CHANGES IN SCHOOL RESULTS IN EQAO ASSESSMENTS FROM 2006 TO 2010

by

Anita Ram

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

©Copyright by Anita Ram 2012
Many accountability systems use data from large-scale assessments to make judgements about school performance. In Ontario, school performance is often assessed using the percentage of proficient students (PPS). The purpose of this study was to shed light on the degree and frequency of changes from year to year in the percentage of proficient students, at a school, in the areas of reading, writing and mathematics for both grades 3 and 6 in Ontario from 2006 to 2010. A second purpose was to assess the influence of cohort size on the variability in scores from year to year. Once schools not having data for 5 consecutive years and outliers were omitted secondary data analysis was used to examine nearly 3000 schools in each subject and grade. For the first part of the study, descriptive statistics and frequencies were the main method of examination. In the second part of the study, variance scores and correlations were used in order to understand the relationship between changes in PPS and cohort size. Findings revealed that changes in school scores from year to year are very large for many schools. Approximately 50 percent of schools experienced changes in PPS greater than 10 percent in any given year. When examining how often, from 2006 to 2010, a school experienced a similar amount of change – generally, both the smallest and largest change categories had a larger percentage of schools experiencing a similar amount of change for two and three years. Very seldom did schools experience the same degree of change in PPS across all 5 years. Results from correlations revealed a significant and inverse relationship between average cohort size and variability in PPS. Considering over 80 percent of schools have 60 or fewer students in a cohort
the unpredictability in PPS may prove to be quite frustrating to schools and confusing to
stakeholders. Annual PPS scores appear to be a poor indicator of real school performance, and
their use to rank or rate schools should be avoided. Recommendations are made about using PPS
to report school level results for EQAO, schools and the public.
Acknowledgements

This paper would not have been possible without the help and support of many individuals. I would first like to express my deepest gratitude to my supervisor Dr. Ben Levin for his generosity with his time, knowledge and support throughout each phase of the dissertation writing process. Despite his many obligations his willingness to put his students first never wavered. I would also like to thank my faculty advisor Dr. Ruth Childs for her thorough feedback and for providing me with opportunities to enhance my research and data analysis skills. Also, I extend my gratitude to Dr. Joe Flessa, Dr. Louis Volante and Dr. Eunice Jang for their time, thoughtful suggestions and insights that contributed to the final product.

I am eternally indebted to my wonderful parents who are an endless source of love, support, and encouragement no matter what I choose to do. Not to mention having put up with me being a graduate student for many years. I extend this gratitude to my brother and the rest of my family – I can now answer your commonly asked question of “Are you done, yet?” with a yes!

To all of my friends at and outside of OISE (students and professors alike) including the Datahost family - I would like to thank all of you for providing me with opportunities to grow both personally and professionally. Your continued moral support, encouragement and willingness to listen and lend a hand throughout the years has always kept me moving forward.

Lastly, I would like to thank the Education Quality and Accountability Office (EQAO) for providing the data used in this research. The opinions presented in this paper do not necessarily reflect the opinions of EQAO.
# Table of Contents

CHAPTER ONE. INTRODUCTION

- Purpose of Study ........................................ 7
- Research Questions ..................................... 7
- Definition of Terms ..................................... 8

CHAPTER TWO. LITERATURE REVIEW

- Large-Scale Assessments ............................... 10
  - Purpose of Large-Scale Assessments ............... 10
  - Education Accountability ............................ 11
  - Large-Scale Assessments in Canada ................. 12
  - Large-Scale Assessments in Ontario ............... 13
  - Education Quality and Accountability Office (EQAO) .......... 15
  - Score Reporting ...................................... 16
  - Misuse of Test Results .............................. 20
  - Error, reliability and validity ....................... 21
  - Fluctuations in Achievement Scores ............... 25
  - Limitations of Literature ........................... 29
  - Influence of Cohort Size on Achievement Scores .... 31

CHAPTER THREE. METHODOLOGY

- Research Design ....................................... 36
  - Secondary data analysis ............................ 36
- Overview of EQAO Dataset .......................... 37
  - Scoring procedure .................................. 38
  - Equating procedure .................................. 39
- Overview of Curriculum ............................. 39
- Population ............................................ 40
- Outcome Measure .................................... 40
- Analysis .............................................. 41
  - Organization of data ................................ 41
- Analytic Approach .................................... 42
  - Descriptive statistics ............................... 42
  - Correlational analysis .............................. 43

CHAPTER FOUR. RESULTS

- Descriptive Results .................................. 44
  - Reading ............................................. 44
  - Writing ............................................. 46
  - Mathematics ........................................ 47
- Cross-year Correlations of PPS ..................... 50
  - Reading grade 3 ................................... 50
  - Reading grade 6 ................................... 50
  - Writing grade 3 .................................... 51
  - Writing grade 6 .................................... 51
  - Mathematics grade 3 ................................ 52
  - Mathematics grade 6 ............................... 52
List of Tables

Table 1.1 Percentage of Grade 3 and 6 Students at or Above the Provincial Standard in Reading, Writing, and Mathematics from 2006 to 2010 .............................. 5

Table 2.1 Overview of Assessment Programs Across Canada by Province/Territory ........................................ 14

Table 3.1 Level Descriptions From the Ontario Curriculum .................. 38

Table 3.2 Skills Assessed on the Grades 3 and 6 Literacy and Numeracy Assessment ........................................ 40

Table 3.3 Number of Schools and Percentage of Outliers in Each Subject and Grade ........................................ 42

Table 4.1 Descriptive Statistics for Grade 3 and 6 Reading PPS .............. 45

Table 4.2 Descriptive Statistics of Annual Changes in PPS for Grade 3 and 6 Reading ........................................ 46

Table 4.3 Descriptive Statistics for Grade 3 and 6 Writing PPS .............. 47

Table 4.4 Descriptive Statistics of Annual Changes in PPS for Grade 3 and 6 Writing ........................................ 48

Table 4.5 Descriptive Statistics for Grade 3 and 6 Mathematics PPS ...... 49

Table 4.6 Descriptive Statistics of Annual Changes in PPS for Grade 3 and 6 Mathematics ........................................ 50

Table 4.7 Correlation of Reading PPS for Grade 3 and 6 ....................... 51

Table 4.8 Correlation of Writing PPS for Grade 3 and Grade 6 .............. 51

Table 4.9 Correlation of Mathematics PPS for Grade 3 and Grade 6 ....... 52

Table 4.10 Correlations Across Grades and Subjects From 2006 to 2010 ...... 68

Table 4.11 Number of Schools in Each Cohort Size Category ................ 69

Table 4.12 Descriptive Statistics of Variance Scores ......................... 70
List of Figures

Figure 1.1 Percentage of proficient grade 3 students across 20 randomly selected schools in reading from 2006 to 2010 .............................................................. 6
Figure 2.1 Distribution of the number of students in each EQAO achievement category using two schools with the same percentage of proficient students of 67 ....... 19
Figure 4.1 The percentage of schools by size of change in PPS from 2006 to 2010 for grade 3 reading ................................................................. 54
Figure 4.2 Frequency of annual change in PPS from 2006 to 2010 for grade 3 reading .... 55
Figure 4.3 The percentage of schools by size of change in PPS from 2006 to 2010 for grade 6 reading ................................................................. 56
Figure 4.4 Frequency of annual change in PPS from 2006 to 2010 for grade 6 reading ... 57
Figure 4.5 The percentage of schools by size of change in PPS from 2006 to 2010 for grade 3 writing ................................................................. 58
Figure 4.6 Frequency of annual change in PPS from 2006 to 2010 for grade 3 writing ... 59
Figure 4.7 The percentage of schools by size of change in PPS from 2006 to 2010 for grade 6 writing ................................................................. 60
Figure 4.8 Frequency of annual change in PPS from 2006 to 2010 for grade 6 writing ... 61
Figure 4.9 The percentage of schools by size of change in PPS from 2006 to 2010 for grade 3 mathematics ...................................................... 62
Figure 4.10 Frequency of annual change in PPS from 2006 to 2010 for grade 3 mathematics ................................................................. 63
Figure 4.11 The percentage of schools by size of change in PPS from 2006 to 2010 for grade 6 mathematics ...................................................... 64
Figure 4.12 Frequency of annual change in PPS from 2006 to 2010 for grade 6 mathematics ................................................................. 65
Figure 4.13 The percentage of school by patterns of annual change for all grades and subjects from 2006 to 2010 .............................................................. 67
Figure 4.14 Average variance by cohort size for grade 3 and 6 reading, writing, and mathematics ................................................................. 70
Figure 5.1 Average of school PPS across multiple subjects ...................................................... 80
Figure 1A Distribution of grade 3 reading PPS for 2006 ...................................................... 97
Figure 2A Distribution of grade 3 reading PPS for 2007 ...................................................... 97
Figure 3A Distribution of grade 3 reading PPS for 2008 ...................................................... 98
Figure 4A Distribution of grade 3 reading PPS for 2009 ...................................................... 98
Figure 5A Distribution of grade 3 reading PPS for 2010 ...................................................... 99
Figure 4H  Distribution of annual change in grade 6 reading PPS, 2009-2010.............. 118
Figure 1I  Distribution of annual change in grade 3 writing PPS, 2006-2007............... 119
Figure 2I  Distribution of annual change in grade 3 writing PPS, 2007-2008............... 119
Figure 3I  Distribution of annual change in grade 3 writing PPS, 2008-2009............. 120
Figure 4I  Distribution of annual change in grade 3 writing PPS, 2009-2010............... 120
Figure 1J  Distribution of annual change in grade 6 writing PPS, 2006-2007............... 121
Figure 2J  Distribution of annual change in grade 6 writing PPS, 2007-2008............... 121
Figure 3J  Distribution of annual change in grade 6 writing PPS, 2008-2009............... 122
Figure 4J  Distribution of annual change in grade 6 writing PPS, 2009-2010............... 122
Figure 1K  Distribution of annual change in grade 3 mathematics PPS, 2006-2007....... 123
Figure 2K  Distribution of annual change in grade 3 mathematics PPS, 2007-2008....... 123
Figure 3K  Distribution of annual change in grade 3 mathematics PPS, 2008-2009....... 124
Figure 4K  Distribution of annual change in grade 3 mathematics PPS, 2009-2010....... 124
Figure 1L  Distribution of annual change in grade 6 mathematics PPS, 2006-2007....... 125
Figure 2L  Distribution of annual change in grade 6 mathematics PPS, 2007-2008....... 125
Figure 3L  Distribution of annual change in grade 6 mathematics PPS, 2008-2009....... 126
Figure 4L  Distribution of annual change in grade 6 mathematics PPS, 2009-2010....... 126
Figure 1M  Distribution of grade 3 reading variance scores.................................... 127
Figure 2M  Distribution of grade 6 reading variance scores.................................... 127
Figure 3M  Distribution of grade 3 writing variance scores.................................... 128
Figure 4M  Distribution of grade 6 writing variance scores.................................... 128
Figure 5M  Distribution of grade 3 mathematics variance scores.............................. 129
Figure 6M  Distribution of grade 6 mathematics variance scores.............................. 129
List of Appendices

Appendix A: Distribution of Grade 3 Reading PPS .............................................. 97
Appendix B: Distribution of Grade 6 Reading PPS .............................................. 100
Appendix C: Distribution of Grade 3 Writing PPS .............................................. 103
Appendix D: Distribution of Grade 6 Writing PPS .............................................. 106
Appendix E: Distribution of Grade 3 Mathematics PPS ................................... 109
Appendix F: Distribution of Grade 6 Mathematics PPS ................................... 112
Appendix G: Distribution of Annual Change in Grade 3 Reading PPS ............. 115
Appendix H: Distribution of Annual Change in Grade 6 Reading PPS ............. 117
Appendix I: Distribution of Annual Change in Grade 3 Writing PPS .............. 119
Appendix J: Distribution of Annual Change in Grade 6 Writing PPS .............. 121
Appendix K: Distribution of Annual Change in Grade 3 Mathematics PPS ....... 123
Appendix L: Distribution of Annual Change in Grade 6 Mathematics PPS ....... 125
Appendix M: Variance Score Distributions ......................................................... 127
CHAPTER ONE

INTRODUCTION

The educational restructuring that occurred in Ontario between the years 1993 and 2003 is described as turbulent (Leithwood, Fullan & Watson, 2003; Taylor, 2005). The election of the Conservative government under Premier Mike Harris in June, 1995, led to a series of changes in education in Ontario which included: curriculum changes; the change from a 5-year to a 4-year high school program; centralization of funding; reduction in the number of school boards; removing principals from teacher unions; and increased school board reporting requirements (Taylor, 2005). With increasing public dissatisfaction, labour disruptions, increasing private school enrolment, and a high turnover among teachers a new government was elected in 2003 (and re-elected in 2007) (Levin, 2007). One of the goals of the government, which began in 2003, was to improve elementary school literacy and numeracy outcomes – with the intention of demonstrating achievement of good outcomes for students (Levin, 2007).

Tremendous pressures have been placed on public services, like education, to cope with the changing social, economic, political, and cultural forces amidst the structural changes that have occurred over the past 30 years; for example, changing demographics, distribution of wealth, international competitiveness, ethnic diversity, changing values and public dissatisfaction with governments (McEwan, 1995). To demonstrate accountability to stakeholders for educational performance and educational costs, provinces use a variety of tools (McEwan, 1995); large-scale assessments were developed in part to keep up with the changing demands of society.

In Ontario, the Education Quality and Accountability Office (EQAO) develops and implements tests to elementary school students. Test results, which provide information on the standing of students at a school in reading, writing and mathematics, are distributed to parents,
teachers, principals, and school boards and are made available to the public (EQAO, 2011). In a world increasingly relying on accountability, many schools today, across the world, use data obtained from large-scale assessments as a basis for their school accountability, improvement and planning. Large-scale assessments provide individual and system data on student achievement (EQAO, 2011) and are common in using data to promote student success. The results of large-scale tests not only have an impact on many students, teachers, and schools, but also policymakers, parents, and ultimately the community (Sicoly, 2002).

Each year, in Ontario, millions of dollars are spent on EQAO testing. Although the EQAO assessments do not have an influence on students’ marks or their advancement to the next grade, the test scores may be used to: allocate resources and support to schools that have a larger proportion of students not meeting the provincial standard (Klinger, 2011; People for Education, 2009); modify or add school programs; track improvement over time to monitor school performance; and judge the quality of schools (Sicoly, 2002) (i.e. Is the school doing a good job of educating students in the school?). In some cases, test results may be published in newspapers ranking schools and some may use the results to imply teacher effectiveness (Baker et al., 2010) which may lead to enquiry by the public and anxiety, anger, embarrassment and criticism among school staff (Smith, 1991).

The School Information Finder - a resource created for interested citizens on the Ministry of Education, Ontario website - allows visitors to learn more about publically funded schools in Ontario by searching by city, school board, postal code or school name (Ministry of Education, 2010). Website visitors have access to such information as grade 3 and 6 achievement in reading, writing and mathematics, the number of pupils in each class in the school and information about the student population (e.g. students whose first language is not English, students who receive
special education services) (Ministry of Education, 2010). Although there has been some controversy in displaying detailed information about schools and their students and the layout of the website, the information has not been taken down because the statistics from provincial education surveys, on the website, are meant to provide information on publicly funded schools in Ontario for those who are interested in learning more (Bowden, 2009; Brown, 2009; Howlett, 2009; Pettibone, 2009). And organizations such as The Fraser Institute - an independent research and educational organization - publish Report Cards every year detailing how a school has ranked in academics in comparison to other schools (Cowley, Easton & Thomas, 2010). This ranking is also meant to aid parents in selecting schools for their children in addition to evaluating a school’s yearly performance (Cowley, Easton & Thomas, 2010). Many have argued the usefulness of such rankings (e.g. Earl, 1995; Johnson, 2005; Klinger, 2011; Pascal, 2005; Rogers, 2011; Rogers & Klinger, 2005) (See Cowley, Easton & Thomas, 2010 for calculation of Fraser Institute Report Card rating).

