Using Large-Scale Assessment Scores to Determine Student Grades

Tess Miller
University of Prince Edward Island

Abstract

Many Canadian provinces provide guidelines for teachers to determine students' final grades by combining a percentage of students' scores from provincial large-scale assessments with their term scores. This practice is thought to hold students accountable by motivating them to put effort into completing the large-scale assessment, thereby generating a more accurate assessment of their ability. This study examined teachers' perceptions of the accountability framework underpinning large-scale assessments—in particular, teachers' beliefs and practices related to using students' provincial assessment scores to determine final grades. Questionnaires were distributed to teachers and follow-up interviews were conducted. Findings revealed that teachers did not entirely endorse the practice of using large-scale assessment results to determine student grades; instead, they appeared to be applying the guidelines while at the same time tweaking students' scores as needed to ensure everyone received a passing grade (i.e., at least 50%) in their course. Further, teachers were drawing from the large-scale assessment instrument to guide their instruction.

Keywords: Large-scale assessment, accountability, grade 9 mathematics, student grades

Précis

Plusieurs provinces canadiennes fournissent des lignes directrices aux enseignants afin qu'ils intègrent un pourcentage du résultat des élèves à des évaluations provinciales à grande échelle à celui de la cession en cours, pour déterminer les notes finales. Cette pratique a été examinée, de même que d'autres pratiques d'enseignants pour lesquelles l'évaluation à grande échelle influençait l'enseignement. Des questionnaires ont été distribués aux enseignants, et des entrevues de suivi ont été menées. Les résultats ont révélé que les enseignants ne souscrivaient pas entièrement à la pratique, mais semblaient appliquer les lignes directrices, tout en peaufinant les résultats des élèves, selon les besoins, afin que tous reçoivent au moins la note de passage (50 %) dans leurs cours. En outre, les enseignants s'inspiraient de l'instrument d'évaluation à grande échelle dans l'orientation de leurs pratiques d'enseignement et d'évaluation.

Introduction

The primary purpose of large-scale assessments is to promote student achievement by holding accountable those responsible for educating students (Chudowsky & Pellegrino, 2003; Decker & Bolt, 2008; Klinger, Rogers, Miller, & DeLuca, 2008; Klinger, DeLuca, & Miller, 2008). Typically, school boards, schools, and teachers are the key stakeholders in this accountability framework, but more recently, students have been drawn in as well. In some provinces (e.g., Alberta, Newfoundland & Labrador, Ontario, and Prince Edward Island), school boards mandate guidelines that call for incorporating a percentage of students' criterion-referenced, large-scale assessment (LSA) score with their overall grade in a course. This practice is thought to hold students accountable by motivating them to put effort into completing the LSAs, thereby generating a more accurate assessment of their abilities (van Barneveld & Brinson, 2011). For instance, in British Columbia (BC), the Grade 10 and 11 provincial assessments in core subjects must be included as 20% of students' grades (BC Ministry of Education, 2004), and in Prince Edward Island (PE) teachers were required to include scores from the Grade 9 provincial mathematics assessment as 10% of their students' grades (PE Department of Education and Early Childhood Development, 2010). In other jurisdictions, incorporating LSA scores as part of students' grades is at the discretion of teachers. For example, in Ontario (ON), teachers have the option of marking some or all items on the provincial Grade 9 mathematics assessment before returning student response sheets to the official marking board for scoring (ON Ministry of Education, 2009). However, there is no province-wide policy to regulate the percentage of students' grades that come from the provincial assessment. Similarly, in Alberta (AB), teachers are encouraged, but not required, to mark and include the scores from Grades 3, 6, and 9 provincial assessments in core subjects as part of students' grades (AB Education, 2007). Compounding the variation in these practices, the procedures used to select and score items included in students' grades also varied. Some provinces (e.g., PE) provide detailed scoring criteria that teachers must use, while other provinces (e.g., ON) permitted teachers to develop their own scoring criteria—for instance, by determining which items to use and how many, as well as how marks are allocated. Given the number of differences surrounding the ways in which LSA scores are used to determine student grades, teachers' perspectives on the uses of LSAs are likely to vary from province to province.

Research focusing on the validity of using LSA scores to determine student grades and the extent to which they are used is beginning to emerge in some provinces (e.g., Newfoundland and ON) (Fushell, 2011; Koch, 2011a; van Barneveld & Brinson, 2011; van Barneveld, King, & Nadon, 2011). In Newfoundland (NL), for example, 20% of a student's grade is based on his or her score in the Grade 9 Assessment of Mathematics, and the majority (60%) of teachers in NL support this practice. However, there are no instructions guiding teachers in determining "what items or components to include, standards to apply, or procedures for marking" (Fushell, 2011, p.6). Likewise, in ON, an absence of guidelines surrounding the use of students' LSA scores to determine final grades resulted in teachers using different items and different weighting schemes (Simon, van Barneveld, King, & Nadon, 2011); these researchers subsequently recommended guidelines to standardize this practice. A similar study in ON (Koch, 2011b) also reported variability in the use of students' LSA scores to determine final grades. In that study, it was found that some teachers were not using LSA scores at all, whereas other teachers used up to 20% of students' LSA scores to determine students' final grades in mathematics. From the students' perspective, van Barneveld and Brinson (2011) found that many students were not aware of whether their LSA scores counted towards their grade in the course; consequently, these researchers called for greater clarity and consistency in communicating this practice to all stakeholders.

In response to the need for more research in this area, in particular an examination of practices in other jurisdictions, the present study aimed to investigate (i) the extent to which LSA scores were being used to determine student grades, and (ii) other ways in which LSAs influenced teachers' practices. More specifically, the Grade 9 provincial assessment of mathematics in PE was selected as the context for this study because PE has implemented guidelines for standardizing the use of LSAs in determining final grades (PE Department of Education and Early Childhood Development, 2013). To obtain an understanding of teachers' dispositions towards the accountability aspect of LSAs, the purpose of the first research question was to gain general insight into teachers' beliefs and issues related to LSAs. The question posed was: "To what extent do teachers' beliefs about LSAs affect their reported practice of using LSA scores?" The second research question specifically sought to examine the use of LSAs in determining student grades. The question posed was: "To what extent do teachers in PE incorporate students' scores from the LSA of Grade 9 mathematics with students' term scores to calculate final grades?"

