.* Richmond, VA: Metropolitan Educational Research Consortium, Virginia.

Nelson, B. & Sassi, A. (2007). What math teachers need most. *Education Digest: Essential Readings Condensed for Quick Review, 72(6),* 54-56.

Ontario Ministry of Education. (2005). *The Ontario Curriculum Grades 1 to 8- Mathematics.* Toronto: Ministry of Education.

Popham, W. (1995). *Classroom assessment: what teachers need to know*, Boston: Allyn & Bacon.

Rodriguez, M. (1999). Linking classroom assessment practices to large-scale test performance. Ph.D. dissertation. East Lansing: University of Michigan. Retrieved from ProQuest Digital Dissertations.

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

Santa Cruz, Rafaela M. (2009). Giving voice to English language learners in mathematics. *NCTM News Bulletin.* Retrieved on January 16, 2009 from http://www.nctm.org/news/content.aspx?id=16895.

Stiggins, R. (1994). *Student-centered classroom assessment.* New York: Merrill.

Stiggins, R. & Conklin, F. (1992). *In teachers' hands: investigating the practices of classroom assessment.* Albany: State University of New York Press.

Stiggins, R. & Chappuis, S. (2005) Putting testing in perspective: it's for learning. *Principal Leadership, 6(2)*, 16-20.

Tuttle, H. (2008). Formative assessment through homework. Retrieved November 17, 2008, from http://www.authenticeducation.org/bigideas/article.lasso?artId=63&-session=Auth:8D95D4D40533c2028AMsFEF5FFBB.

Wormeli, R. (2006). *Fair isn't always equal: assessing & grading in the differentiated classroom.* Portland, Maine: Stenhouse Publishers.