In Ontario, the Ontario Focused Intervention Partnership (OFIP) identifies 3 sets of schools requiring assistance:

OFIP 1 – schools where less than 34% of students are achieving at Level 3 (the provincial standard) or Level 4 in reading, in two of the past three years.

OFIP 2 – schools where 34%-50% of students are achieving at Level 3 or 4 in reading and results have been static or declining based on trends over the past three years.

OFIP 3 – identifies schools where 51%-74% of students are achieving at Level 3 or 4 in reading but results have been static or declining based on three-year trends. (Ministry of Education, 2008).
Once these schools are identified boards may use OFIP funds for their improvement strategies such as:

- dedicated blocks of uninterrupted time for literacy and numeracy, a school improvement team that meets regularly to review school data and plan next steps, a mechanism to regularly monitor the growth and progress of specific students, a capacity-building strategy for school staff to improve instructional effectiveness, a common assessment tool for primary and junior grades used in all of the schools in the OFIP strategy, adequate resources to ensure a comprehensive literacy program, teacher moderation of samples of student learning which leads to instructional planning, school and classroom organization and scheduling, a school culture that makes school improvement a whole school priority, professional development and capacity-building for teachers and principals aligned to the School Effectiveness Framework and the school improvement plan that strengthens instructional practice, parental involvement and community engagement. (Ministry of Education, 2008).

Although, provincially, achievement on EQAO tests may show a seemingly consistent trend in year to year performance (Table 1.1) (EQAO, 2010), within schools, achievement may be more susceptible to fluctuations in scores across school years. For example, using the percentage of proficient students (PPS) to evaluate results in EQAO reading from 2006 to 2010, one school may have 89 percent of their students achieving proficiency (Level 3 or higher) whereas in the following year the percentage may increase to 93 percent and drop the year after to 90 percent (or more drastic changes in some cases) (Figure 1.1). Some may understand these
Table 1.1

Percentage of Grade 3 and 6 Students at or Above the Provincial Standard in Reading, Writing, and Mathematics from 2006 to 2010

<table>
<thead>
<tr>
<th>Grade</th>
<th>Subject</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Reading</td>
<td>62</td>
<td>62</td>
<td>61</td>
<td>61</td>
<td>62</td>
</tr>
<tr>
<td>3</td>
<td>Writing</td>
<td>64</td>
<td>64</td>
<td>66</td>
<td>68</td>
<td>70</td>
</tr>
<tr>
<td>3</td>
<td>Mathematics</td>
<td>68</td>
<td>69</td>
<td>68</td>
<td>70</td>
<td>71</td>
</tr>
<tr>
<td>6</td>
<td>Reading</td>
<td>64</td>
<td>64</td>
<td>66</td>
<td>69</td>
<td>72</td>
</tr>
<tr>
<td>6</td>
<td>Writing</td>
<td>61</td>
<td>61</td>
<td>67</td>
<td>67</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>Mathematics</td>
<td>61</td>
<td>59</td>
<td>61</td>
<td>63</td>
<td>61</td>
</tr>
</tbody>
</table>

Adapted from EQAO (2010).

fluctuating results to mean an inadequate school leading the public to interpret the quality of the school and staff as not doing a very good job at educating students in the school. “The tendency to equate good [test] scores with good schools and good teaching is an urge that is difficult to resist….Enormous pressure may be put on schools to do well or to demonstrate incremental improvement over time.” (Sicoly, 2002, p. 174) Relying on PPS to inform school-based decisions, with variations in PPS from one year to the next, may be detrimental if the scores are more a result of extraneous factors than an accurate reflection of school achievement and meaningful changes in performance (Kane & Staiger, 2002; Sicoly, 2002; Smith & Smith, 2005; Turque, 2009). Failing to acknowledge fluctuations in test scores over time and not examining the degree with which this occurs may lead to misinterpretations - successes and failures may be mistaken for natural fluctuations (Smith & Smith, 2005).
There is a growing importance of school achievement scores from large-scale assessments and increasing concern about fluctuating school level scores; therefore, inconsistent results from large-scale testing could lead to the adoption of practices that do not have real merit, further confusing which strategies are appropriate for school reform (Kane & Staiger, 2002). Theoretically, school level performance is expected to improve by improving educational practices based on results from large-scale testing.

Considering that in many countries school results are used to judge schools, the topics of stability and change are becoming increasingly more relevant and so is the need for studies examining school level results over a longer period of time (Doolaard, 2002). This study examines variations in school level PPS across 5 consecutive years and compared them across subjects and grades. Examining school-level PPS sheds light on how often fluctuations in PPS
occur and to what extent. Without such information it is difficult to discern a baseline for any changes in school level scores from year to year.

**Purpose of Study**

The purpose of this study was to examine changes in school level PPS from 2006 to 2010 to reveal the degree and frequency of changes in the percentage of proficient students, at a school, in the areas of reading, writing, and mathematics for both grade 3 and 6 elementary schools in Ontario. EQAO data was used from the literacy and numeracy assessments administered to Grade 3 and Grade 6 students in reading, writing and mathematics.

**Research Questions**

1. What is the degree and frequency with which PPS changes, at a school, across 5 years for grade 3 and 6 reading, writing and mathematics?
2. Is there a pattern of change, in PPS, specific to grade 3 or grade 6 reading, writing, or mathematics?
3. Is there a relationship between change in PPS across 5 years and the size of the cohort writing the assessment for grade 3 and grade 6 reading, writing and mathematics?

In this study, the school is used as the unit of analysis.
Definition of Terms

*Accommodations.* “‘Accommodations’ [for EQAO testing] are forms of support and services that enable students with special needs to demonstrate their competencies in the skills being measured by the test. Accommodations change only the way in which the test is administered or the way in which the person tested responds to the components of the test. They do not alter the content of the test or affect the validity or reliability of the test.” (EQAO, 2011).

*Accountability.* “[T]he demand that a particular individual or institution assume some responsibility and demonstrate it in a certain form” (Smith & Fey, 2000, p. 335).

*Annual Change.* The difference between two scores, such as PPS, a year or more apart.

*Assessment.* Relates to the process of collecting information from a number of sources including, but not limited to, paper and pencil activities, assignments, oral questioning, observation and providing students with descriptive feedback (Ontario Ministry of Education and Training, 1999).

*Consistency (of achievement).* “[C]onsiders the correlation between rankings of schools based on different criterion variables.” (Doolaard, 2002, p. 774).

*Cross-year stability coefficient.* Is “the correlation between assessment scores for a specific subject and grade during two consecutive years” (Sicoly, 2002, p. 175).

*Expectations.* “The statements of the knowledge and skills that students are expected to learn and demonstrate in their work and in the activities used to assess their achievement.” (Ontario Ministry of Education and Training, 1999, p. 149).

*EQAO dot score.* Refers to the score achieved by a student on the province-wide testing in Ontario. The scale of the EQAO score ranges from 1.1 to 4.9 and increases in increments of
0.2 (e.g. 1.3, 1.5, 1.7 etc.) (EQAO, 2011). These are not reported publicly but are available in student level results.

**Fluctuate.** Means “to shift back and forth uncertainly” (Merriam-Webster, 2012).

**Modifications.** [For EQAO testing] are changes to content and to performance criteria and are not permitted as they would affect the validity and reliability of the assessment results (EQAO, 2011).

**Percentage above cut score (PAC).** The percentage of students scoring above a pre-determined cut score.

**Percentage of proficient students (PPS).** The percentage of students, at a school, achieving at or above the expected standard (levels 3 and 4 in Ontario). In this paper, this term is used interchangeably with school score and achievement score.

**Special Provisions for English Language Learners.** “‘Special provisions’ [for EQAO testing] are adjustments to the setting and/or timing of the assessment for English language learners. These provisions do not affect the validity or reliability of the assessment results for these students.” (EQAO, 2011, p.10).

**Standardization.** “[I]mplies the uniformity of procedure in administering and scoring the test.” (Anastasi & Urbina, 1997, p. 6).

**Unidimensionality.** “[R]elates to whether results of schools and classes are the same for different grades, subjects, years and groups of students. In other words, are results of schooling consistent and stable?” (Doolaard, 2002, p. 774).
CHAPTER TWO

LITERATURE REVIEW

The use of large-scale assessments has changed over the past few decades and in many cases scores from large-scale assessments are used to inform decisions about schools. This literature review addresses the history and purpose of large-scale assessments, methods used by studies that measure year to year changes in achievement, the quality and gap in research, score reporting, use and misuse of testing and the influence of cohort size on changes in school level achievement scores. The analyses conducted in this study will help contribute to a body of literature that does not yet have a commonly accepted definition for classifying ‘normal’ or expected change in PPS across years, subjects and grades.

Large-Scale Assessments

Large-scale assessments are wide-spread and used frequently to measure achievement and can be used to provide the profession and the public with information on the achievement of students and schools (Crundwell, 2005; Earl, 1999).

Purpose of Large-Scale Assessments

In the United States, standardized tests have been in use as early as 1840 (Koretz, 2002). Standardized tests, after World War II, were mainly used to evaluate individual students and less frequently to evaluate curricula (Koretz, 2002). McEwen (1995) explains the purpose of large-scale assessment programs is to observe student achievement to provide stakeholders with information about the progress of students’ learning. Similarly, Koretz (2002) describes that the introduction of standardized tests was to hold educators accountable by tracking performance of education systems, while Airasian, Madaus & Pedulla (1979) expressed the purpose to be “organized efforts to make sure public school students are able to demonstrate their mastery of
certain minimum skills needed to perform tasks they will routinely confront in adult life” (p. 5). Almost all countries with centralized educational curriculum use their large-scale testing programs to monitor improvement of school performance over time or certify student achievement in relation to national standards (OECD, 2004), implement and evaluate educational interventions, policies, and the effectiveness of a school (Klinger, DeLuca, & Miller 2008) or (in rare cases) provide salary incentives to teachers having higher student scores (Kane & Staiger, 2002). An examination of assessment results is expected to lead to improved educational practices (e.g. adjustment of teaching practices or focus on staff development efforts) which in turn may result in an improvement in student performance (Sicoly, 2002).

Examinations have long been a part of schooling and viewed as an equitable way to identify the best candidates for scarce opportunities for instance to direct students into various programs or appropriate work positions (Earl, 1999; Nagy 2000). In other countries, like England, France and Hong Kong, centralized assessment programs have meant key examinations for students at pivotal points, like A-levels, the Baccalaureat, and the public examinations respectively (Earl, 1999). In Canada, education is a provincial responsibility. Canadian provinces and territories participate in national (e.g. Pan-Canadian Assessment Program (PCAP)) and international (e.g. Trends in International Mathematics and Science Study (TIMSS), Programme for International Student Assessment (PISA)) assessments. These provide external points of reference for comparing results across Canada and around the world (Earl, 1999; McEwan, 1995).

**Educational Accountability**

Over the past 30 years large-scale testing has been the basis for educational accountability (Popham, 1999). According to Kane and Staiger (2002), a school accountability
system is usually comprised of three components: “testing students, reporting to the public on school performance, and rewards or sanctions based on some measure of school performance or improvement” (p. 92). In Ontario, the testing initiative is meant to “evaluate, report and improve the performance of students” and to “make the education system more responsible to the public it serves” (Ontario Ministry of Education and Training, 1995, p.1). Essentially education stakeholders, like taxpayers, want information related to how well schools are doing in comparison to the expectations set forth in the Ontario curriculum to demonstrate that they are “getting reasonable value for their educational dollar” (Leithwood, Edge & Jantzi, 1999, p. 9). Information is collected for improvement of individual student learning and education as a whole; it is also meant for teachers and parents to use to improve learning for all students (EQAO 2004). The reports published by EQAO can be used by parents, educators, policymakers and boards of education to monitor the effectiveness of the education system over time (EQAO 2004) and support improvement planning by schools, school boards and the Ontario Ministry of Education (EQAO, 2011). In Ontario, schools are expected to have the majority of students achieving at a proficiency level of 3 (out of 4) or higher; more specifically, the Ontario government in 2004, set a goal on province wide testing for reading, writing, and mathematics that at least 75% of all grade 6 students would achieve level three standard by 2008 (Johnson, 2005; Levin, 2007).

Large-Scale Assessments in Canada

In Canada, during the 1980’s and 1990’s, the number and types of large-scale assessment programs increased (Earl, 1999; Klinger, DeLuca & Miller, 2008). For instance, every province and territory has at least one provincially administered large-scale assessment program (Table 2.1). Although each province and territory is in charge of the development of curriculum and the
assessment of student achievement, large-scale assessments have become central in influencing instruction, curriculum and policy in Canadian education (Klinger, DeLuca, & Miller, 2008). In this study, the focus is on the primary and junior assessment programs in Ontario and in order better to understand these it is useful to review the history of large-scale assessments in Ontario.

**Large-Scale Assessments in Ontario**

Ontario was one of the first provinces to administer entrance examinations for secondary school (Nagy, 2000). In the 1950’s standardized exit examinations were administered to Grade 13 students (Royal Commission on Learning, 1994) to determine final school marks, graduation, and for admission to university (Earl, 1995). This process, however, changed in the 1960’s and a combination of marks from exams and scores from teachers were used instead of relying solely on provincial assessments (Klinger, DeLuca, & Miller, 2008). In the next two decades, when the United States and other Canadian provinces were intensifying assessments, Ontario did not; assessments were done at the district level and headed by educators based on curriculum guidelines as per provincial policy documents (Earl, 1995). The focus was on teacher assessments of student achievements (Klinger, DeLuca, & Miller, 2008) as the teacher’s judgment, with the appropriate evaluation techniques, was viewed as the most effective form of evaluation (Royal Commission on Learning, 1994).
Table 2.1

*Overview of Assessment Programs Across Canada by Province/Territory*

<table>
<thead>
<tr>
<th>Province/Territory</th>
<th>K-2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alberta</td>
<td>LM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>British Columbia</td>
<td>LM</td>
<td>LM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LMS</td>
<td>OA</td>
<td>AL</td>
</tr>
<tr>
<td>Manitoba</td>
<td>LM</td>
<td>M</td>
<td>M</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ALM</td>
<td></td>
</tr>
<tr>
<td>New Brunswick-Anglophone</td>
<td>L</td>
<td>L</td>
<td>M</td>
<td>L</td>
<td>M</td>
<td>L</td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>New Brunswick-Francophone</td>
<td>L(K)L(2)</td>
<td>M</td>
<td>LMS</td>
<td>MS</td>
<td>LS</td>
<td>LMOA</td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>Newfoundland &amp; Labrador</td>
<td>LM</td>
<td>L</td>
<td></td>
<td></td>
<td>LM</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northwest Territories</td>
<td>LM</td>
<td></td>
<td></td>
<td></td>
<td>LM</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nova Scotia</td>
<td>LM</td>
<td>L</td>
<td></td>
<td></td>
<td>L</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nunavut</td>
<td>LM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ontario</td>
<td>LM</td>
<td>LM</td>
<td></td>
<td></td>
<td>M</td>
<td>L</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prince Edward Island</td>
<td>LM</td>
<td>LM</td>
<td></td>
<td></td>
<td>LM</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quebec</td>
<td>LM</td>
<td></td>
<td></td>
<td></td>
<td>LM</td>
<td></td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saskatchewan</td>
<td>L</td>
<td>LM</td>
<td></td>
<td></td>
<td>LM</td>
<td>L</td>
<td>ALM</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yukon</td>
<td>LM</td>
<td>LM</td>
<td></td>
<td></td>
<td>LM</td>
<td>LMO</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A - Academic; L - Language; M - Mathematics; O - Social Science; S - Science, K - Kindergarten, 2 - Grade 2 (Adapted from Klinger, DeLuca, & Miller, 2008)
There were many allegations that many students at the end of their schooling did not possess adequate skills for entry into the work world, so assessment of student achievement became a top priority in the agenda of accountability (Earl, 1995). The concern for the quality of education was initially rooted in the political and economic changes influencing the 1980’s and 1990’s in Ontario and worldwide (Adams & Kirst, 1999; Earl, 1995; Klinger, DeLuca & Miller, 2008); the public wanted more information on the efficiency of the school system and educators and policy-makers required more information to inform their decisions to improve the quality of education (Earl, 1995). There was also an increased interest in participating in and developing large-scale assessments in Ontario in the 1990’s (Bartley & Lawson, 1999). Ontario joined other Canadian provinces in struggling with the issues of assessment and accountability (Earl, 1995). The creation of EQAO was recommended by the Royal Commission on Learning in 1994 for this very purpose.