Prince Edward Island Context

Prince Edward Island is a small province with a population of approximately 141,000 (11% French speaking); slightly more than half of the population is located in rural settings (PE Government, 2010). PE has a growing population of immigrants (350 in 2007, 400 in 2008, 740 in 2009, and 1200 in 2010), but other than in 2010, this growth rate has not exceeded the rates in other provinces (Statistics Canada, 2010).

In the fall of 2010, PE's publicly funded, semi-private kindergarten program, begun in 2000, moved to the public sector, enabling a seamless kindergarten (K) to Grade 12 program.

PE's educational program is divided into three school boards separated into three divisions (i.e., primary, middle, and senior) of varying divisional splits (e.g., the primary division can have Grades K to 3, K to 5, or K to 9), rather than divisions based on students' cognitive developmental stages.

Although the education system in PE is similar to other jurisdictions' systems, students in PE have consistently scored at or near the bottom on national and international assessments in science, mathematics, and literacy (Programme for International Student Assessment (PISA) [Science], 2006; PISA [Mathematics], 2003; PISA [Literacy], 2003; Student Achievement Indicators Program (SAIP) [Writing], 2002; SAIP [Mathematics], 2001).[1] There is no literature or empirical research attempting to explain this phenomenon. In 2006, the PE Department of Education created a task force commissioned to explore the state of education in PE (Kuriel, 2005). Based on the Kuriel report, an additional five million dollars was allocated to revise the curriculum and develop provincial assessments to ensure "the curriculum is being covered and outcomes are being achieved" (PE Government, 2006). Since the implementation of provincial assessments in PE, student achievement on PISA's international assessment has continued to lag behind results from other provinces, particularly in mathematics, where PE scored the lowest of all Canadian provinces and lowest of all 65 participating countries (PISA [Mathematics], 2009). Further evidence of the alarming state of mathematics education in PE are

1 PISA assesses 15-year-olds in reading, mathematics, and science across Canada. SAIP assessed 13- and 16-year-olds in mathematics, reading, and science across Canada; PCAP—the Pan-Canadian Assessment Program—replaced SAIP in 2007.

the more recent scores on the national Pan-Canadian Assessment Program (PCAP), where PE retained its position as the lowest scoring province (PCAP [Mathematics], 2010). These statistics highlight the severity of mathematics education in PE and suggest that the accountability framework of LSAs is having little, if any, effect on promoting student achievement in that province.

Results on these national and international assessments should be sounding an alarm for the province of PE; the children of PE are not being prepared for the 21st century. Unlike in the past, societies in the 21st century will rely on innovation, and the backbone of innovation is skills in mathematics and engineering (Peterson, Woessmann, Hanushek, & Lastra-Anadón, 2011). Much is to be learned about the state of education in PE, and the intention of this is study is to begin peeling back the complex layers by exploring the context of LSAs in general and, more specifically, the accountability framework that calls for incorporating a percentage of students' scores from LSAs to determine final grades.

Background to Large-Scale Assessments in PE

LSAs are referred to as "common assessments" in PE and were proposed for Grades 3, 6, and 9 as well as at the secondary level, in language arts and mathematics. In the spring of 2007, the PE Department of Education administered the first LSAs in Grade 3 reading and writing and Grade 9 mathematics. In the following year, a Grade 6 reading assessment was introduced. In 2009, the writing component of the Grade 6 literacy assessment was introduced to the roster, along with a Grade 3 mathematics assessment. In 2010, a Grade 6 mathematics assessment was developed for field-testing and administration in the fall of the same year. Although this staggered introduction of LSAs may have alleviated the intensity of introducing LSAs all at the same time or all in one grade level (e.g., Grade 3), it did not parallel the introduction of new curriculum. For example, the Grade 9 mathematics assessment was first introduced in 2007, but the curriculum reform in mathematics did not occur until three years later. As of 2011, curriculum reform had only occurred in selected grades and subjects, and provincial assessments had not been developed for any grades or subjects in the secondary division (note: Grade 9 falls in the intermediate division).

Of particular interest to this study is PE's Grade 9 Assessment of Mathematics, which was administered in June of each year. As previously noted, the first instrument was administered in 2007 and contained 62 criterion-based items (Miller, 2010), of which 56 were multiple-choice and 16 were short answer (dichotomously scored). This instrument was revised in 2008 to remove poor performing items and add new items. The revised instrument contained 58 items (48 multiple-choice and 10 short answer – all dichotomously scored). Given the absence of larger problem-solving items (e.g., *tasks*, as used in Ontario), the difficulty level of items on the Grade 9 Assessment of Mathematics was at the knowledge and application level (Miller, 2010). Student achievement (provincial average) on this LSA of mathematics was 59%, 62%, and 64% for the period 2008 to 2010. A more detailed analysis (distribution of students by achievement levels or cognitive complexity) was not available on the Department of Education's website. In 2010/2011, a new curriculum was introduced, along with a revamped LSA to be administered in June 2011 as a pilot instrument for the first year.

In 2009/2010 (i.e., the period of this study), a total of 48 Grade 9 mathematics teachers in PE were responsible for scoring their own students' Grade 9 Assessment of Mathematics. Students recorded answers on a response sheet separate from the assessment booklet. Using students' response sheets and an answer key provided by the Department of Education, teachers scored students' LSA of Grade 9 mathematics. It was believed that this procedure for scoring would maintain the security of the instrument and at the same time provide teachers with immediate scores for use in determining students' final grades (C. Wood, personal communication, April 21, 2011). Teachers did not receive any release time for scoring the Grade 9 Assessment of Mathematics; hence, the time lapse between administering and scoring the LSA could vary depending on teachers' schedules. Once the scoring of the LSA was completed, teachers were required to return the assessments to the Department of Education for official scoring by a marking board (C. Wood, personal communication, April 21, 2011).