**Education Quality and Accountability Office (EQAO)**

EQAO, an independent agency of the Ontario government, was developed in 1995 after the Ontario Royal Commission on Learning had consulted considerably with parents, students, educators and taxpayers of Ontario. The commission concluded that large-scale assessments in Ontario would help appease public demand for greater quality and accountability in the publicly funded school system. EQAO, in response, has developed large-scale assessment programs for the province of Ontario for grades 3, 6, 9, and 10; the Grade 3 and Grade 6 Literacy and Numeracy Assessment, the Grade 9 Numeracy Assessment, and the Ontario Secondary School Literacy Test respectively (EQAO, 2011). EQAO’s responsibilities include the scoring, reporting, constructing, and administering of the assessment programs (Royal Commission on
Learning, 1994) as well as coordinating Ontario’s participation in other large-scale assessments such as PISA and PCAP (Volante, 2007). More specifically, the mandate of EQAO is to:

- Develop tests, and manage their administration in cooperation with school boards;
- Evaluate the testing results and report them to the public;
- Make recommendations to the government to improve the quality of education;
- Manage Ontario’s participation in international testing;
- Conduct research on best classroom practices;
- Collect quantitative and qualitative data to evaluate the effectiveness of the educational system;
- Report publicly each year on testing results and related system evaluation. (Leithwood, Edge & Jantzi 1999, p. 113).

Assessments are developed by EQAO with teachers and private contractors (Klinger, DeLuca, Miller, 2008). The purpose of EQAO testing in Grades 3 and 6 is to report yearly on the extent to which students are meeting curriculum expectations in reading, writing, and mathematics, and to aid in improvement planning, target setting of schools, and curriculum planning (Klinger, DeLuca, Miller, 2008).

Score Reporting

According to Braun, Chudowsky and Koenig (2010), different types of “test-based evaluation models” are used to answer different questions applicable to policy (p. 3). For instance, status models provide an overview of average student performance which may be compared to a target of expected achievement (Braun et al., 2010). A status model is used to assess the percentage of students achieving the provincial standard in a given year. Cohort-to-cohort change models address the degree of change in test results, for a school, at two time
points. Since this model involves comparing two different groups of students, policy questions using this type of evaluation model would relate to comparing the performance of one group of grade 3 students one year to another group of grade 3 students, at the same school, the following year (Braun et al., 2010).

In Ontario, at the school level, achievement on EQAO testing is reported as the percentage of proficient students (PPS). The percentage of proficient students became a popular way of reporting scores using the National Assessment of Educational Progress (NAEP) in the 1990’s (Rothstein, Jacobsen, & Wilder, 2006) and has been the primary way of gauging decisions related to school accountability under the No Child Left Behind (NCLB) Act. Ho (2008) explains that there are costs, statistically and otherwise, to focusing on proficiency by only offering a limited and unrepresentative snapshot of large-scale test scores and gap trends. Ho (2008) argues that PPS represents a goal and “encourages higher order interpretations about the progress of students and schools that are limiting and often inaccurate” (p.351). Relying heavily on proficiency as a way of reporting scores leads to overly sensitive statistics and policy responses for those students near the proficiency cut score. This approach of reporting proficient students is especially sensitive, to small changes in the cut score, when comparing results between groups or across years (Ho, 2008). By reducing distributional change, in education, to a single number, such as PPS, much of the information is missed (Barton & Coley, 2008; Ho, 2008). Ho (2008) explains that if “a single summary statistic is to be chosen as a foundation for reporting, it should be the average” (p. 356). Barton and Coley (2008) argue that although using the average is useful since it is derived from the scores of all students, averages provide little information about the achievement of the population at different points along the score distribution and can cover up important differences within a population.
Unlike Ontario, states may vary under the NCLB Act: (a) the trajectories they select to move from the baseline percent proficient or above in 2002 to the 100% proficient goal in 2014 and (b) the minimum number of students required for reporting of disaggregated subgroup results (Porter, Linn & Trimble, 2005, p. 32). The influence of these design changes can mean not meeting Adequate Yearly Progress (AYP) in some cases (Porter, Linn & Trimble, 2005). For example, different states with the same data can be seen having very different trajectories due to the differences in each state’s proficiency cut score (Ho, 2008). There are 5 levels to the argument that Ho (2008) poses where the choice made at Level 1 is seen to impact, at the higher levels, interpretations of equity, progress and policy. The five levels are:

Proficiency cut scores are judgmental; Trend and gap magnitudes depend on proficiency cut scores; Gap trend magnitudes depend dramatically on proficiency cut scores; These dependencies are not straight forward; and These dependencies may not be used, cynically or nobly, as policy tools (Ho, 2008, p. 352).

In Ontario, there are four achievement levels as defined in The Ontario Curriculum which are used by EQAO to report student achievement in reading, writing and mathematics (EQAO, 2011). The provincial standard of achievement is Level 3; students performing below the provincial standard are achieving at Level 1 or Level 2 and students achieving at Level 4 indicate above provincial standard achievement. Some students may fall into three additional reporting categories: exempt; not enough evidence for Level 1 (NE1); and no data (EQAO, 2011). EQAO scores are measured on the basis of provincial standards where all schools are encouraged, in some cases, by providing needed resources or programs to increase the proportion of students reaching Level 3. Reporting the percentage of proficient students only reveals information about students who are just at or above the cut point (e.g. the part of the score distribution where the
change is reported from year to year), but by doing this information about students further down
the achievement distribution is neglected (Barton & Coley, 2008). Furthermore, two schools
having the same PPS may mean two completely different distributions. For example, as is seen
in Figure 2.1, both schools have 67 percent of students in their school achieving at the provincial
standard or higher, by examination one can observe there to be clear differences in the
percentage of students in each EQAO achievement category which would otherwise go
unrecognized. Ram and Levin (2011) explain that using one indicator - the percentage of
proficient students - in the public analysis of the EQAO provincial results can be potentially
misleading. “In recent years, the proportion of Ontario students at Level 1, a very low level, has
fallen by more than 50 percent. This is as important as the increase at Level 3, yet goes virtually
unreported.” (Ram & Levin, 2011, p. 12)

![Figure 2.1. Distribution of the number of students in each EQAO achievement category using two schools with the same percentage of proficient students of 67.](image-url)
Misuse of Test Results

Rating schools in order to rank them based on yearly standardized test scores and making this information public is a hotly debated issue in education. The initiation of province wide testing, by the Ontario Ministry of Education and Training, was to collect student achievement data and make improvements to the curriculum; it was not to evaluate specific schools or school boards or rank schools (Earl, 1995; EQAO, 2000). Tests, according to Pascal (2005) “are not designed, nor are they useful, for ranking and comparing all individual schools or school boards” (p. 1) and “ultimately harmful to the health of our education system” (EQAO, 2000, p.2). Through ranking, people may make simplistic comparisons or generalized conclusions without considering the factors affecting achievement (EQAO, 2000). Every school may encounter challenges that are specific to students in that school and this may have a drastic impact on school results (Taylor, 2007). For example, schools with fewer students taking the test may lead to high variance from year to year having a big impact on ranking (Linn & Haug, 2002). Rankings may exaggerate small differences where many schools have very similar results (Ho, 2008). It is common practice for schools to be ranked according to results, by the press, and conduct school by school comparisons sometimes leading schools to defend and validate their results (Simner, 2000).

The Fraser Institute, well known for ranking schools in Canada (Cowley & Easton, 2003), “have been seriously questioned for their failure to adequately take into account school context” when ranking (Rogers & Klinger, 2005, p. 28). Johnson (2005) argues that presenting school level results without context about the school may lead parents to place their children in schools that rank high, but perform poorly in relation to schools with similar socio-economic background; for instance, not understanding “how uneven the playing field is” (p. 215). Parents
selecting their child’s school based on rankings, when often rankings do not provide the full picture, may also lead to problems with student enrollment at schools with low rankings (Taylor, 2007). Many states use results from high stakes testing to assess students, teachers, and schools and make judgments solely based on test scores showing that much of the emphasis is put on the product and not the process of how students learned to read or whether they could already read (Smith & Smith, 2005).

**Error, reliability and validity**

It is important to acknowledge that error is inevitable in large-scale assessments (Briggs, 2011; Ho, 2008; Hollingshead & Childs, 2011; Kane & Staiger, 2002; Kane, Staiger, & Geppert, 2001; Linn & Haug, 2002; Milanowski, 1999; Wu, 2010). According to Wu (2010), there are three main types of error associated with large-scale assessments: Measurement error; Sampling error; and Equating error. These errors affect both the student and school levels differently. At the student level, measurement error is linked to inaccuracies associated with the test and is usually large when reporting individual student results due to basing results on one test. Also, fewer items on a test may lead to a larger measurement error (Wu, 2010). According to Trochim (2006), the true test score theory of a student’s score on a test (X) is made up of the true score (T) and error (e):

\[ X = T + e \]

The error can be broken down further into two types of error that may influence a student’s test scores: random error (er) and systematic error (es) making the equation:

\[ X = T + e_r + e_s \]

Random error influences observed scores up or down for individuals randomly and may be caused by individual differences in mood on testing day. Systematic error is sometimes
viewed as a bias in measurement because it is caused by factors that influence the measurement of a variable systematically. For instance, a loud disturbance just outside the testing environment would affect the performance of all students and as a result, systematically lowering all scores. When averaging scores at the school level, for instance, random error only influences the variability around the average score; in other words, the sum of random errors in a distribution would be zero as there would be as many negative errors as positive ones, ideally. While systematic error affects the average score (Trochim, 2006). In other measurement literature, random and systematic measurement errors may sometimes be referred to as non-persistent and persistent errors respectively (e.g. Kane & Staiger, 2002).

The degree of measurement error is also dependent upon the validity and reliability of the test (Trochim, 2006). The reliability of a test measures how accurately the test assesses the construct(s) and whether a similar score would be produced if that student was to write the test again (Thorndike, 2005). Factors that may influence the reliability of the test include variability of individual students and the number of items on the test (Thorndike, 2005). According to Wolfe, Childs and Elgie (2004) one large difficulty with the Ontario EQAO assessment is the limited number of test items. A high reliability coefficient and low standard error of measurement (information included in EQAO’s technical report), for instance, indicate a more reliable assessment (Thorndike, 2005). Highly reliable assessments are necessary in order to make useful decisions based on the results (Thorndike, 2005).

EQAO administers criterion-referenced assessments that align with the provincial curriculum (Volante, 2007). Criterion-referenced tests are traditionally focused on a specific range of information being taught in a program of instruction (Thorndike, 2005); this type of test is designed to compare an individual’s knowledge and skills to a provincial standard (Rankin,
2003). A norm-referenced test on the other hand, is used to compare the test score of one person to that of a larger group of similar age or grade level (Nagy, 2000). According to Nagy (2000), “public thinking is norm-referenced, despite the criterion-referenced nature of performance assessments.” (p. 275).

Validity in another key component, alongside reliability, that helps to determine whether a test measures what it is intended to measure and that relevant inferences may be made from the test scores (Thorndike, 2005). Ensuring test items are not too easy or too hard to avoid ceiling and floor effects respectively (Briggs, 2011) (in other words test items and content) and the type of test being administered (e.g. multiple-choice, open response) are related to the validity of a test (Thorndike, 2005). More specifically, construct validity which is “the extent to which the test may be said to measure a theoretical construct or trait” (Anastasi & Urbina, 1997, p. 126) has two threats: construct underrepresentation and construct irrelevant test variance (Thorndike, 2005). Construct underrepresentation relates to leaving something out of the test that theoretically should be included making the test narrow and the construct irrelevance test variance relates to including something in the test that is theoretically irrelevant. Both threats to construct validity reduce the ability to accurately interpret and use test scores (Thorndike, 2005). Also, content validity, which is important for assessing the validity of criterion-referenced tests “relates to the test score and all of the factors that affect it, including clarity of directions and adequacy of scoring procedures.” (p. 192). For instance, if very little time is provided to students to attempt all of the items on the test, then the speed of performance becomes part of the content of the test (Thorndike, 2005). Overall, if the error is too great, the validity and reliability of the measure become questionable and interpretations from analyses conducted with such data may be skewed (Trochim, 2006).
Both typical measurement error and errors in standard setting are involved in EQAO testing. Initially, Brennan and Lockwood (1980) examined the relationship between both errors of measurement and errors in standard setting related to item effect as uncorrelated; however, Kane and Wilson (1984) suggest that the two errors may be correlated. Kane and Wilson (1984) found the covariance of these two errors were verification for the validity of interpretation for criterion-referenced tests and standard setting procedures. In other words, a negative covariance value would suggest that the “criteria being used by judges to set the cutoff score are not consistent with the attribute being measured by the items” and a positive value would decrease the probability of misclassification of students due to a decrease in total error variance (Kane & Wilson, 1984, p. 114). Ideally, there should be no misclassifications (Thorndike, 2005).

According to Wu (2010), measurement error is necessary to consider when reporting individual student results; however, it may not be as great a concern when averaging scores to report school level results because the average includes a larger number of students (Wu, 2010). As a result, sampling error would be of more concern when examining scores at the school level. “The key distinction here is that measurement error at the student level is assumed to have a functional relationship with the number of test items that students have been administered; at the teacher or school level, measurement error is assumed to have a functional relationship with the number of students.” (Briggs, 2011, p. 31). Fluctuations in school achievement scores over time may reflect “real” improvement or decline in school performance as well as random variations (Willms & Raudenbush, 1989). For example, Briggs (2011) states that “regardless of the quality of instruction to which they are exposed, it may be the case that some cohorts of students are simply “better” or “worse” than others” (p. 31). According to Kane and Staiger (2002), however, differences in test scores from year to year are in most cases due to non-persistent factors such as
a new cohort of students associated with testing each year (Kane, Staiger, & Geppert, 2001; Kane & Staiger, 2002). On average, if errors are random, an increase in the number of students included in the average may lead to a more accurate score (Wu, 2010). Similarly, scores varying from year to year may be a result of two main sources: measurement error and sampling error (testing a different cohort each year) where sampling error is much more instrumental in fluctuations of school scores than measurement error (Hill and DePascale, 2003).

Lastly, equating allows for results from different assessment administrations to be directly compared by aligning the results (Wu, 2010). According to Wu (2010), the error associated with equating, in some cases, may outweigh both measurement and sampling errors because equating error impacts both the student and school levels. These sources of error are important to recognize to ensure scores are accurately interpreted. All of these errors may be compounded with other sources of error, such as the lack of consistency in test administration procedures, marker reliability or inter-rater reliability where teachers may be receiving incorrect information for their students’ responses, item and test bias against particular groups of students (Wu, 2010), teaching to the test (Popham, 2001) and the appropriateness of the statistical analyses carried out (Wu, 2010); however, choosing to use PPS as a reporting method may inflate these errors (Ho, 2008) “even if the true value were actually constant over time.” (Briggs, 2011, p.31). These types of testing related concerns are necessary to address in order to use results to claim that student learning is improving (Volante, 2007).

**Fluctuations in Achievement Scores**

For schools to implement school improvement and planning recommendations efficiently based on large-scale assessments the data should be reliable and a high degree of consistency is required for school scores to be used (Sicoly, 2002). The issue of consistency, or reliability, of
yearly school level scores on any measure is an important one and is cited across multiple bodies of literature. Although there is mention of cross-year stability coefficients in some school effectiveness literature (e.g., Crone, Lang, Good & Brophy, 1986; Ma, 2001, Mandeville & Anderson, 1987; Teddlie, & Franklin, 1995) there is little mention of student performance in the context of large-scale assessments (Sicoly, 2002) as most school effectiveness studies of unidimensionality examine stability of school effects (the relationship between schooling inputs and student outcomes) (Willms & Raudenbush, 1989) (e.g., Kyrakides & Creemers, 2008; Luyten, 1994; Teddlie & Stringfield, 1993; Willms & Raudenbush, 1989).

Considering a different cohort of students, with varying abilities, are placed in schools with differing resources and challenges specific to each school the outcome may be to anticipate changes in school level test results from year to year. Yearly test scores for elementary schools in Washington, D.C. in 2009 showed that some schools had an increase in PPS while others schools had a decrease by at least 10 points on the D.C. Comprehensive Assessment (Anderson & Brown, 2009). These types of school to school and year to year changes may make it difficult for some principals to “instil a culture of student achievement in many schools where it has long been missing” (Anderson & Brown, 2009, p. 1). Similarly, in an examination of student progress in England, results showed that of “schools below the target in English in 2002, 40% were above the floor target in 2003. Of those above in 2002, 12% were then below in 2003.” (National Statistics, 2004, p. 3). Haney (2002) also observed decreases in the percentage of proficient students in one year after gains in the previous school year when examining mathematics test score averages and change scores for 977 schools in Massachusetts writing the grade 4 Massachusetts Comprehensive Assessment System (MCAS). For the years 1998 to 2001, test score gains in one testing period were “often” followed by losses in the next; although, there was
a general decrease in the average change scores, the difference was highest between the first two years and least for the last two years used in the analysis. At the school level, changes in score averages ranged from -22 to +25 and the largest range of scores was between 2000 and 2001 (-22 to 18). The correlation of school mathematics test scores across years were very high at 0.86 (1998 to 1999) and 0.87 (1999 to 2000 and 2000 to 2001) (Haney, 2002) when compared to results for some regions including schools in Ontario in the following study by Sicoly (2002).

For a more extensive analysis of school level fluctuations, Sicoly (2002) examined the stability of PPS writing large-scale assessments using data from 21 states and 2 provinces (Ontario and Alberta). Assessment results were based on two consecutive years between 1997 and 2001 and the number of schools in the jurisdictions in the sample varied from 124 to 3316. Data from EQAO in Ontario was comprised of school level reading, writing and mathematics scores from 390 schools with results from grade 3 students and 315 schools with results from grade 6 students in the Toronto District School Board for the years 1999 and 2000. Data from the Alberta Achievement Testing Program included mathematics, reading, and writing scores from 305 grade 3 schools and 278 grade 6 schools for the years 1997 and 1998. Results of cross-year stability coefficients, across subjects and grades, varied from 0.41 to 0.94. Average cross-year stability coefficients across subjects and grades were lowest for Vermont (0.51), Maine (0.61), Alberta (0.63), and Ontario (0.65) and highest for Georgia (0.81) and Maryland (0.81). The median stability coefficient for mathematics and reading across all systems was 0.78, while only 44 percent of the reading stability coefficients were 0.70 or higher – the median correlation coefficients were obtained by each jurisdiction contributing one coefficient per subject and an average was calculated in jurisdictions where the same subject was assessed more than once.

---

1Toronto data: Coefficients for grade 3 reading, writing and mathematics were 0.62, 0.58 and 0.64 respectively. Coefficients for grade 6 reading, writing and mathematics were 0.72, 0.69 and 0.66 respectively.
Interestingly, the researcher found that some test types proved to have more stable scores than others; higher cross-year stability coefficients were demonstrated by some states that had students write norm referenced, performance based, and combined multiple-choice and constructed responses tests (Sicoly, 2002). Not considering the size of the cohort writing the test, from year to year, in relation to cross-stability coefficients, is seen as a major flaw in the Sicoly (2002) study.

Another approach to examining school level fluctuations was to use change scores instead of PPS as was done by Sicoly (2002). Using grade 4 reading achievement data from The Colorado Student Assessment Program for the years 1997 to 2000 Linn and Haug (2002) found on average, across all four years, more than 50 percent of the students were in the proficient category (3rd of 4 categories) or higher. The yearly percentage of students who made gains in the proficient or advanced levels varied from school to school. Correlation coefficients for school performance were calculated across the four school years and ranged from 0.796 to 0.837. In other words schools doing well one year were likely to do well the following year; however, schools with the highest proportion of students (80 percent) scoring proficient or advanced in the first year of the analysis were more likely to have smaller gains than those schools with a lower percentage of proficient or advanced students (10 percent) in the first year. The authors examined this by correlating PPS in one year with change scores from the following year and found correlations to range from -0.35 to -0.23 indicating that a high PPS in 1997, for instance, was related to lower change scores between 1997 and 1998 (Linn & Haug, 2002). A more probable association to changes in PPS, seen from year to year, may be a result of a new cohort every year. By conducting correlations of change scores, Linn and Haug (2002) further explain that schools that gain from the first year to the second year will most likely show a decline in the
third year just as schools which show a decline from the first year to the second year will show a gain in the third. In the latter case, schools may be identified as needing assistance, but will likely have gain scores in the third year therefore making the assistance provided to these schools appear more successful than is actually the case (Linn & Haug, 2002). The degree to which scores will decline or increase depend on the specific school and data provided, but the implication the authors make are that there are fluctuations in school scores from year to year. Knowing the size of a gain or loss in percent of students scoring either proficient or advanced, from 1997 to 1999, does not relate to change from 1998 to 2000. The authors explain that the large amount of the variability in school change scores (not quantified in this study) relates to “noise” and schools making a big shift in achievement in one change cycle are not likely to repeat their performance in the next cycle. The “noise” observed in year-to-year changes in the percentage of students scoring at the proficient or advanced level is more prominent in smaller schools (Linn & Haug, 2002).

Limitations of Literature

During the literature search it became apparent that very few studies have examined school level data in the way that is done in this study. There were very few studies that were directly related to the research questions posed in this study exemplifying the gap that exists in explaining school level scores from year to year and the importance of conducting such research. The number of studies examining stability of results rooted in school effectiveness and school improvement research has grown over the past few decades. However, most examine the results of high school students from the US or UK using a value-added approach to examine both student and school level achievement and focus on the overall improvement of achievement or the rate of improvement and less on the variability (e.g., Benikson, Hattie & Robinson, 2010;
Gray, Goldstein & Jesson, 1996; Gray, Goldstein & Thomas, 2001; Thomas, Peng & Gray, 2007). Longitudinal studies, on the other hand, may neglect information about the years in between (e.g., Doolaard, 2002), which are just as important, by choosing to examine two years many years apart. There still remains, however, a certain amount of uncertainty regarding what is viewed as “normal” changes in PPS from year to year, which is a gap in the literature that this study aims to address.

Sicoly (2002) argued that cross-year stability coefficients across 2 consecutive years provide an efficient way to evaluate the overall reliability of school level data from large-scale assessments. Using a stability coefficient of 0.80 as the standard for large-scale assessments of student performance is recommended by Sicoly (2002), while Willms and Raudenbush (1989) provide a range of 0.60 to 0.90 as being “relatively stable” and correlations lower than 0.30 as “small”. Doolaard (2002) questions the interpretation of correlations between years as there is not any “criterion to discern stability from instability” (p. 774). Stability coefficients used in some of the studies reviewed are useful in providing an overview of achievement across years for all schools, but provide little information for individual schools, such as the direction and degree of change in achievement from year-to-year. For example, a correlation of 0.80 would indicate that if PPS is high in 2009 PPS is likely to be high in 2010 approximately 64 percent of the time. The amount of change in PPS cannot be inferred. An individual school may have an increase in PPS by 10 percent in one year while another school may have a decrease in PPS by 20 percent – this type of information is not available from correlations.

The assumption in earlier school effectiveness literature was that of unidimensionality or that a school had consistent or stable results from one year to the next year, or in another class, subject, or grade, but this may not always be the case as a result of the different analysis methods
that have been used in studies (Doolaard, 2002; Scheerens & Bosker, 1997) and variable assessments which may be imperfect, with differing content and formats which may lead to inconsistencies between studies. Studies use a variety of datasets paired with several methodologies across different time points which may make it difficult to compare and interpret results. The question then becomes, why is consistency assumed to be present in schools with changing cohorts with varying ability? Accordingly, should there not be an expectation of change more than consistency?

**Influence of Cohort Size on Achievement Scores**

Many researchers have attributed smaller sample sizes to be correlated with increased variability in achievement scores across years (Haney, 2002; Linn & Haug, 2002; Kane & Staiger, 2002; Turque, 2009). In other words, fewer students taking the test can lead to high variance in school results from year to year and thus have a big impact on school ranking (The Alberta Teachers’ Association, 2006). Most states in the United States, in recent years, have built accountability systems around using school level test scores; however, due to the average elementary school containing approximately 69 students per grade level, such measures may be inaccurate (Kane & Staiger, 2002).

Schools with fewer than 100 students tested per grade, in Massachusetts, showed changes in average scores of up to 15 to 20 points (MCAS scale from 200 to 280) whereas, schools with 150 or more students being tested per grade only showed a 5 point or less change in average scores (Haney, 2002). Furthermore, three of four schools that had the largest gains in the 1998/1999 school year, showed a decline in average grade 4 mathematics scores the following school year (1999/2000). A commonality between these four schools was that they were all small schools each having had less than 100 students tested. The author tracked other schools that
showed a 10 point or greater increase in average school scores one year with significant drops 
the following year and concluded that this was a result of having a small number of students 
tested (Haney, 2002). Linn and Haug (2002) also reported that the variance observed in year-to-
year changes in the percentage of students scoring at the proficient or advanced level was greater 
in smaller schools. For this examination, they split schools into 5 categories (e.g., 30 or fewer, 
31-60, 61-90, 91-120, or 121 or greater) depending upon the number of students in each school. 
Schools with a larger population being tested were less likely to be at either extreme (increases 
or declines) of students scoring proficient or advanced, which was not the case for schools with a 
smaller population being tested (Linn & Haug, 2002).

A similar finding was made by Kane and Staiger (2002) where they found much more 
variance in test scores among smaller schools than among larger schools. Reading and 
mathematics test scores from approximately 300,000 students in North Carolina elementary 
schools (grades 3, 4, and 5) from the 1992/93 school year to the 1998/99 school year were 
matched across the years using student date of birth, race and gender. In examining the 
“volatility” of school level test scores, the researchers took into consideration the sampling 
variability (a different population of students each year), and extraneous factors that may 
influence performance on the test. Variance in mathematics and reading test scores was 
approximately 50 percent larger for the smallest 20 percent of schools than for the largest 20 
percent (Kane & Staiger, 2002). Nash (2002) revisited volatility of test scores by examining 
trends in test score variance from 1998 to 2000 by average enrolment for both grades 3 and 5. 
She found that grades with smaller enrolments showed more instability of test scores. More 
specifically, as enrolment increased for each grade level the standard deviation of change scores 
decreased for both mathematics and English across all three school years. Using the formula
developed by Kane and Staiger (2002), Nash (2002) was able to determine that changes in 
student achievement were attributable to non-persistent factors for both grades 3 and 5 
mathematics (72%, 63%) and grades 3 and 5 English (88%, 70%) scores from one year to the 
next respectively. Morgan (2002), on the other hand, states that with small samples “the fraction 
of students scoring at the Proficient or Advanced level can fluctuate from year-to-year by 10% - 
20% or more even in the absence of other influences.” (p. 3).

The studies reviewed thus far in this section, indicate that cohort size plays an important 
role in the fluctuations in PPS. Hollingshead and Childs (2011) examined this relationship in two 
ways, first using the cohort of students from each school and secondly by randomly selecting 
students from a pool of students to create artificial schools. Applying similar principles as Ho 
(2008, 2009) to grade 6 EQAO data, Hollingshead and Childs (2010), examined the “effect of 
reporting-group size on the stability of the percentage of students achieving at or above a 
standard” using reading, writing, and mathematics scores reported to parents between the years 
2008 and 2009 (p. 36). Their findings showed that the percentage of students scoring at level 3 or 
higher, at a school, was unstable for small schools and small randomly drawn groups 
(Hollingshead and Childs, 2010).

In part 1 of their study, Hollingshead and Childs (2010), examined schools which had 
students complete the grade 6 assessment in reading, writing and mathematics in 2008 and 2009. 
The schools, in 2008, were grouped into categories based on the size of grade 6 classes (e.g., 11- 
20, 21-40, 41-80, 81-160, 161 or more) and the difference in the percentage of students above the 
cut score (PAC) or PPS, between the years 2008 and 2009, was examined (Hollingshead and 
Childs, 2010).
In part two of the Hollingshead and Childs (2011) study, students were selected randomly from the population of grade 6 students to form 10 sets of samples of differing group size (ignoring student characteristics). Each set contained 1000 samples of two equal group sizes (e.g. 15, 30, 60, 120…7680). Results indicated that, in both cases, variance in PAC increased as the size of the group decreased. The differences observed between groups could be attributed to sampling error because the randomly drawn samples were drawn from the same population. In other words: “There is a 5% probability that an observed difference of 33% [in percent reaching level 3] between groups could occur just by chance.” (Hollingshead & Childs, 2011, p. 40) In a typical school, there are 2 classrooms (approximately 60 students) Hollingshead (2010) revealed using PPS to examine achievement over 2 years “there is a 90% probability that a difference of approximately 2% could occur just by chance, and also a 68% probability that a 3% difference could occur just by chance.” (p. 29). One of the potential drawbacks acknowledged by Hollingshead and Childs (2011) was, in their study, randomly drawing samples from Ontario which does not take into consideration that school populations are not randomly drawn and as a result may have more stable results between years (Hollingshead & Childs, 2011).

The purpose of this literature review was to frame the research questions posed in chapter one and to examine the research studies that have been conducted related to this topic. School results have been shown to vary from year to year, and in some cases, significantly. This has been shown to be related to the number of students at a school writing the test. Upon reviewing current literature on changes in school level achievement, it was evident that there was a significant lack of information related to changes in PPS across multiple consecutive years for grade 3 and 6 reading, writing and mathematics. The gap in research, for this topic, implicates
the importance of conducting this research study to provide information on expected changes in PPS and the relationship of cohort size on variance in PPS across multiple consecutive years.
CHAPTER THREE
METHODOLOGY

This chapter will present a description of the data used in this study and expand on the methods used to investigate each research question posed in chapter one.

Research Design

Secondary data analysis. Secondary data analysis (SDA) involves applying statistical techniques to existing data (Kiecolt & Nathan, 1985; Trochim, 2006a). Data may be stored in electronic databases to later access and analyze as may be done in governments, schools, and other organizations. SDA has the many benefits of saving time, money and resources and often allows the scope of the study to be extended considerably (Trochim, 2006a) because it allows researchers to have access to data that would otherwise be difficult for them to collect (Hyman, 1972; Kiecolt & Nathan, 1985).

Secondary data analysis also has some disadvantages: specific variables required by researchers may only be available in aggregate, index or scaled form; researchers may not be aware of coding or data entry errors made in the original data; data collection procedures may not be sufficiently documented making it difficult to determine errors in the data and limiting interpretation of the data; items on instruments used may not be an accurate measure of the constructs the researcher is interested in measuring or they may have been poorly operationalized or there may be too few samples for the desirable statistical analysis (Kiecolt & Nathan, 1985). Some of these limitations, however, can be avoided by using datasets intended for secondary analysis (Kiecolt & Nathan, 1985) such as the EQAO dataset which is used in this study.
Overview of EQAO Dataset

In keeping with the Principles for Fair Student Assessment Practices for Education in Canada (1993), yearly assessments of reading, writing, and mathematics, in both grades 3 and 6, were developed to measure how well students achieve select expectations outlined in The Ontario Curriculum (EQAO 2011). The Literacy and Numeracy Assessment for grades 3 and 6 takes place during a two-week period in late May and early June (EQAO, 2012). The test is administered by the student’s regular teacher. The test is six hours long and consists of three booklets (2 for language and 1 for mathematics) (EQAO, 2012). The assessments contain both open-response and multiple-choice questions (EQAO 2011). Results of the assessment are made public in the fall (EQAO, 2012). “Two sets of results are reported: those based on all students and those based on participating students. Students without data and exempted students are not included in the calculation of results for participating students.” (EQAO, 2011, p. 40).