Upon tabulating students' scores, teachers were required to allocate 10% of students' achievement on the LSA to their final grade in the course (PE Department of Education and Early Childhood Development, 2013). Students' final grade would then be composed of 10% LSA score and 90% term score. To receive a credit in Grade 9 mathematics, students must receive a passing grade of 50%. However, there was another guideline that prevented students from failing Grade 9 mathematics as a result of a low score

on the Grade 9 assessment test (C. Wood, personal communication, April 21, 2011). This guideline stems from a social promotion policy that is common in the early and middle grades (Leckrome & Griffith, 2006). This policy is based on the notion that it is better for students to proceed to the next grade level, regardless of academic ability, to avoid the negative effects of retention on students' self-esteem.

Based on the social promotion policy and the policy guiding teachers to determine final grades using 10% of students' LSA score, there is likely to be variability in determining students' final grades, particularly for those students near the cut score of 50%. It is unknown whether teachers adjust students' Grade 9 Assessment of Mathematics score or students' overall term score to arrive at a passing grade of 50% when students' grades would otherwise fall below.

The process of engaging teachers in scoring LSAs is related to using student achievement scores in determining students' final grades. As noted previously, teachers use students' answer sheets and an answer key to score the assessment. This method of scoring LSAs can potentially maintain the security of the instrument; however, it is unknown whether teachers refer to the actual instrument if they find a number of students responding incorrectly to a particular item or set of items, to gain insight into problem areas. It is suspected that teachers in other provinces who are not provided an answer key for scoring (e.g., ON) may need to examine the LSA items, especially if teachers have the discretion of deciding which items to include in determining student grades.

PE's LSA program is relatively new in comparison to programs in other provinces and territories that have been administering LSAs since prior to 2000. Being the last of the provinces and territories to implement LSAs, PE has the opportunity to learn from other provinces but also has its own challenges connected with implementation. The climate surrounding LSAs in PE is similar to what occurred in other provinces when LSAs were introduced: teachers there expressed resistance (Darling-Hammond, 2004; Kohn, 2000; Popham, 2001, 2004; Smith & Fey, 2000, Volante, 2004), as has PE's teachers' association (Horne, 2008).

Hundreds of thousands of dollars are spent each year on LSAs in PE with the aim of holding stakeholders accountable for educating the public. Little is known about whether the LSA initiative stemming from the task force report in 2006 is influencing teachers' instructional practices. Based on the most recent PISA and PCAP reports, we know that the state of education in PE is suffering. In fact, regressing may be a better de-

scriptor, given that student achievement on the 2009 PISA was lower than on the previous PISA (in 2006). As noted previously, there has only been one study (i.e., Miller, 2010) focusing on PE's LSA that examined the functionality and cognitive complexity of items on the Grade 9 Assessment of Mathematics. Further, there has been little research on the impact of low-stakes provincial assessments on teachers' professional development (Gambell & Hunter, 2004). This study will provide insight into PE teachers' dispositions towards LSAs and on the specific practice of using LSAs to determine students' grades. It is suspected that most (if not all) teachers' will report allocating 10% of students' final grade to the Grade 9 Assessment of Mathematics; however, the method of determining final grades for students' who do not meet the 50% mark is likely to vary, as is the reported impact of LSA on teachers' instructional practices.

Method

A two-phase approach employing questionnaires and interviews was used to survey teachers' beliefs, issues, and practices related to using an LSA (i.e., Grade 9 Assessment of Mathematics) to determine students' final grades. These two data sources were drawn together to respond to the research questions posed in the study.

Questionnaire

Survey items (see Appendix A for a copy of the survey) documented teachers' demographic information (e.g., grades taught, highest degree, school characteristics, etc.) and explored teachers' beliefs about LSAs (37 items) and issues related to LSAs (24 items). These items were drawn from an existing questionnaire exploring LSAs in Canada (Klinger, Rogers, Miller & DeLuca, 2008), while new items were created to focus on uses of the Grade 9 Assessment of Mathematics in the PE context. The five-point Likert scale designed to measure teachers' beliefs and identify issues presented a response continuum with endpoints labeled *not appropriate* to *very appropriate* (beliefs) and *not an issue* to *very serious issue* (issues). Such a scale would allow participants to attach meaning to the endpoints by partitioning the distance between the endpoints into equal units (Lam & Klockars, 1982). The stem for items exploring teachers' beliefs about LSA stated: "How appropriate do you believe are the following purposes and uses of the Grade 9 Common

Assessment?" The intention of this construct was to examine teachers' perceptions related to the purpose of LSAs in light of the intended purpose of promoting student achievement by holding educators accountable.

The stem for items exploring issues related to LSA stated: "How serious do you feel the following issues are for supporting the education of your students?" The intention of these items was to expand on findings from the first question. Items in this section of the questionnaire were intended to be more descriptive, in that they would highlight potential areas (issues) that could subsequently be addressed (e.g., professional development opportunities).

Two pre-service teachers volunteered to complete a "think-aloud" for the purpose of ensuring modifications made to the questionnaire reflected the PE context. Pre-service teachers were given general instructions to simply think aloud and verbalize their thoughts (Ericsson & Simon, 1993). Changes to the questionnaire focused predominantly on language. The phrase *Large-Scale Assessment*, for example, was used in a couple of instances but was not completely understood by pre-service teachers. This phrase was subsequently replaced by *Common Assessment* to better align the vocabulary to the PE context.

Questionnaire Analysis

A research assistant manually entered data and a second research assistant re-examined all data for potential entry errors; no errors were found. The data set was then analyzed to identify items that contained multiple missing responses that may have been caused by poorly worded or ambiguous items. Although not all questionnaires were completed entirely, a visual examination of the data did not reveal any patterns of missing responses that could have indicated a poorly performing item. All records were at least 75% complete, and for the records with missing responses, the data was handled using the listwise option in PASWStatistics 18.0, which excluded cases only when they were missing data for a specific analysis (Howell, 2007).