An administration guide is sent to schools for training teachers to administer EQAO tests including a guide for providing accommodations and support to students with special education needs and English Language Learners (ESL) so that they can fully participate in the assessment (EQAO, 2011). Students are only exempt from participating in the assessment if they are unable to participate in all or part of the assessment even with accommodations or support (e.g. student had to be read to by a teacher or if terms have to be defined) (EQAO, 2011).

To evaluate students’ work, teachers who mark the tests use achievement levels from the Ontario curriculum (Table 3.1). These levels of achievement are generated and used by EQAO to report student results and align with the four levels of the Ontario curriculum expectations put forth by the Ministry of Education (EQAO, 2011). Large-scale assessments in most cases have scores at both the student and school level. In relation to student level scores, there are many
indicators of reliability such as internal consistency measured by Cronbach’s alpha. Given the use of school level scores within various accountability programs a test should be highly reliable. A reliability coefficient between the values of 0.70 and 0.80 indicates modest consistency and values higher than 0.80 indicate a highly reliable test (Lavrakas, 2008). The Cronbach’s alpha for student level scores, on the EQAO assessment range from 0.85 and 0.88 for reading, 0.78 to 0.82 for writing, and 0.87 to 0.89 for mathematics (EQAO, 2011). The standard error of measurement for reading, writing and mathematics is: 3.31, 1.95, and 3.47 for grade 3 and 3.12, 1.98 and 3.77 for grade 6 respectively (EQAO, 2011).

Table 3.1

*Level Descriptions from the Ontario Curriculum*

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>The student has demonstrated the required knowledge and skills. Achievement exceeds the provincial standard.</td>
</tr>
<tr>
<td>3</td>
<td>The student has demonstrated most of the required knowledge and skills. Achievement meets the provincial standard.</td>
</tr>
<tr>
<td>2</td>
<td>The student has demonstrated some of the required knowledge and skills. Achievement approaches the provincial standard.</td>
</tr>
<tr>
<td>1</td>
<td>The student has demonstrated some of the required knowledge and skills in limited ways. Achievement falls much below the provincial standard.</td>
</tr>
<tr>
<td>NE1</td>
<td>Not enough evidence for level 1</td>
</tr>
</tbody>
</table>

Adapted from EQAO (2012).

**Scoring procedure.** The first step in the scoring process referred to as range finding, is to define and illustrate the range of performances within each code of the scoring rubrics (EQAO, 2011). This is done by a committee of educators with expertise in scoring and the education system in Ontario. Anchor papers are identified through coding of student responses to train

---

2 These are for the 2009/2010 school year.
scorers. Before scoring begins, every scorer undergoes extensive training so that each scoring leader, scoring supervisor and scorer is able to interpret and apply the scoring materials similarly. EQAO uses strict scoring procedures to ensure the reliability of assessment results. All responses are scored by trained scorers; multiple-choice items for the primary assessment are keyed in manually by scorers from booklets. Scoring rubrics and anchors are used to score open-response items and writing prompts and each item is scored in relation to its best match with one of the code descriptors in the rubric for the item and its anchors (EQAO, 2011).

**Equating procedure.** Once the tests have been scored, the tests are equated. Test equating adjusts changes in difficulty of the test from year to year (Kolen & Brennan, 2004) since, for security purposes, EQAO constructs a different test every year (EQAO, 2011). Even though EQAO ensures that test content and statistical specifications are similar from year to year, equating ensures that students in one year are not given an unfair advantage over students in another year; therefore, eliminating potential for changes in achievement levels in student performance to be attributed to differences in test difficulty (EQAO, 2011).

**Overview of Curriculum**

Both the grade 3 and grade 6 tests are based on the expectations in the Ontario curriculum which outline the knowledge and skills students should have acquired by the relative stages of their schooling (EQAO, 2011) (Table 3.2).
Table 3.2

Skills Assessed on the Grades 3 and 6 Literacy and Numeracy Assessment

<table>
<thead>
<tr>
<th>Subject</th>
<th>Example of skills assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>• Understand explicit and implicit information and ideas in a variety of text types</td>
</tr>
<tr>
<td></td>
<td>• Make inferences</td>
</tr>
<tr>
<td></td>
<td>• Make connections between own personal knowledge and experience with what they read.</td>
</tr>
<tr>
<td>Writing</td>
<td>• Understand assigned tasks</td>
</tr>
<tr>
<td></td>
<td>• Communicate with the reader</td>
</tr>
<tr>
<td></td>
<td>• The use of correct spelling, grammar and punctuation</td>
</tr>
<tr>
<td></td>
<td>• Organize main ideas and supporting details</td>
</tr>
<tr>
<td>Mathematics</td>
<td>• Knowledge and skill across five mathematical strands</td>
</tr>
<tr>
<td></td>
<td>o number sense and numeration</td>
</tr>
<tr>
<td></td>
<td>o geometry and spatial sense</td>
</tr>
<tr>
<td></td>
<td>o measurement</td>
</tr>
<tr>
<td></td>
<td>o patterning and algebra</td>
</tr>
<tr>
<td></td>
<td>o data management and probability</td>
</tr>
</tbody>
</table>

(EQAO, 2011)

Population

The population for this study includes all elementary schools in Ontario participating in the English version of the EQAO assessments from 2006 to 2010.

Outcome measure

The outcome measures pertain to the percentage of proficient students (PPS) at a school in reading, writing, and mathematics for both grades 3 and 6 from 2006 to 2010 – that is, the percentage of students scoring at level 3 or higher. When generating PPS for each school, from student records, students who were in the exempt or missing category were not included in
analyses, however, students in the “not enough evidence” category were included. Outliers in PPS were identified (i.e. absolute PPS z-score value of greater than 3) and deleted. Changes in PPS from year to year and variance scores were then calculated using PPS and were also examined in this study. In order to examine fluctuations in PPS, a variance score was computed which measures how much scores deviate from year to year within a school across 5 years, in this case, and was calculated for each subject in each grade from the years 2006 to 2010 using the formula $\sigma^2 = \sum(X - \mu)^2/N$.

Analysis

**Organization of data.** Data were obtained from EQAO by submitting an application for data access acquired from the EQAO website. Once access was granted, files were received in excel format with a different dataset for each year ranging from 2006 to 2010. The files were then converted into SPSS format and all individual datasets were then merged into one large file. Schools with fewer than 5 years of data in reading, writing, or mathematics were omitted. Table 3.3 displays the final sample sizes by subject and grade and the number of outliers omitted from the study. The larger the population, the higher the z-score criterion is for excluding an observation being an outlier (Cosineau & Chartier, 2010). In order to reduce the effect of extreme scores, PPS scores were standardized and any z-score values greater or less than 3.0 were classified as outliers and omitted. In other words, schools with a score that was above 99.6 percent and below 0.4 percent of the population was omitted.
Table 3.3

*Number of Schools and Outliers in Each Subject and Grade*

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Number of schools</th>
<th>Number of outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>3</td>
<td>2918</td>
<td>40</td>
</tr>
<tr>
<td>Writing</td>
<td>3</td>
<td>2908</td>
<td>52</td>
</tr>
<tr>
<td>Mathematics</td>
<td>3</td>
<td>2923</td>
<td>58</td>
</tr>
<tr>
<td>Reading</td>
<td>6</td>
<td>2829</td>
<td>45</td>
</tr>
<tr>
<td>Writing</td>
<td>6</td>
<td>2830</td>
<td>45</td>
</tr>
<tr>
<td>Mathematics</td>
<td>6</td>
<td>2856</td>
<td>17</td>
</tr>
</tbody>
</table>

**Analytic Approach**

This study focused on the examination, description, correlation and clarification of PPS in grade 3 and 6 reading, writing and mathematics from 2006 to 2010. For the first part of the study, descriptive statistics and frequencies were the main method of examination. In the second part of the study, variance scores and correlations were used in order to better understand the relationship between changes in PPS and cohort size.

**Descriptive statistics.** Descriptive statistics, such as the mean, standard deviation, minimum and maximum values and frequencies provide information about the distribution of data points and it becomes easier to identify any outliers or variables with unusual characteristics which can then be addressed. This study includes descriptive statistics for PPS, annual change, variance and the average number of students in each cohort. Annual change was calculated by subtracting PPS for adjacent years (e.g., Year2 PPS – Year1 PPS) where a positive value indicates an increase in PPS in the following year. When examining the frequency of change, absolute annual change was used because the main concern was the magnitude of change.
In order to examine whether the way PPS changed from year to year, at a school, was specific to a grade level or subject, changes in PPS were coded; for instance, an increase in PPS from 2006 to 2007 was coded as ‘1’ and a decrease as a ‘0’. These values were then merged to form a pattern of change (increase/decrease) across all 5 years for each subject and grade. For example, ‘0111’ indicates a decrease in PPS from the first to the second year and an increase in PPS across all other years and ‘1000’ would indicate the exact opposite.

**Correlational analyses.** A correlation describes the amount of variance that is shared between two variables. In this study PPS was correlated across all 5 years. Variance scores were correlated with the size of the cohort writing the assessment for grades 3 and 6 reading, writing and mathematics. The mean size of the cohort across 5 years was used when correlating cohort size with the variance scores. Correlations were also used to determine the strength of the relationship in PPS across subjects and grades.
CHAPTER FOUR

RESULTS

The initial dataset from EQAO contained data for 3935 schools. Once schools writing the French version of the assessment, those schools not having data for 5 consecutive years and outliers were omitted from the analysis there were slightly fewer than 3000 schools for each year, grade level and subject. The following sections are organized by subject (reading, writing, and mathematics) and include results for both grades 3 and 6 in each section. To facilitate answering the research questions results are presented accordingly: Descriptive results; Degree and frequency of changes in PPS; and Variance in PPS and average cohort enrolment.

Descriptive Results

The descriptive results section contains means, standard deviations, minimum and maximum values for achievement scores, annual change in achievement and correlations for achievement scores. The results, for both grades 3 and 6 are presented in the order of reading, writing, and mathematics (Please see Appendices A to F and Appendices G to L for PPS and annual change distribution histograms, respectively).

The following subsections pertain to descriptive results for the reading, writing and mathematics portions of the EQAO assessment. School level PPS and annual change in PPS for both grades 3 and 6 from the years 2006 to 2010 are presented and compared for each subject.

Reading. Average reading PPS for grade 3 reading was highest for 2006 and decreased slightly from then onwards and increased again in 2010 (Table 4.1). The standard deviation was lowest for 2006 indicating the spread of scores was not as large as it was for 2009. The average reading PPS for grade 6 was highest in 2010 and has steadily increased from 2007 and the standard deviation has steadily declined. Overall, the reading score means were higher and the
standard deviations were lower for grade 6 than for grade 3. The provincial PPS was lower and more stable for both grades 3 and 6 than PPS means.

Table 4.1

*Descriptive Statistics for Grade 3 and 6 Reading PPS*

<table>
<thead>
<tr>
<th>Grade 3</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provincial PPS</td>
<td>62</td>
<td>62</td>
<td>61</td>
<td>61</td>
<td>62</td>
</tr>
<tr>
<td>Mean</td>
<td>66.73</td>
<td>65.32</td>
<td>64.74</td>
<td>63.25</td>
<td>64.34</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>14.76</td>
<td>15.20</td>
<td>15.32</td>
<td>16.09</td>
<td>15.29</td>
</tr>
<tr>
<td>Minimum</td>
<td>21.43</td>
<td>21.21</td>
<td>19.23</td>
<td>15.00</td>
<td>17.24</td>
</tr>
<tr>
<td>Maximum</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade 6</th>
<th>64</th>
<th>64</th>
<th>66</th>
<th>68</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provincial PPS</td>
<td>64</td>
<td>64</td>
<td>66</td>
<td>68</td>
<td>70</td>
</tr>
<tr>
<td>Mean</td>
<td>68.02</td>
<td>67.54</td>
<td>69.12</td>
<td>71.95</td>
<td>73.51</td>
</tr>
<tr>
<td>Minimum</td>
<td>23.33</td>
<td>23.53</td>
<td>25.00</td>
<td>28.00</td>
<td>30.56</td>
</tr>
<tr>
<td>Maximum</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

The average change in school PPS, for grade 3 reading, was negative across all years except for 2010 (Table 4.2). The average annual change in PPS, for grade 6 reading, was positive across all years except for 2007. The standard deviations for annual change in grade 6 reading was lower than for grade 3 as was the range of scores. Overall, the annual change averages were higher for grade 6 than for grade 3.
Table 4.2

*Descriptive Statistics of Annual Changes in PPS for Grade 3 and 6 Reading*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-1.42</td>
<td>-0.58</td>
<td>-1.49</td>
<td>1.09</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>15.61</td>
<td>15.51</td>
<td>15.69</td>
<td>15.39</td>
</tr>
<tr>
<td>Minimum</td>
<td>-70.92</td>
<td>-61.92</td>
<td>-62.50</td>
<td>-54.45</td>
</tr>
<tr>
<td>Maximum</td>
<td>65.38</td>
<td>68.00</td>
<td>58.33</td>
<td>63.33</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade 6</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-0.48</td>
<td>1.59</td>
<td>2.82</td>
<td>1.56</td>
</tr>
<tr>
<td>Minimum</td>
<td>-50.15</td>
<td>-51.70</td>
<td>-60.00</td>
<td>-56.25</td>
</tr>
<tr>
<td>Maximum</td>
<td>56.00</td>
<td>56.06</td>
<td>56.25</td>
<td>50.38</td>
</tr>
</tbody>
</table>

**Writing.** For grade 3 writing PPS, the average score increased from 2006 to 2010. The standard deviation was highest in 2006 (Table 4.3). The average PPS for grade 6 writing scores was highest in 2010 and decreased in both 2007 and 2009. The standard deviations, across all years, were lower for grade 6 than grade 3, but the range of scores varied. Overall, the mean scores were higher for grade 3 than for grade 6 except for 2008. The provincial PPS was lower than the means derived from the data used in this study; however, the provincial PPS increased year after year, while grade 6 means fluctuated slightly from year to year.
Table 4.3

*Descriptive Statistics for Grade 3 and 6 Writing PPS*

<table>
<thead>
<tr>
<th></th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provincial PPS</td>
<td>64</td>
<td>64</td>
<td>66</td>
<td>68</td>
<td>70</td>
</tr>
<tr>
<td>Mean</td>
<td>67.21</td>
<td>67.49</td>
<td>68.89</td>
<td>70.72</td>
<td>72.10</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>16.08</td>
<td>15.63</td>
<td>15.89</td>
<td>15.12</td>
<td>15.63</td>
</tr>
<tr>
<td>Minimum</td>
<td>18.52</td>
<td>19.05</td>
<td>21.21</td>
<td>20.00</td>
<td>23.08</td>
</tr>
<tr>
<td>Maximum</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Grade 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provincial PPS</td>
<td>61</td>
<td>61</td>
<td>67</td>
<td>67</td>
<td>70</td>
</tr>
<tr>
<td>Mean</td>
<td>64.57</td>
<td>62.85</td>
<td>69.13</td>
<td>68.83</td>
<td>71.21</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>15.64</td>
<td>15.60</td>
<td>14.29</td>
<td>14.82</td>
<td>15.54</td>
</tr>
<tr>
<td>Minimum</td>
<td>16.00</td>
<td>15.22</td>
<td>25.00</td>
<td>23.08</td>
<td>22.22</td>
</tr>
<tr>
<td>Maximum</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Annual changes in PPS for grade 3 writing were on average positive and the highest mean was in 2009/2010 and the lowest in 2006/2007 (Table 4.4). Annual change in grade 6 writing PPS was on average negative for 2006/2007 and 2008/2009. The highest mean in grade 6 was in 2007/2008. The standard deviations, for annual change, were smaller for grade 6 writing PPS than grade 3 writing PPS.
Table 4.4

*Descriptive Statistics of Annual Changes in PPS for Grade 3 and 6 Writing*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grade 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.28</td>
<td>1.40</td>
<td>1.19</td>
<td>2.02</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>15.34</td>
<td>16.24</td>
<td>15.29</td>
<td>15.13</td>
</tr>
<tr>
<td>Minimum</td>
<td>-56.84</td>
<td>-57.94</td>
<td>-67.65</td>
<td>-62.73</td>
</tr>
<tr>
<td>Maximum</td>
<td>77.27</td>
<td>66.67</td>
<td>65.88</td>
<td>66.67</td>
</tr>
<tr>
<td><strong>Grade 6</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-1.72</td>
<td>6.3</td>
<td>-0.30</td>
<td>2.37</td>
</tr>
<tr>
<td>Minimum</td>
<td>-52.78</td>
<td>-45.45</td>
<td>-66.67</td>
<td>-77.78</td>
</tr>
<tr>
<td>Maximum</td>
<td>53.90</td>
<td>65.28</td>
<td>60.91</td>
<td>60.87</td>
</tr>
</tbody>
</table>

**Mathematics.** Average grade 3 mathematics PPS decreased in both 2008 and 2010 and was highest in 2009 (lowest standard deviation) (Table 4.5). Grade 6 mathematics PPS decreased in 2007 and 2010 and displayed the highest mean in 2009. The standard deviations for grade 6 scores were higher than grade 3 scores and the means were lower, across all years, for grade 6. The provincial PPS was slightly lower than the mathematics PPS averages in this study for both grades 3 and 6. Provincial PPS displayed an increasing trend for grade 3 mathematics. Provincial PPS for grade 6 mathematics scores had more variability and was lowest in the 2007 school year. Overall, the means in this study fluctuated slightly from year to year in comparison to the provincial PPS.
Table 4.5

*Descriptive Statistics for Grade 3 and 6 Mathematics PPS*

<table>
<thead>
<tr>
<th></th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provincial PPS</td>
<td>68</td>
<td>69</td>
<td>68</td>
<td>70</td>
<td>71</td>
</tr>
<tr>
<td>Mean</td>
<td>72.66</td>
<td>72.67</td>
<td>71.36</td>
<td>73.01</td>
<td>72.81</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>16.01</td>
<td>16.33</td>
<td>16.61</td>
<td>16.58</td>
<td>15.82</td>
</tr>
<tr>
<td>Minimum</td>
<td>25.00</td>
<td>22.22</td>
<td>20.00</td>
<td>20.59</td>
<td>22.73</td>
</tr>
<tr>
<td>Maximum</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Grade 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provincial PPS</td>
<td>61</td>
<td>59</td>
<td>61</td>
<td>63</td>
<td>61</td>
</tr>
<tr>
<td>Mean</td>
<td>64.32</td>
<td>61.76</td>
<td>63.28</td>
<td>64.70</td>
<td>61.61</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>17.71</td>
<td>18.28</td>
<td>18.82</td>
<td>17.80</td>
<td>18.93</td>
</tr>
<tr>
<td>Minimum</td>
<td>11.11</td>
<td>7.14</td>
<td>7.14</td>
<td>11.11</td>
<td>6.67</td>
</tr>
<tr>
<td>Maximum</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Annual changes for grade 3 mathematics PPS were, on average, lowest for the 2006 to 2007 school years and highest for the 2008 to 2009 schools years. Mean annual change alternated between a positive and negative mean indicating fluctuating scores from year to year. The average annual change for grade 6 mathematics was highest for 2007/2008 and lowest for 2009/2010. Overall, the standard deviation for annual change in mathematics was lower for grade 3 than for grade 6 except for the 2007 to 2008 schools years.
Table 4.6

*Descriptive Statistics of Annual Changes in PPS for Grade 3 and 6 Mathematics*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.01</td>
<td>-1.31</td>
<td>1.65</td>
<td>-0.21</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>16.26</td>
<td>16.40</td>
<td>15.91</td>
<td>15.40</td>
</tr>
<tr>
<td>Minimum</td>
<td>-58.73</td>
<td>-72.78</td>
<td>-69.23</td>
<td>-62.54</td>
</tr>
<tr>
<td>Maximum</td>
<td>65.00</td>
<td>71.43</td>
<td>61.22</td>
<td>56.25</td>
</tr>
<tr>
<td>Grade 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-2.56</td>
<td>1.52</td>
<td>1.42</td>
<td>-3.09</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>16.33</td>
<td>16.13</td>
<td>16.13</td>
<td>16.15</td>
</tr>
<tr>
<td>Minimum</td>
<td>-68.89</td>
<td>-77.78</td>
<td>-60.83</td>
<td>-59.67</td>
</tr>
<tr>
<td>Maximum</td>
<td>75.00</td>
<td>68.57</td>
<td>85.00</td>
<td>73.68</td>
</tr>
</tbody>
</table>

**Cross-year Correlations of PPS**

**Reading grade 3.** Correlations of PPS across school years were moderate and statistically significant, but highest for consecutive years (Table 4.7). The smallest correlation was between 2006 and 2010.

**Reading grade 6.** Correlations for PPS across years were positive and statistically significant. The highest correlations were for consecutive years and the lowest were between 2006 and 2009 and 2006 and 2010.
Table 4.7

Correlation of Reading PPS for Grade 3 (above diagonal) and 6

<table>
<thead>
<tr>
<th>Year</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>-</td>
<td>0.46*</td>
<td>0.41*</td>
<td>0.39*</td>
<td>0.38*</td>
</tr>
<tr>
<td>2007</td>
<td>0.54*</td>
<td>-</td>
<td>0.48*</td>
<td>0.43*</td>
<td>0.41*</td>
</tr>
<tr>
<td>2008</td>
<td>0.47*</td>
<td>0.54*</td>
<td>-</td>
<td>0.50*</td>
<td>0.47*</td>
</tr>
<tr>
<td>2009</td>
<td>0.46*</td>
<td>0.50*</td>
<td>0.54*</td>
<td>-</td>
<td>0.52*</td>
</tr>
<tr>
<td>2010</td>
<td>0.46*</td>
<td>0.48*</td>
<td>0.49*</td>
<td>0.55*</td>
<td>-</td>
</tr>
</tbody>
</table>

*Correlation is significant at the 0.01 level (2-tailed).

**Writing grade 3.** All correlations were positive and significant (Table 4.8). The largest correlations were between consecutive years except for the correlation between the years 2007 and 2008; the correlation between the years 2006 and 2009 was higher.

**Writing grade 6.** All correlations were over 0.50 except for the correlation between 2006 and 2010 which was slightly lower. The correlations for PPS were all significant and positive indicating an increase in PPS one year was related to an increase in PPS the following year.

Table 4.8

Correlation of Writing PPS for Grade 3 (above diagonal) and Grade 6

<table>
<thead>
<tr>
<th>Year</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>-</td>
<td>0.53*</td>
<td>0.44*</td>
<td>0.52*</td>
<td>0.39*</td>
</tr>
<tr>
<td>2007</td>
<td>0.59*</td>
<td>-</td>
<td>0.47*</td>
<td>0.44*</td>
<td>0.43*</td>
</tr>
<tr>
<td>2008</td>
<td>0.52*</td>
<td>0.57*</td>
<td>-</td>
<td>0.54*</td>
<td>0.48*</td>
</tr>
<tr>
<td>2009</td>
<td>0.50*</td>
<td>0.55*</td>
<td>0.57*</td>
<td>-</td>
<td>0.55*</td>
</tr>
<tr>
<td>2010</td>
<td>0.49*</td>
<td>0.53*</td>
<td>0.56*</td>
<td>0.58*</td>
<td>-</td>
</tr>
</tbody>
</table>

*Correlation is significant at the 0.01 level (2-tailed)*
**Mathematics grade 3.** Correlations for grade 3 mathematics PPS were positive and statistically significant across all years with the highest correlations between consecutive years and between the years 2006 and 2009 (Table 4.9).

**Mathematics grade 6.** PPS scores were moderately correlated and significant across all years. The lowest correlation was between the years 2006 and 2010.

Table 4.9

*Correlation is significant at the 0.01 level (2-tailed).

**Summary of Correlations**

**Grade 3.** Reading (0.38 to 0.52), writing (0.39 to 0.55) and mathematics (0.39 to 0.55) correlations shared a similar range of values, where in almost all cases the correlations were highest for consecutive years and moderate. The lowest correlations in all cases were between the years furthest apart.

**Grade 6.** The correlations were slightly higher for grade 6 subjects than for grade 3 subjects, overall. The range of correlations for reading, writing, and mathematics were 0.46 to 0.55, 0.49 to 0.59, and 0.47 to 0.62 respectively. The largest range of correlations and the highest correlations for consecutive years were for mathematics scores.
Degree and Frequency of Change in PPS

This section focuses on the change in the percentage of proficient students from 2006 to 2010. For each subject and grade, 2 figures are reported. The first pertains to the amount of change in PPS across years and the second displays the number of years a school experienced a similar degree of change in PPS from 2006 to 2010.

Reading grade 3. Figure 4.1 displays the degree of change in PPS, for schools, and it is evident that most schools experienced between 0 and 4 percent change in PPS from 2006 to 2010. The percentage of schools in each successive change category gradually decreased from the smallest change category with the fewest percentage of schools experiencing between 26 and 30 percent change in PPS. Within each percentage of change category, the percentage of schools varied slightly depending upon the year. For example, in the 2006 to 2007 school year, the smallest percentage of schools experienced change in PPS between 0 and 2 percent. Approximately 6 percent of schools had a change in PPS that was greater than 30.
Figure 4.1. The percentage of schools by size of change in PPS from 2006 to 2010 for grade 3 reading.

Smaller changes in PPS of less than 6 percent were most likely to occur for one or two years and sometimes three years (Figure 4.2). Changes in PPS between 6 and 20 percent occurred, in most cases, for one year. Larger changes in PPS of 21 percent or more occurred most frequently for one year and sometimes two years. Very few schools experienced the same amount of change for four years.
Figure 4.2. Frequency of annual change in PPS from 2006 to 2010 for grade 3 reading.

**Reading grade 6.** Approximately 12 and 14 percent of schools had a change in PPS ranging from 0 to 2 percent (Figure 4.3). The percentage of schools in each change category gradually declined from the lowest to highest change categories where the smallest percentage of schools experienced change between 28 and 30 percent. Approximately 4 percent of schools experienced change in PPS larger than 30 percent.

Overall, the degree of change for the largest percentage of schools was between 0 and 10 percent (approximately 50 percent). There were approximately 30 percent of schools that experienced between 11 and 20 percent change in PPS.
Figure 4.3. The percentage of schools by size of change in PPS from 2006 to 2010 for grade 6 reading.

The highest percentage of schools (approximately 80 percent) having changes in PPS between 16 and 20 percent did so for 1 year (Figure 4.4). Changes in school PPS of less than 6 percent occurred, in most cases, for 1 or 2 years – this was the only change category with a similar percentage of schools in both the 1 and 2 year category. Very few schools experienced the same amount of change in PPS for 4 years. The highest percentage of schools experiencing change for 3 years was in the 0 to 5 percent change in PPS category.
Figure 4.4. Frequency of annual change in PPS from 2006 to 2010 for grade 6 reading.

Writing grade 3. Figure 4.5 displays the breakdown of the degree of change in PPS; many schools experienced change in PPS between 0 and 2 percent. The percentage of schools in the 2 to 4 PPS change category, however, was lower for the 2007 to 2008 and 2008 to 2009 school years. In general there was a decrease in the percentage of schools that experienced greater change in PPS; although, the percentage of schools within each change category was variable for each year making the trend less distinct in the smaller change categories. Approximately 6 percent of schools experienced change in PPS that was greater than 30 percent; the percentage of schools in this change category was highest between the 2007 and 2008 school years.
Figure 4.5. The percentage of schools by size of change in PPS from 2006 to 2010 for grade 3 writing.

Over 80 percent of schools had a change in PPS between 16 and 20 percent for one year (Figure 4.6). The highest percentage of schools experiencing a similar amount of change for two or three years was in the 0 to 5 percent change in PPS category. Very few schools experienced a similar amount of change in PPS for four years.
Figure 4.6. Frequency of annual change in PPS from 2006 to 2010 for grade 3 writing.

Writing grade 6. In Figure 4.7, although there is a decrease in the percentage of schools in each change category as the size of change increases, there is some variability across years within each change category. For example, from the 2008 to 2009 school years, a larger percentage of schools experienced change in PPS between 0 and 2 percent; the same was also true for the 2 to 4 percent and 4 to 6 percent change categories. A larger percentage of schools from the 2007 to 2008 school years experienced changes in PPS between 28 and 30 percent and approximately 6 percent of schools had changes in PPS greater than 30 percent.
**Figure 4.7.** The percentage of schools by size of change in PPS from 2006 to 2010 for grade 6 writing.

In each change category, the largest percentage of schools had a similar amount of change in PPS for one year (Figure 4.8); the highest percentage of schools was in the 16 to 20 change category and the lowest percentage of schools was in the 0 to 5 percent change category. The percentage of schools having a similar amount of change in PPS for two, three and four years was largest in the 0 to 5 percent change category. Overall, very few schools experienced a similar amount of change for 4 years.
Figure 4.8. Frequency of annual change in PPS from 2006 to 2010 for grade 6 writing.

Mathematics grade 3. Figure 4.9 shows that the percentage of schools experiencing smaller amounts of change gradually declined as the size of the change category increased. In other words, a very small percentage of schools experienced 28 to 30 percent change in PPS. In the first four categories of change, the highest percentage of schools was in the 2009/10 school year. Approximately 20 percent of schools experienced changes in PPS greater than 20 percent.
Figure 4.9. The percentage of schools by size of change in PPS from 2006 to 2010 for grade 3 mathematics.

The majority of schools in each change in PPS category experienced a similar amount of change for one year (Figure 4.10). There were a larger percentage of schools in the 0 to 5 and 21 or more change in PPS categories that experienced a similar amount of change for 2 years compared to the other change categories. Very few schools had a similar amount of change in PPS for four years.
Mathematics grade 6. The percentage of schools in each of the change categories gradually declined from the first to the last change category (Figure 4.11). In the first three categories of change the highest percentage of schools was in 2006/07. In the very last change category the largest percentage of schools experiencing greater than 30 percent change in PPS was between the 2009 and 2010 school years.

Figure 4.10. Frequency of annual change in PPS from 2006 to 2010 for grade 3 mathematics.
Figure 4.11. The percentage of schools by size of change in PPS from 2006 to 2010 for grade 6 mathematics.

Over 80 percent of schools in the 16 to 20 percent change category had a similar amount of change for one year (Figure 4.12). The largest percentage of schools experiencing a similar amount of change for two and three years was in the 0 to 5 percent change in PPS category. A very small percentage of schools had a similar amount of change for four years.
Figure 4.12. Frequency of annual change in PPS from 2006 to 2010 for grade 6 mathematics.

Generally, the largest percentage of schools experienced small changes in PPS and the percentage of schools in each change category gradually decreased as the size of the change in PPS increased. Across all subjects and grades, the highest percentage of schools experienced change in PPS, across all years, between 0 and 10 percent. Approximately 30 percent of schools, however, had a change in PPS between 11 and 20 percent, nearly 13 percent of schools experienced changes between 21 and 30 percent and approximately 7 percent of schools had changes larger than 30 percent.