To determine whether items in the scale exploring beliefs related to the accountability purpose of LSAs were all measuring the same underlying construct, commonly known as the scale's internal consistency, Cronbach's alpha was calculated. This scale was considered reliable, with an alpha coefficient of 0.95. Descriptive statistics were then used to describe the characteristics of the sample, commencing with skewness and kurtosis values, which were calculated but only reported if an anomaly was found. Frequencies

(number of responses) and percent were then calculated for all items. Lastly, means and standard deviations were calculated for all continuous items.

To explore the relationship between the two categorical variables (i.e., "Do you feel the provincial testing programs have any value? (Yes/No)" and "As a teacher, do you actually use the results? (Yes/No)") a cross-tabulation was used. To test for differences in beliefs related to the accountability purpose of LSAs and the same two categorical variables stated above, a Mann-Whitney U Test was used to compare the medians of both groups. Although this non-parametric test does not make assumptions about the underlying population distribution and is less sensitive to differences between groups, it is purposeful when working with small data sets, as was the case in this study. It was hypothesized that teachers with strong beliefs about using LSAs for the purpose of accountability would report using the results and finding value in them.

Questionnaire Participants

In 2009, 113 teachers taught mathematics in grades 7, 8, or 9 (referred to in PE as the middle or intermediate years). Although some teachers taught only Grade 9 mathematics, most taught mathematics to at least two grade levels and some to all three, which is indicative of the small school settings in PE. Forty-eight of these teachers were registered by the Department of Education as being teachers of Grade 9 mathematics.

In April 2009, 85 of the 113 middle years mathematics teachers in PE attended a provincial conference on mathematics and assessment. The paper-and-pencil questionnaire was distributed to the 85 teachers attending the conference. Participants were reminded throughout the day to complete the questionnaire; no remuneration was given. Of the 67 teachers who returned the questionnaire, 24 out of a possible 48 taught Grade 9 mathematics. The beliefs, issues, and reported practices of these Grade 9 teachers were the focus of this study. This sample comprised 50% of the Grade 9 mathematics teachers in PE, which was representative given that a small sample provides proportionately more information for a small population than for a large population (Lenth, 2001).

Interviews

Semi-structured interviews were designed to expand on areas connected to the questionnaire. Four interview questions were created following the analysis of the questionnaire.

The first interview question ("On average, how long does it take to score one student's Grade 9 Common Assessment of Mathematics?") was designed to ease participants into the interview process while at the same time providing some background related to the scoring of the test. The next two questions (a copy of the interview questions are found in the section on findings) were intended to confirm findings from the questionnaire and expand on experiences related to scoring the Grade 9 Assessment of Mathematics. The last question was also intended to expand on findings from the questionnaire by exploring other ways, not addressed in the questionnaire, that teachers use the Grade 9 Assessment of Mathematics to enhance practice.

Interview Participants

Teachers were strategically selected by organizing schools into three categories: (a) well above the provincial average, (b) at or within plus or minus three points of the provincial average, and (c) well below the provincial average. Two teachers were randomly selected from each category and contacted by either e-mail or telephone. They also received a letter of information that described the scope of the study. Only four teachers agreed to participate. A number of teachers did not return our telephone calls, and three who did contact us refused to participate, citing political issues.

For those who agreed to participate in an interview, an agreeable time and place was determined. All interviewees consented to having the interview audio recorded. Each teacher participating in an interview received a follow-up thank-you note, along with a $25 gift card.

Interview Analysis

The interviews were audio recorded and transcribed. Conceptual analysis was used to determine the presence of common words or phrases in the transcriptions so as to draw inferences about the participants' views. The coding commenced with predefined categories but was flexible to allow for the addition of unforeseen categories as well as differences in phrasing (e.g., "I'm just going with the 10%" and "the score is what it is" both endorse the practice of using 10% of the LSA result to determine a student's grade).

Findings

Questionnaire Demongraphics

Two of the 24 teachers taught in the French Immersion Program. Teachers were from schools with a variety of socioeconomic status characteristics, such as family income, academic achievement, and geographic locations (i.e., urban, suburban, semirural, or rural). Only one teacher indicated their school received some form of Department of Education initiative or intervention, while 14 teachers indicated their school did not receive any initiatives and nine teachers did not know whether the school received any initiatives.

Beliefs Related to the Accountability Purpose of LSAs

A descriptive analysis of the 35 items surveying beliefs related to the accountability purpose of LSAs revealed a clustering of responses at the low end of the five-point Likert scale. The highest response on the scale for all 35 items was three. This clustering at the low end of the scale resulted in low mean scores and relatively small standard deviations ($M = 1.98$, $SD = 0.445$) for the total scale score. These response patterns clearly indicated this sample of Grade 9 mathematics teachers did not strongly believe in using LSAs for accountability purposes. Further inquiry was required to examine the implications of such strong views. Specifically, there was a need to examine how teachers dealt with the juxtaposition between this view and the provincial guideline calling for incorporating 10% of students' score from the LSA to determine students' final grades. To expand on this finding, interview questions were created to further explore this practice.

Next, a cross-tabulation was used to probe the relationship between the two categorical variables (i.e., "Do you feel the provincial testing programs have any value? (Yes/No)" and "As a teacher, do you actually use the results? (Yes/No)"). This test was significant ($p = 0.015$), with 73.9% (17 out of 23 who responded) reporting they felt the LSA had value and they actually used the results (see Table 1). These findings seem to contradict the negative views towards the LSA that teachers reported on the scale above. In one measure (i.e., the total scale score of beliefs towards the accountability purpose of LSAs), teachers do not overwhelmingly support the accountability purpose of LSAs, but on this second measure (i.e., cross-tabulation), the majority of teachers indicated LSAs had value and they used the results.