The largest percentage of schools, in each change category, had a similar amount of change for one year. Both the smallest and largest change categories had a larger percentage of schools experiencing a similar amount of change for two and three years. Very few schools had similar changes in PPS across all of the years examined.

A closer look was taken at schools experiencing larger changes in PPS (greater than 21) consecutively from 2006 to 2010 and the average cohort size for those schools ranged from 11 to 35 students (one classroom). Some schools with larger cohorts of more than 90 students also
experienced larger changes between 10 and 20 PPS, but very seldom did this happen for more than 1 year.

**How PPS Changes within Schools**

Figure 4.13 displays the percentage of schools, for all grades and subjects experiencing various patterns of change over the 5 years. When increases and decreases were coded as a ‘1’ and ‘0’ respectively there were schools in all 16 patterns. For instance, very few schools were in the ‘0000’ category (a school that had a decrease in PPS each year from 2006 to 2010) or in the ‘1111’ category (a school that had an increase in PPS each year from 2006 to 2010). Approximately 3 percent of schools had gains every year in grade 6 writing, for example. As would be expected, a higher percentage of schools experienced an alternating decrease/increase (0101) pattern. Over 18 percent of schools experienced the 0101 pattern in grade 6 writing and over 14 percent of schools experienced the alternating increase/decrease (1010) pattern for grade 3 mathematics.
Figure 4.13. The percentage of school by patterns of annual change for all grades and subjects from 2006 to 2010.

Table 4.10 displays the correlations of PPS across subjects and grades for each school year. Grade 3 reading PPS correlated highly with grade 3 writing and mathematics PPS, while the correlations between grade 3 writing and mathematics were slightly lower between the 2006 and 2007 school years. This was not the case for grade 6 subjects; all the subjects had higher correlations indicating that as school PPS increased in one subject an increase in PPS in the other two subjects was likely. When subjects were correlated across grades, the correlations were much lower indicating that an increase in reading PPS for grade 3 was not strongly related to an increase in grade 6 reading PPS, for instance.

---

3 Patterns with no change were seen in less than 1 percent of cases and as a result not included in Figure 4.13.
Table 4.10

Correlations Across Grades and Subjects from 2006 to 2010

<table>
<thead>
<tr>
<th>School Year</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading-Writing</td>
<td>0.75</td>
<td>0.76</td>
<td>0.76</td>
<td>0.80</td>
<td>0.74</td>
</tr>
<tr>
<td>Reading-Mathematics</td>
<td>0.76</td>
<td>0.74</td>
<td>0.74</td>
<td>0.75</td>
<td>0.74</td>
</tr>
<tr>
<td>Writing-Mathematics</td>
<td>0.67</td>
<td>0.68</td>
<td>0.72</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td>Grade 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading-Writing</td>
<td>0.78</td>
<td>0.77</td>
<td>0.76</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td>Reading-Mathematics</td>
<td>0.77</td>
<td>0.74</td>
<td>0.74</td>
<td>0.76</td>
<td>0.74</td>
</tr>
<tr>
<td>Writing-Mathematics</td>
<td>0.74</td>
<td>0.72</td>
<td>0.73</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td>Grade 3 and 6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>0.42</td>
<td>0.42</td>
<td>0.41</td>
<td>0.44</td>
<td>0.48</td>
</tr>
<tr>
<td>Writing</td>
<td>0.46</td>
<td>0.45</td>
<td>0.40</td>
<td>0.45</td>
<td>0.44</td>
</tr>
<tr>
<td>Mathematics</td>
<td>0.44</td>
<td>0.42</td>
<td>0.43</td>
<td>0.43</td>
<td>0.48</td>
</tr>
</tbody>
</table>

Variance in PPS and Average Cohort Enrolment

In order to determine the strength of the relationship between the variance in the percentage of students achieving proficiency from 2006 to 2010, at a school, and the number of students writing the assessment from 2006 to 2010, a correlational analysis was conducted (please see Appendix M for variance score distributions). The results revealed negative correlations between variance scores and average cohort enrollment for grade 3 reading ($r = -0.38, p < 0.01$), writing ($r = -0.38, p < 0.01$), and mathematics ($r = -0.35, p < 0.01$) and grade 6 reading ($r = -0.30, p < 0.01$), writing ($r = -0.31, p < 0.01$), and mathematics ($r = -0.31, p < 0.01$). The negative correlations indicate that an increase in average cohort size is related to a decrease in the variance in PPS across 5 years.
To further test the hypothesis that a decrease in cohort size would be related to more variability in PPS, the average variance score was plotted for 5 distinct cohort size categories (e.g., 10-30, 31-60, 61-90, 91-120, 121 or more). Table 4.11 displays the number of schools in each cohort size category. Most schools had either 10 to 30 students (1 classroom) or 31 to 60 students (2 classrooms) writing the assessment, while very few schools had 121 or more students (approximately 5 classrooms) writing the assessment. From figure 4.14 it is evident that an increase in cohort size is related to a decrease in average variability up until the 121 or more cohort size category where the average variance for mathematics increased for both grades 3 and 6. Table 4.12 displays the descriptive data of variance scores. The mean variance scores were highest for mathematics and lowest for grade 6 reading.

Table 4.11

<table>
<thead>
<tr>
<th>Number of Schools in Each Cohort Size Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort Size</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>10 – 30</td>
</tr>
<tr>
<td>31 – 60</td>
</tr>
<tr>
<td>61 – 90</td>
</tr>
<tr>
<td>91 – 120</td>
</tr>
<tr>
<td>121 or greater</td>
</tr>
</tbody>
</table>
Figure 4.14. Average variance by cohort size for grades 3 and 6 reading, writing and mathematics.

Table 4.12

Descriptive Statistics of Variance Scores

<table>
<thead>
<tr>
<th></th>
<th>Reading</th>
<th>Writing</th>
<th>Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Grade 3</td>
<td>Grade 6</td>
<td>Grade 3</td>
</tr>
<tr>
<td>N</td>
<td>2918</td>
<td>2829</td>
<td>2908</td>
</tr>
<tr>
<td>Mean</td>
<td>131.93</td>
<td>106.52</td>
<td>137.77</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>117.77</td>
<td>94.64</td>
<td>127.52</td>
</tr>
<tr>
<td>Minimum</td>
<td>1.06</td>
<td>0.94</td>
<td>1.35</td>
</tr>
<tr>
<td>Maximum</td>
<td>979.64</td>
<td>878.05</td>
<td>1183.08</td>
</tr>
</tbody>
</table>
CHAPTER FIVE: DISCUSSION

In this chapter, the results from the analyses presented in the previous chapter will be discussed. This chapter begins with an overview of the purpose of this study, the research questions posed in chapter one and a summary of the analyses conducted. The results of each research question will be discussed in a separate section. In each section results are summarized, interpreted and the significance is discussed. This section also includes the educational importance and implications of the findings and future studies.

Overview of Study

One of the reasons EQAO testing was put in place in Ontario was to hold schools accountable for their results by demonstrating progress in schools to stakeholders. Most often the percentage of proficient students, at a school, is used to make decisions about schools. Provincially, PPS is increasing with small changes year after year; however, at the school level some schools may be ranked highly one year but not the following year. At the school level, results may be more susceptible to variations. It was apparent from such inconsistencies that school results needed further exploration.

The purpose of this study was to examine changes in school level PPS across 5 years (2006 to 2010) to reveal the degree and frequency of changes in the percentage of proficient students, at a school, in the areas of reading, writing, and mathematics for both grades 3 and 6 elementary schools in Ontario. A second purpose was to assess the influence of cohort size on the variability in scores from year to year.

A descriptive analysis was conducted to explore the degree and frequency of change in PPS alongside correlations to examine the relationship of PPS from 2006 to 2010. The first research question addressed the size of year to year changes in PPS. Once this was established it
was important to examine how PPS fluctuated across years, within schools, which was the focus of the second research question. From the results of these two research questions it was evident that there were large fluctuations in school PPS from year to year across all subjects and grades. Therefore, the third research question examined the relationship between the variance in PPS within a school and cohort size. To examine this, variance scores were computed, for each grade and subject, and correlations and variance scores were examined.

The research findings establish that school scores are not as stable as provincial PPS and indeed are not very stable at all. School scores change from year to year, and for many schools these changes can be very large. Part of this variance is related to the average size of the cohort writing the test. Similar to the findings of Ho (2008), the rankings made based on yearly PPS cannot, as shown in this study, be seen as reliable. For instance, schools ranked solely on the percentage of proficient students in a school year may be surprised to see their school among the top ranked schools one year, but not the next. Even small increases or decreases frequently seen in this analysis (e.g. 5 percent or more), may mean the difference between being one of the top ranked schools or not. For other schools experiencing even larger changes from year to year, the variations may prove to be quite frustrating to schools and confusing to stakeholders.

**Descriptive Statistics**

PPS means in this study were higher than provincial PPS. Correlating PPS across years resulted in moderate correlations which were higher for consecutive years than for years further apart. This indicates that PPS from year to year is not very reliable and using PPS as the basis for accountability programs may prove to be inaccurate. Many of the studies on stability of test scores reviewed earlier used a correlational analysis and some used the correlations to estimate the degree of error attributed to cohort changes across years. If schools were shifting in PPS from
year to year it is appropriate to assume that the correlations would not be very high, which is the case as is shown by moderate correlations. In this study, correlations were moderate and lower than those found in the Sicoly (2002) study. Both in this study and the Sicoly (2002) study correlations were much lower than those reported by Haney (2002) which ranged from 0.86 to 0.87; however, Haney (2002) used average test scores and not PPS. Thus, using test score averages may be more stable than using PPS as is explained by Ho (2008).

Although correlations provide a general picture of the variance in PPS across years, they provide minimal information about the amount of change from year to year.

**Degree and Frequency of Changes in PPS**

The literature examining year to year changes in PPS is minimal and findings vary from study to study, but one thing is clear – PPS does change from year to year, and often by very large amounts. Findings indicate that approximately 88 percent of schools had changes larger than 2 percent, 77 percent of schools had changes larger than 4 percent, and approximately 50 percent of schools experienced changes in PPS greater than 10 percent in any given year.

When examining how often, from 2006 to 2010, a school experienced a similar amount of change – changes less than 6 percent were likely to be experienced by schools more than once in the 5 years. Overall, schools experiencing less than 6 percent change in PPS did so similarly for 1 and 2 years. Schools experiencing change in PPS between 6 and 10 percent and more than 21 percent did so once and sometimes twice. Schools were a lot less likely to experience change between 11 and 20 percent more than once. Nearly 80 percent of schools experiencing change in PPS between 11 and 15 percent did so for one year and approximately 20 percent of schools were likely to experience this level of change for 2 years. Very seldom did schools experience any change in PPS across all 5 years. When comparing the 16 to 20 percent change in PPS
category to the 21 percent or more change in PPS category, there were a higher percentage of schools experiencing change for 1, 2, 3 and 4 years in the latter category; whereas, in the former category the majority of schools (over 80 percent) experienced a similar amount of change for one year.

According to Hill and DePascale (2003), Kane and Staiger (2002) and Kane, Staiger Geppert (2001) many of these changes from year to year are a result of measurement error and sampling error – more so sampling error where a different cohort is tested each year. Hill and DePascale (2003), however, believe that variance in year to year scores remains even after accounting for non-persistent factors. The remaining variance may be a result of equating error mentioned by Wu (2010), which may outweigh both measurement and sampling errors because it is influential at both the student and school levels. Accordingly, much of the change in PPS may be random given that the majority of schools experienced large levels of change at least once from 2006 to 2010.

**How PPS Fluctuates within Schools**

Once it was recognized that schools were experiencing shifts in PPS from year to year, the nature of these fluctuations was the focus of the second research question. Of the 16 types of change across all subjects and grades, the most frequent patterns of change were ‘decrease-increase-decrease-increase’ (0101), ‘increase-decrease-increase-decrease’ (1010) and ‘decrease-increase-increase-decrease’ (0110). In an attempt to quantify change within each type of change category, no clear patterns were found. For instance, some schools with the ‘0101’ pattern had a decrease of 5 percent in PPS followed by an increase of 2 percent in PPS and so on. All schools increased and decreased in differing increments with no clear pattern emerging across the 5 years. A continually increasing pattern (‘1111’) was seen to occur very infrequently. In other
words, very few schools were able to continually increase their PPS scores, most were increases followed by decreases or the opposite.

Unidimensionality relates to the consistency of schooling (i.e. Are results the same for different grades, subjects and years at a school?). The findings of the correlations run within subjects ranged from 0.67 to 0.80 for grade 3 and from 0.71 to 0.78 for grade 6 indicating students doing well in one subject also did well in the other subjects being. In theory, if scores from testing are used to gauge a school’s ability to educate students, then a successful school should ideally display high levels of student achievement (or PPS) regardless of the grade, subject or year being examined in that school; however, results from correlating PPS within the same grade were higher than correlating PPS across grades (3 and 6). When correlating grade 3 reading scores, for instance, with grade 6 reading scores the results were much lower. The correlations ranged between 0.40 and 0.48, indicating that the variance in achievement for both grades 3 and 6 was only shared to a moderate degree. This may be due to differing students, tests, teachers, resources allotted to grade 3 and 6 classrooms, and/or the size of the classroom. As a result, classroom and student level factors may be more important than school level factors in addressing the differences in PPS within a school.

Variance in PPS and School Enrolment

To determine whether the size of the cohort being tested was related to the variance in PPS, a variance score was calculated using 5 years of data resulting in one score for each subject within each grade. The variance score quantified how much PPS in one year deviated from the PPS in the previous year; a higher variance score indicated more variability in PPS and a lower variance value indicated smaller changes in PPS from 2006 to 2010. In order to determine whether average cohort size, from 2006 to 2010, was related to variance in PPS a correlational
analysis was conducted. Correlations were negative indicating an increase in cohort size was related to a decrease in variance scores; however, the correlations were moderate. Many of the studies reviewed explored the relationship between cohort size and changes in student outcomes; although, in this study variance scores were used instead to examine changes in PPS across 5 years. Overall, the studies reviewed showed a commonality with the results of this study – a decrease in cohort size was related to an increase in changes in PPS. For instance, in a larger cohort of 100 students each student’s score is equivalent to 1 percent; in other words, if everyone in that cohort achieves Level 3 except for one individual the PPS for that school would be 99 percent. In contrast, in a class with 10 students, each student accounts for 10 percent where 1 individual not achieving proficiency would lead to a PPS of 90 percent explaining why smaller schools may have higher variance scores.

In order to further delve into the relationship of average cohort size and variability in PPS, the average variance was mapped against average cohort size. If the hypothesis that there is more variability in PPS in smaller cohorts is true, then greater variability will be seen in small cohort sizes and as a result a larger average variance score – the average variance should decrease with increasing cohort size. The results indicate that average variance continues to decrease as average cohort size increases. There was a slight increase in average variance for grade 3 and 6 mathematics in the 121 or more cohort size category; this may because there are fewer schools in the 121 or more school cohort size category where one school having greater fluctuations may influence the distribution and as a result the average variance. Just as in a small cohort of students one or two students may drastically change the PPS and influence PPS variability across years. Despite this, the average variance, for grade 3 and 6 mathematics for the 121 or more cohort size category is still much smaller than the smaller cohort size categories
(e.g., 10 to 30, 31 to 60 and 61 to 90). From this data, schools with less than 60 students writing the assessment should expect more variability in PPS. According to Morgan (2002), $\sqrt{N/N}$ can be used to estimate confidence intervals of fluctuations. For example, a school with a cohort of 100 students writing the test would anticipate their scores to fluctuate by 10 percent. In this study, although, schools with smaller cohorts did experience more variance in scores than schools with larger cohorts, there were also some schools with large cohorts displaying large changes in PPS and small schools displaying small changes in PPS for some years.