Table 1: Cross-Tabulation: Value vs. Use

		Do you feel that provincial testing programs have any value? (Item A)		
		Yes	No	**Total**
As a teacher, do you actually use the results? (Item B)	Yes	17	0	**17**
	No	3	3	**6**
	Total	**20**	**3**	**23**

To test for differences in the beliefs scale examining teachers' perceptions of the accountability purpose of LSAs and the same two categorical variables stated above (i.e., Items A and B), a Mann-Whitney U Test was performed to compare the medians of both groups. It was hypothesized that teachers who felt the accountability nature of the LSA was inappropriate (i.e., responding at the low end of the scale) would indicate the LSA had no use or value. These comparisons were not significant ($p < 0.05$). It is possible that this less sensitive, non-parametric test failed to detect differences between the two groups, which may have been further exacerbated by the clustering of responses at the low end of the Likert scale.

Issues Related to Provincial Assessment

Of the 24 items examining issues related to the Grade 9 Assessment of Mathematics, 11 items had a mean score of three or higher. These items were identified as issues for Grade 9 mathematics teachers. The 11 items are presented in Table 2.

Teachers identified the practice of incorporating students' score on the Grade 9 Assessment of Mathematics (Table 2, Item 6) as being an issue. This finding is of interest because teachers were required to follow a Department of Education guideline that called for the practice of incorporating students' LSA scores with their term scores. Given this juxtaposition, the extent to which teachers followed this guideline was selected for follow-up in the interviews. What remained unanswered was whether teachers actually followed the guideline, since schools do not use a common mark-recording software that would automatically include 10% of a student's score on the Grade 9 Assessment of Mathematics with the student's term score.

Table 2: Issues Related to the Provincial Common Assessment

Item #	<- Not an issue . . . Very serious issue -> 1	2	3	4	5	M	SD
1. Provincial testing narrows the teaching of the curriculum at the grade levels at which the test is given.	4 (16.7)	4 (16.7)	3 (12.5)	7 (29.2)	4 (16.7)	3.14	1.42
3. Results of provincial testing are used to publicly rank schools.	5 (20.8)	3 (12.5)		6 (25.0)	7 (29.2)	3.30	1.65
4. Results of provincial testing are used to evaluate teacher effectiveness.	3 (12.5)	3 (12.5)	3 (12.5)	5 (20.8)	7 (29.2)	3.48	1.47
5. Provincial tests are used as an accountability tool.	1 (4.2)	2 8.3)	6 25.0)	9 (37.5)	2 (8.3)	3.45	1.00
6. Results from the provincial tests are encouraged to be included in the students' final grade.	2 (8.3)	1 (4.2)	5 (20.8)	9 (37.5)	3 (12.5)	3.5	1.15
7. Provincial test results provide a "snap-shot" of what students know and can do.	3 (12.5)	1 (4.2)	6 25)	8 (33.3)	3 (12.5)	3.33	1.24
8. Teachers teach towards the test.	3 (12.5)	2 (8.3)	6 25.0)	7 (29.2)	2 (8.2)	3.15	1.23
9. Classroom activities are limited to the learning expectations assessed on Provincial Assessments.	2 (8.3)	4 (16.7)	9 (37.5)	3 (12.5)	3 (12.5)	3.05	1.16
10. Classroom assessment instruments reflect item format and content on provincial tests.	2 (8.3)	3 (12.5)	7 (29.2)	6 (25.0)	1 (4.2)	3.05	1.08
16. Students are excluded from participating in order to improve school results.	4 (16.7)		4 (16.7)	6 (25.0)	5 (20.8)	3.42	1.47
22. The press ignores the limitations of results when publishing rankings of schools based on provincial test results.	1 (4.2)	1 (4.2)	2 (8.3)	5 (20.8)	12 (50.0)	4.24	1.14

Although the most contentious issue did not relate to the 10% guideline (i.e., Table 2, Item 22), it is important to acknowledge that teachers strongly believed the press ignored the limitations of the results when publishing school rankings. It is interesting that teachers rank this item as being more of an issue than the other 10 items in Table 2. One may speculate that the accountability aspect of publicizing the Grade 9 Assessment of Mathematics results was having more impact on teaching practices than the opportunities provided through the assessment itself (i.e., support curriculum implementation [PAB4][2]; focused instruction on the provincial curriculum [PAB5]; or improved and enhance teaching [PAB7]). Items 3, 4, and 5 also explored the accountability issues of the Grade 9 Assessment of Mathematics and similarly have mean scores higher than the other issues presented in Table 2.

While it is important to highlight teachers' issues related to the Grade 9 Assessment of Mathematics, it is equally important to identify areas teachers did not feel were issues. Based on the 24 items focusing on issues, 10 items had scores less than three. Of these items, three refer to uses of the Grade 9 Assessment of Mathematics. These items are summarized in Table 3.

Table 3: Non-Issues Related to the Grade 9 Assessment of Mathematics

Item #	<- Not an issue . . . Very serious issue ->					M	SD
	1	2	3	4	5		
18. Due to the nature of the reported results, the data cannot be used to support instruction.	5 (20.8)	5 (20.8)	7 (29.2)	3 (12.5)	1 (4.2)	2.52	1.17
19. Teachers do not know how to interpret the assessment results.	7 (29.2)	6 (25.0)	5 (20.8)	2 (8.3)	1 (4.2)	2.24	1.18
20. Teachers do not know how to use the assessment results.	5 (20.8)	9 (37.5)	3 12.5)	3 (12.5)	1 (4.2)	2.33	1.16

Based on these findings, teachers indicated the results can be used to support instruction, and they reported knowing how to interpret and use the results. This area was explored further in the interviews (i.e., interview Item 4) by asking teachers to describe ways in which they were using results from the LSAs.

2 Note: The code [PAB] refers to items organized by the construct, Provincial Assessment Beliefs. The number following the code refers to the item number for that construct.

Interviews

Of the four teachers who accepted the invitation to participate in the study, Teacher 1 was from a school with eight classes of Grade 9 students (for a total of 167 students), whose school average on the Grade 9 Assessment of Mathematics was the same as the provincial average (i.e., 64%). Teacher 2 was from a smaller school with only 88 Grade 9 students, corresponding to five Grade 9 classes; their average on the Grade 9 Assessment of Mathematics was 58%. Teachers 3 and 4 were from even smaller schools, with, respectively, 16 and 21 individuals equating to one class of Grade 9 students. The average on the Grade 9 Assessment of Mathematics for these schools exceeded the provincial average (77% and 69%, respectively). Unfortunately, the limited timeframe prevented our gaining input from a teacher in a large school (i.e., with three or more Grade 9 classes) whose students' average on the Grade 9 Assessment of Mathematics exceeded the provincial average; such data would have better represented the views of teachers in different school sizes.

Interview Question 1: On average, how long does it take to score one student's Grade 9 Common Assessment of Mathematics?

Three teachers indicated the process of scoring the Grade 9 Assessment of Mathematics took approximately three to 10 minutes per student. In comparison, the fourth teacher indicated he spent approximately half an hour per student. This teacher elaborated further noting that, teachers at his school received a full day to administer and score the Grade 9 Assessment of Mathematics. Given the number of students in Teacher 4's class, it would not be possible to spend 30 minutes scoring each test as well as administer the test in one day. It is possible that Teacher 4 over estimated the time spent scoring.

Interview Question 2: Teachers are required to score the Grade 9 Assessment of Mathematics. Tell us about this experience.

All four interviewed teachers indicated their experience scoring the Grade 9 Assessment of Mathematics was positive; however, interview probes provided more insight into how the experience influenced their practice. Specifically, teachers identified two areas of influence: adoption of the multiple-choice format on classroom assessments and reviewing items on the Grade 9 Assessment of Mathematics to inform instructional practice.

All four teachers reported they adopted the multiple-choice format on classroom assessments. Teacher 3 commented, "They do some multiple-choice throughout the year on the tests that I give them." Another teacher indicated they have "put in a section of multiple-choice and short-answer questions [in almost every chapter], that gets away from how we would normally test on the process" (Teacher 1). One teacher acknowledged not using the multiple-choice format but rather changed classroom practice to reflect the item format on the Grade 9 Assessment of Mathematics. This teacher also noted that the 2011 curriculum resources provided a databank and blackline masters of multiple-choice questions, making it easier for teachers to include this item format on their classroom assessments. Another teacher noted that the low literacy rate in PE might hinder students' success on multiple-choice items, given the amount of reading involved in comparison to a constructed response question. This teacher stated "a lot of them not only have difficulty with math but they have difficulty with literacy, and all the multiple-choice questions… you have to be able to read and understand what is being asked" (Teacher 1).

Although the practice of reviewing items to identify students' common areas of strengths and weaknesses infringed on the security of items, the practice may be viewed as a professional development opportunity by teachers who do not understand the sensitivity of item security. When asked, all teachers indicated that they reviewed items on the LSA. One reported examining items to determine "whether the question was valid or not" (Teacher 1,) in addition to examining items on which students did not score well. Another teacher reported reviewing the items because "it gives me an idea of where we've been successful and where to strengthen our students" (Teacher 2). Teacher 3 reported reviewing the LSA in terms of item difficulty. The rationale given for this practice was to guide the range of item difficulty presented on classroom tests. The fourth teacher, who initially denied reviewing items on the LSA, felt the information gained from doing so "is not applicable to the current class" (Teacher 4); hence, there was no need to review items. However, this teacher also reflected, "I should be doing more of that." The interviewer then asked this teacher: "Do you look back on the assessment questions in terms of content . . . whether you have covered an area or not covered an area enough?" The teacher replied:

> We get them [Grade 9 Assessment of Mathematics] well in advance, so we know what they are going to be tested on. And I mean, if you are scrambling for time, sometimes you will set aside something you may have wanted to spend a week

> on. Now you can spend less than a week, maybe half a week. I remember a couple of years ago, there was a question on a sphere, and we don't generally go over geometry until the very end [of the term]. So then when June hits, because you are in Grade 9 you are pushing for the closing celebration and everything else is going on. So what I ended up doing last year, I just—which we are allowed to do, I put the formula up on the board. I did one problem in advance that taught them how to do a sphere. (Teacher 4)

Even though students receive a formula sheet along with the assessment booklet (C. Wood, personal communication, April 21, 2011), this teacher felt the need to emphasize the formula for a sphere, likely because the topic was so new to students.

Although all teachers indicated directly or indirectly that they examined items from the LSA, there was insufficient evidence to indicate whether the viewing of items (before or after administering the LSA) influenced their class average on the LSA. However, it is possible that this practice may be contributing to grade inflation of provincial LSA scores, thereby generating higher provincial averages for student achievement in mathematics ability than what is reported on national or international assessments. It is also important to note that the provincial, national, and international assessments of mathematics ability may be measuring different content areas and cognitive abilities. Much more research is needed in this area to draw any decisive conclusions related to this practice.

Interview Question 3: On a previous questionnaire, the majority of PE teachers indicated that the Grade 9 Common Assessment of Mathematics was not appropriate to determine students' final grades. How do you deal with this juxtaposition between what you believe and what is called for in practice?

All teachers appeared to be resigned to the guideline directing them to incorporate 10% of students' scores on the Grade 9 Assessment of Mathematics with their term scores in the course. The second teacher's response resonated well with the other teachers: "It's pretty straightforward. We're told it's going to be 10%, so you have to prepare the students as best you can . . . that's the way it is." Teacher 1 concurred with the guideline, stating: "It's a sign of the times. Everybody is clamouring for this common assessment because everybody is looking for accountability." Echoing this consensus, Teacher 4

suggested there was a place for the 10% guideline because of students' and parents' heightened focus on grades.

A second theme among the four interviewees was that "students' mark on the exam was relatively similar to what they had been getting all along. Maybe a little bit lower. So I agree that 10% is fine for marking them, and you still give them a chance to make their marks up in assignments, and I tend to give little quizzes, so that kind of thing as well" (Teacher 3). This perspective indicated that teachers were not opposed to allocating 10% of the Grade 9 Assessment of Mathematics score to students' term score, but compensated for lower scores by giving, as described by the teacher, opportunities to increase students' term grade. In sum, it appeared that teachers may not be wholeheartedly endorsing the 10% guideline, but based on the responses from the four teachers interviewed, they were implementing the guideline.