**Educational Importance and Implications**

Presently, the most widely used method of reporting school achievement in Ontario is to use the percentage of proficient students; in other words, reporting the percentage of students, at a school that meets (level 3) or exceeds (level 4) the provincial standard. The results from this study show that annual PPS fluctuates from year to year, in many cases quite dramatically, especially for schools with a small cohort. Considering 80 to 90 percent of schools (depending on the subject and grade) have 60 or fewer students in a cohort the use of annual PPS appears to be misleading and unreliable for the majority of schools.

The results from testing are used to make decisions about the school, such as instruction strategies for teachers, strategies to address areas of weakness in the subject areas tested, adjust or create school improvement plans, allot resources, and make changes to policies and programs. In Ontario, OFIP identifies schools requiring assistance and places them into one of three programs based a static or declining PPS over two or three consecutive years. With frequent large and random changes in PPS, for schools, this could mean the adoption of practices that do not have real merit, or worse neglecting schools that require assistance (Kane & Staiger, 2002).
Opinions regarding the usefulness of external assessments vary. Many teachers disagree with standardized testing at the elementary school level (Johnson, 2005). In particular, teachers disagree with making PPS available to the public and then using this score to evaluate the effectiveness of a school (Johnson, 2005). Unfortunately, some may continue to use test results to misinform the public through rating and ranking schools based on yearly EQAO scores, which may then be used by parents to choose what they think is an appropriate school for their child. While reporting annual scores to the public may be useful to those interested in a schools’ progress, this has become a major issue as annual scores are being used inappropriately. Because these scores have large amounts of variation from year to year, public reporting can also bring unfair stress and anxiety to school staff. Without having knowledge about variability in results and how school scores may be exaggerated using the current reporting method of percentage above a cut score, the confusion in year to year variance of annual scores may decrease public confidence in test results and their usefulness.

With over 20 billion dollars spent each year on education in Ontario for over 2 million students it is imperative that public measures of results are highly reliable. Continuing to use a reporting method that has been shown to have large changes especially in small cohorts is not recommended. Using annual PPS can greatly over- or under-estimate real school performance, especially if there are a small number of students in the cohort being evaluated.

**Recommendations.** It is not recommended that annual scores of individual schools be used for any consequential purposes due to the large degree of variance in annual scores. Specifically, I make recommendations for EQAO, schools and the public.

**EQAO.** EQAO should identify very clearly in all of their reporting the degree of fluctuations in year to year results at a school especially for small cohorts. Also, EQAO should
emphasize the dangers in using school-level results to draw any conclusions about school performance especially using one year of data.

An alternative might be to include students in all of the EQAO levels instead of solely focusing on students who are proficient. Currently, schools may make great strides in moving students from level 1 to level 2 but this would practically go unnoticed because the emphasis on reporting results focuses on students in both levels 3 and 4. Using averages has also been shown to be a better indicator of achievement because there may be a reduction in variance producing more reliable results; however, there are different averages that may be used to compute an annual score for a school.

Building on the cohort-to-cohort change model explained by Braun et al. (2010), Briggs (2011) suggests averaging both grade 3 and 6 results, at a school, for each subject to reduce some of the year to year variability in PPS; by including more students in the average, the measurement and sampling error is reduced. Also, by averaging results from both grades over many years, there is overlap in students, which Briggs (2011) proposes “enhances the stability of school level statistics from year to year” (p. 6). Similarly, Linn and Haug (2002) suggest combining data across multiple grades, subject areas or years to increase the accuracy of school level results (Figure 5.1). The inclusion of a larger number of students in the average (i.e., across grades) also includes more teachers, providing a sense of shared responsibility for results. Although averaging results across grades and subjects deemphasizes the relative strengths and weaknesses of a school in certain grades or subject areas, this does not prevent separate reporting of results (Linn & Haug 2002) to school personnel for planning.
**Figure 5.1.** Average of school PPS across multiple subjects

**Schools.** It is strongly advised that improvement planning not be based solely or primarily on PPS, but rather used in conjunction with other information. A more comprehensive report for schools to use for improvement planning could also include averages of results across grades, subjects and years in an effort to reduce the variability in scores.

For school improvement and planning purposes, in a world increasingly relying on data to inform accountability systems, educators have access to very large amounts of data. For instance, the Ontario Statistical Neighbours information system can provide answers to over 80 billion queries (The Literacy and Numeracy Secretariat, 2007). As a result, it is the responsibility of educators to know how to correctly use and interpret data in order to improve student learning (Popham, 2004; Stiggins, 2002). In using data to improve student achievement, van Barneveld (2008) offers three recommendations to increase the ability of teachers to effectively use data for school improvement purposes: “Develop professional learning communities focused on reviewing and interpreting data for the purpose of improving student achievement; Leaders must be prepared to guide a process of data investigation that results in improved student learning; and
Induct new teachers into a data-based decision-making culture by linking them with veteran teachers who have experience in using data to enhance student achievement.” (p. 3).

Undoubtedly, schools are more complex than a single numerical score that has been shown to have dramatic fluctuations over time. Using PPS as the basis of an accountability system or to assess the quality of teachers and education at a school is not recommended. An appropriate alternative would be to develop a complete representation of achievement at individual schools including information from educators to better inform decisions about that school.

**Public.** Although annual test scores, according to Crundwell (2005), are released in a way that encourages the rating and ranking of schools, this is strongly not recommended based on the results of this study. Annual scores should *never* be used to compare schools or make any other judgements about the status or progress of a school. This is simply because the scores have large variations from year to year and do not necessarily reflect ‘real’ change. A large part of these fluctuating scores may be the small number of students writing the assessment which is the case for the majority of schools in Ontario.

Comparing schools in this manner may have other consequences for schools, such that schools reporting lower achievement even if the result is not accurate may have difficulty in retaining or attracting experienced teachers (Hargreaves & Fink, 2006) or worse, leading some schools to cheat (Simner, 2000) in an attempt to maintain equilibrium within their school. There is more to a school and the learning that takes place in the school that can ever amount to one score. Therefore, judgements about schools should only be made using a number of indicators that provide a comprehensive view of educational quality.
Conclusion

The complexity of the testing process is undeniable from translating objectives into test items all the while taking into account such things as cost, test time and psychometric properties (Wu, 2010) to consistently administering the test and grading test items to disseminating the results which are then used to make decisions that will benefit schools and ultimately the students. All tests have limitations, no single score is going to tell you everything you need to know about a school, but we can make the information less misleading by choosing how to report the results. Without an effort made to acknowledge and understand variability in school level scores and de-emphasize the focus placed on fluctuating results, it may be difficult in turn to change attitudes towards external assessments and restore public confidence in assessment results. This underlying apprehension within the education community, with respect to external testing, may put added pressure on schools to do well making it difficult to maintain a high morale within the school community. Teachers unenthusiastic about the assessment may lead students to not take the test very seriously or conversely have overly anxious students when teachers overly stress the importance of the test and results further adding to already fluctuating scores.

Education is a long-term investment in the future well-being and prosperity of a nation where “[e]very student deserves a good outcome from his or her public education.” (EQAO, 2005, p.1). According to Wu (2010) avoiding assessment systems that attempt to “solve it all” and instead using focused studies may prove to be “more suited to establish the effectiveness of a particular intervention, or a particular policy change. Smaller, purposeful and targeted assessment programs may achieve a narrow but well-defined set of objectives rather than a large-
scale assessment system that may not be able to meet all the objectives of the assessment.” (p. 24).

The challenges of choosing to focus on the percentage of students above a cut score and the consequences of using potentially misleading results to inform decisions about schools and funding should play a key role in designing accountability systems. “In implementing testing programs, we should stop and reflect on whether objectives are met in the programs, and whether and how improvements can be made.” (Wu, 2010, p. 24). In many cases, misinterpretations may be made due to stakeholders not being made aware of the degree and frequency of change in using PPS to represent school achievement and in unfavourable scenarios being used to make misinformed or inappropriate policy decisions (Wu, 2010).

**Future Studies**

Many studies have examined various influences on achievement; however, few have examined the variability of school achievement levels over time. As a result, future studies could include investigating factors such as student composition (e.g. ESL students, percentage of males and females, students with special needs) and teacher and principal stability at a school that lead some schools to have highly variable scores from year to year and why some schools experience larger changes in PPS than others. From the findings of this study, student and classroom factors may need more attention when attempting to explain changes in PPS from year to year. Instead of solely placing emphasis on results from large-scale assessments, the benefits of using results from classroom assessments in conjunction with EQAO results and other factors in an effort to make the current accountability system a comprehensive representation of the school should be explored.
In moving forward, studies evaluating whether there are methods for public reporting of results that would produce less variability in results such as the use of averages (of EQAO levels) could lead to more reliable scores. Examining the inclusion of students in all of the EQAO levels instead of solely focusing on those who are proficient may also prove to lessen variability in school level scores. Knowing the specific degree of error inherent to all schools as a result of using PPS, including a breakdown of the amount of different types of error (e.g. measurement, sampling and equating errors) would be beneficial in understanding scores and choosing an appropriate alternative to using PPS. Otherwise, schools may seem as though they are improving or declining when in reality their true scores may be quite similar from year to year. Other future research could include examining the relationship between variations in PPS, at a school, and changes in identified OFIP schools and school improvement planning.

In addition, consequential validity has not been systematically examined (Volante, 2007) which relates to the impact of external tests on students and teachers. The consequences of having highly variable school level results over many years on students, teachers, parents, school boards, policy makers and the community has not been well documented. It would be worthwhile to understand the impact this piece of the accountability system has on the education system as a whole.

All of the aforementioned studies would help to understand the reasons behind fluctuations in school level results. These studies encourage moving towards a more accurate and comprehensive method of assessing a schools’ true achievement by determining key components necessary for a complete representation of a school and methods that would produce less variability in school level scores.
References


primary division (Grades 1–3) and junior division (Grades 4–6). Retrieved from http://www.eqao.com


The Alberta Teachers’ Association (2006). Responding to the Fraser Institute’s school rankings:

The Literacy and Numeracy Secretariat (2007). Ontario statistical neighbours: Informing our
strategy to improve student achievement. Retrieved from
http://www.edu.gov.on.ca/eng/literacynumeracy/osneng.pdf

Thomas, S., Peng, W. J. & Gray, J. (2007). Modelling patterns of improvement over time: Value
added trends in English secondary school performance across ten cohorts. Oxford Review

Columbus, OH: Pearson Education Inc.

http://www.socialresearchmethods.net/kb/measerr.php

http://www.socialresearchmethods.net/kb/unobtrus.htm


No. 15. Retrieved from


Appendix A

Distribution of Grade 3 Reading PPS

Figure 1A. Distribution of grade 3 reading PPS for 2006.

Figure 2A. Distribution of grade 3 reading PPS for 2007.
Figure 3A. Distribution of grade 3 reading PPS for 2008.

Figure 4A. Distribution of grade 3 reading PPS for 2009.
Figure 5A. Distribution of grade 3 reading PPS for 2010.
Appendix B

Distribution of Grade 6 Reading PPS

Figure 1B. Distribution of grade 6 reading PPS for 2006.

Figure 2B. Distribution of grade 6 reading PPS for 2007.
Figure 3B. Distribution of grade 6 reading PPS for 2008.

Figure 4B. Distribution of grade 6 reading PPS for 2009.
Figure 5B. Distribution of grade 6 reading PPS for 2010.
Appendix C

Distribution of Grade 3 Writing PPS

Figure 1C. Distribution of grade 3 writing PPS for 2006.

Figure 2C. Distribution of grade 3 writing PPS for 2007.
Figure 3C. Distribution of grade 3 writing PPS for 2008.

Figure 4C. Distribution of grade 3 writing PPS for 2009.
Figure 5C. Distribution of grade 3 writing PPS for 2010.
Appendix D

Distribution of Grade 6 Writing PPS

Figure 1D. Distribution of grade 6 writing PPS for 2006.

Figure 2D. Distribution of grade 6 writing PPS for 2007.
Figure 3D. Distribution of grade 6 writing PPS for 2008.

Figure 4D. Distribution of grade 6 writing PPS for 2009.
Figure 5D. Distribution of grade 6 writing PPS for 2010.
Appendix E

Distribution of Grade 3 Mathematics PPS

Figure 1E. Distribution of grade 3 mathematics PPS for 2006.

Figure 2E. Distribution of grade 3 mathematics PPS for 2007.
Figure 3E. Distribution of grade 3 mathematics PPS for 2008.

Figure 4E. Distribution of grade 3 mathematics PPS for 2009.
Figure 5E. Distribution of grade 3 mathematics PPS for 2010.
Appendix F

Distribution of Grade 6 Mathematics PPS

Figure 1F. Distribution of grade 6 mathematics PPS for 2006.

Figure 2F. Distribution of grade 6 mathematics PPS for 2007.
Figure 3F. Distribution of grade 6 mathematics PPS for 2008.

Figure 4F. Distribution of grade 6 mathematics PPS for 2009.
Figure 5F. Distribution of grade 6 mathematics PPS for 2010.
Appendix G

Distribution of Annual Change in Grade 3 Reading PPS

*Figure 1G.* Distribution of annual change in grade 3 reading PPS, 2006-2007.

*Figure 2G.* Distribution of annual change in grade 3 reading PPS, 2007-2008.
**Figure 3G.** Distribution of annual change in grade 3 reading PPS, 2008-2009.

**Figure 4G.** Distribution of annual change in grade 3 reading PPS, 2009-2010.
Appendix H

Distribution of Annual Change in Grade 6 Reading PPS

*Figure 1H*. Distribution of annual change in grade 6 reading PPS, 2006-2007.

*Figure 2H*. Distribution of annual change in grade 6 reading PPS, 2007-2008.
Figure 3H. Distribution of annual change in grade 6 reading PPS, 2008-2009.

Figure 4H. Distribution of annual change in grade 6 reading PPS, 2009-2010.
Appendix I

Distribution of Annual Change in Grade 3 Writing PPS

Figure 1I. Distribution of annual change in grade 3 writing PPS, 2006-2007.

Figure 2I. Distribution of annual change in grade 3 writing PPS, 2007-2008.
Figure 3I. Distribution of annual change in grade 3 writing PPS, 2008-2009.

Figure 4I. Distribution of annual change in grade 3 writing PPS, 2009-2010.
Appendix J

Distribution of Annual Change in Grade 6 Writing PPS

Figure 1J. Distribution of annual change in grade 6 writing PPS, 2006-2007.

Figure 2J. Distribution of annual change in grade 6 writing PPS, 2007-2008.
Figure 3J. Distribution of annual change in grade 6 writing PPS, 2008-2009.

Figure 4J. Distribution of annual change in grade 6 writing PPS, 2009-2010.
Appendix K

Distribution of Annual Change in Grade 3 Mathematics PPS

Figure 1K. Distribution of annual change in grade 3 mathematics PPS, 2006-2007.

Figure 2K. Distribution of annual change in grade 3 mathematics PPS, 2007-2008.
Figure 3K. Distribution of annual change in grade 3 mathematics PPS, 2008-2009.

Figure 4K. Distribution of annual change in grade 3 mathematics PPS, 2009-2010.
Appendix L

Distribution of Annual Change in Grade 6 Mathematics PPS

Figure 1L. Distribution of annual change in grade 6 mathematics PPS, 2006-2007.

Figure 2L. Distribution of annual change in grade 6 mathematics PPS, 2007-2008.
Figure 3L. Distribution of annual change in grade 6 mathematics PPS, 2008-2009.

Figure 4L. Distribution of annual change in grade 6 mathematics PPS, 2009-2010.
Appendix M

Variance Score Distributions

Figure 1M. Distribution of grade 3 reading variance scores.

Figure 2M. Distribution of grade 6 reading variance scores.
Figure 3M. Distribution of grade 3 writing variance scores.

Figure 4M. Distribution of grade 6 writing variance scores.
Figure 5M. Distribution of grade 3 mathematics variance scores.

Figure 6M. Distribution of grade 6 mathematics variance scores.