To explore the depth of this practice, one of the probes inquired whether teachers would alter the 10% guideline if a student's Grade 9 Assessment of Mathematics score fell far below the overall class score. Teacher 3 replied that this scenario did not apply because there was little difference between their students' Grade 9 Assessment of Mathematics scores and overall term scores. In one case, two students "got under 50 on the exam of Grade 9 Assessment of Mathematics but it didn't make them fail—well, actually, one, but that didn't make or break their year" (Teacher 3). This teacher suggested that course grades were increased to compensate for low scores so that all students were promoted to Grade 10. Teacher 2 stated: "If it says 10%, I'm just going to go with the 10%." Teacher 1 had a different practice: "I would never adjust their score, 'cause the score is what it is. I may adjust how much that is actually worth to the student" (Teacher 1). This line of questioning provided some evidence that teachers were not adjusting students' Grade 9 Assessment of Mathematics score but, in some cases, may be adjusting or reducing the weight of the Grade 9 Assessment of Mathematics to align it with the guideline, preventing students from failing the course as a result of a low score on the Grade 9 Assessment of Mathematics.

Interview Question 4: Describe other ways in which the Grade 9 Assessment of Mathematics influences your instructional practices.

For the most part, this interview question generated comments similar to what had already been shared, as described above. However, Teacher 1 mentioned using a Department of

Education supplement that provided "practical tips on how to approach multiple-choice tests." Teacher 4 reported that the Grade 9 Assessment of Mathematics kept him on track: his goal was to be at specified curriculum checkpoints throughout the course because the Grade 9 exam, Assessment of Mathematics, covered the curriculum. "It's published and it will show what areas [students'] struggle with because I missed [teaching] it or there wasn't enough time to do it." Holding teachers accountable by publicizing results appeared to be impacting teachers' practice. Teacher 2 added that the Grade 9 Assessment of Mathematics had caused anxiety in some students: "I found myself trying to calm the waters, and then again you have ones [students] that are indifferent, so you don't have to worry about them at all. So the ones that are really keen, it's going to be okay, and maybe anxious is a better word." Lastly, Teacher 3 was influenced by the Grade 9 Assessment of Mathematics to "do a lot of questions straight out of the curriculum guide." This teacher also mentioned reviewing the exams from previous years to

> figure out where the time should be spent. For example, fractions [are] one big section that kids struggle with. So I spend extra time on fractions, and I tend to do a lot of whiteboard work with fractions. Quick little ten-minute things in the morning, where I put them up. And I go back throughout the year, on fractions, just as a ten-minute review every day. (Teacher 3)

In hindsight, it is unclear whether this teacher's reference to the *exams* was meant to indicate the Department of Education supplements for Grade 9 mathematics or the actual provincial assessment results showing how well students scored in each strand of mathematics.

Discussion

Research Question 1*:* To what extent do teachers' beliefs about the LSA affect their reported practice of using LSA scores?

PE does not monitor whether teachers are incorporating 10% of students' scores from the LSA of Grade 9 Mathematics into their term scores to produce final grades. Further, no research has explored (i) whether teachers believe in or endorse the accountability aspect

of LSAs, and (ii) the implications of their beliefs on their practices. Findings from this study suggest that while teachers do not entirely endorse using the Grade 9 Assessment of Mathematics in determining students' final grades, they are implementing the guideline.

When asked on the questionnaire whether this practice was appropriate or not, 15 out of 23 teachers who responded to Item PAB16 (i.e., "How appropriate is using the Grade 9 Assessment of Mathematics to help determine final grades?") reported this practice was inappropriate (i.e., rated the appropriateness at the low end of the scale; $M = 2.13$, $SD = 1.12$). This practice was further explored in terms of whether teachers believed the practice was an *issue*. Similar to the item above, responses were not positive: 12 of 23 teachers indicated this practice was a fairly serious issue (i.e., rated as a serious or very serious issue), and only three of 23 teachers rated the practice as a non-issue (IRPA6, $M = 3.5$, $SD = 1.15$). Based on these findings and the overall low belief in the Grade 9 Assessment of Mathematics, teachers were not wholeheartedly endorsing the practice of incorporating 10% of the Grade 9 Assessment of Mathematics scores with students' term scores.

A non-significant Mann-Whitney Test, used to compare the beliefs scale with mean scores of teachers who reported *using* LSA results, indicated no difference in disposition between teachers who reported using the LSA results and teachers who reported they did not. Similarly, there was no significant difference in disposition between teachers who reported the LSA had *value* and those who felt it was not valuable. Although it was hypothesized that teachers who endorsed the LSA—as shown by indicating that they used the LSA results and felt LSAs had value—would be more accepting of the accountability nature of LSAs, this was not the case in this study. As noted previously, it is important to be cautious with these findings, given the clustering of responses at the negative end of the Likert scale used to measure beliefs, as well as the less sensitive nature of this non-parametric test.

The four interviews extended what teachers had reported on the questionnaire. All four teachers responding to interview question 3 (i.e., how they dealt with the juxtaposition between what they believed and what was called for in practice) indicated they were following the guideline even though they did not fully endorse the practice. Two teachers provided further insight into how they implemented the 10% guideline along with a second guideline that prevented students from failing Grade 9 mathematics due to a low score on the provincial assessment. Teacher 1 stated that the score students received on the Grade 9 Assessment of Mathematics was a hard piece of evidence, not to be altered;

however, in the scenario posed during the interview (i.e., when a student's Grade 9 Assessment of Mathematics score was much lower than the student's score in the course), this teacher would lower the weight of the Grade 9 Assessment of Mathematics in calculating the student's overall score. Hence, this teacher acknowledged the reliability and validity of the LSA but still chose to alter its weight as opposed to adjusting the weight of the potentially less reliable term score.

Unlike Teacher 1, Teacher 3 chose to adjust students' term scores. When presented with the same scenario, Teacher 3 described a practice of raising course grades to prevent students from receiving an overall failing grade (i.e., less than 50%), by assigning extra assignments that would allow students to increase their term scores.

It is unclear whether these teachers believed they were operating within both guidelines, because none of the teachers defended their practices by citing either guideline. Practice of the 10% guideline was not clear-cut. Although it appeared teachers were implementing that guideline, there was some evidence of other adjustments to determining students' final grades in response to the guideline preventing students from failing Grade 9 mathematics due to a low score on the provincial assessment. In one case, a teacher lowered the weight of the Grade 9 Assessment of Mathematics to ensure students received a passing grade. In a second case, a teacher described a practice known as grade inflation, whereby students' term scores were raised by providing them with opportunities to complete additional assignments that would produce a higher term score. It is important to note that these two practices involving an adjustment of scores may be isolated to PE. In this province, promotion to Grade 10 (high school) is dependent on a passing score in Grade 9 mathematics. In other provinces where Grade 9 is the first year of high school, students have more opportunities to retake a course if they are not successful. Hence, the manner in which teachers outside of PE determine students' grades may not be influenced by social promotion.

Lastly, most teachers reported the Grade 9 Assessment of Mathematics did not motivate students to work hard. In hindsight, this item (i.e., PAB17) could have been worded differently to investigate whether the practice of using 10% of students' Grade 9 Assessment of Mathematics score (a) motivated students to work hard in the course, (b) motivated students to put forth good effort to complete the Grade 9 Assessment of Mathematics, and (c) produced a more accurate picture of students' abilities in mathematics. To explore the impact of the 10% guideline on students, further research is needed—similar

to the study conducted by van Barneveld and Brinson (2011), wherein a large sample was obtained by incorporating student questionnaires with the LSA. Specifically, it would be beneficial to survey students' views on (i) whether the 10% guideline motivates them to do their best on the Grade 9 Assessment of Mathematics, (ii) the impact of the weight (e.g., 10%) allocated to their LSA (e.g., would a greater weight motivate them more?), and (iii) whether they believe the LSA to be an accurate reflection of their mathematics abilities.

Research Question 2*:* To what extent do teachers in PE incorporate students' scores from the LSA of Grade 9 mathematics with students' term scores to determine final grades?

In the province of PE, teachers are required to score the Grade 9 Assessment of Mathematics using only students' response sheets and an answer key provided by the Department of Education. The purpose of this practice was to provide teachers with immediate student scores for use in determining final grades. The protocol surrounding this practice would likely result in little impact on teachers' instructional practices, given the mechanical nature of indicating whether students' responses are correct or incorrect. Factors that have the potential to impact teaching practices are the support material (i.e., practice assessment booklets) and results (organized by curriculum strand) accompanying PE's LSA. This study discovered an unanticipated impact of the LSA on teaching in PE that involved examining LSA items for content areas and difficulty levels assessed.

Based on responses from the questionnaire, teachers indicated they found some uses for the Grade 9 Assessment of Mathematics, although these uses were not rated at the high end of the five-point Likert scale. Teachers believed the Grade 9 Assessment of Mathematics could improve and enhance teaching ($M = 3.63$, $SD = 1.17$), provide data for data-based decision making at the class level ($M = 3.96$, $SD = 1.12$), and increase teachers' assessment knowledge and skills ($M = 3.48$, $SD = 0.994$). These uses of the Grade 9 Assessment of Mathematics do not necessarily correspond to teachers' scoring of the assessment but, as noted previously, are more likely connected to the preparation material (i.e., practice assessment booklets) and school reports (by curriculum strand) provided by the Department of Education. Given the somewhat positive but moderate responses, more support in these area may have greater impact on teaching and learning. For example, distributing preparation material at the beginning of a term, such as 10 multiple-choice items organized by strand or teaching units, could provide additional

support for this criterion-referenced LSA. Another way to use LSAs to support teaching is by providing an additional report to teachers, identifying common student errors. Such a report focusing on what students did incorrectly and how to correct their thinking may prove to be a useful resource with greater potential to impact teaching than simply reporting student achievement by strand.

Teachers who participated in the interviews all indicated the Grade 9 Assessment of Mathematics was impacting their instructional practices because they were now incorporating multiple-choice items on classroom assessments. The format of the LSA (i.e., primarily multiple-choice) and the provision of sample items in students' practice booklets likely influenced the adoption of this item format on classroom assessments. One teacher reported that the new curriculum resource (i.e., implemented in the fall of 2010) provided sample multiple-choice items in a data bank, facilitating the use of this item format on classroom assessments. It is unclear how frequently teachers use these and other resources and whether or not they create their own multiple-choice questions. Further, it is unknown whether teachers are able to differentiate between poor- and high-quality multiple-choice items or whether there is an over-reliance on multiple-choice items and not enough emphasis on other item formats, as some are claiming in the United States (Sawchuck, 2006). There is still much to be learned about the impact of multiple-choice format on teachers' instructional and assessment practices. More importantly, if LSAs are only influencing the item format teachers use in their daily assessment practices, then one must question whether this is a sufficient impact on teaching to warrant the allocation of funding to LSAs. As discussed above, there are other ways to positively impact teachers' practices through LSAs, but more focus is needed on better utilizing LSA resources to influence teaching practices.

This last point of discussion is rather sensitive and likely to brew heated conversations surrounding testing security and ethical practices. In the interviews, three of the four teachers reported they were examining items on the LSA, but each aiming to glean different information. One teacher acknowledged he had reviewed the LSA prior to administering the assessment and altered his instruction to better prepare students for the test. During the scoring phase of the LSA, two teachers reported examining items on the Grade 9 Assessment of Mathematics for item difficulty level, and three teachers reported examining items on which students scored poorly or well, based on student response patterns.

Are these practices unethical or merely naive on the part of teachers in a province that is new to LSAs? These interviews have revealed a clear lack of coherency regarding fair administration and scoring practices. One explanation for the incoherency could be that PE teachers are new to LSA processes and inappropriately assumed they were permitted to examine items either before or following administration of the LSA. A more plausible explanation is that teachers have not been given sufficient direction and training by provincial authorities, along with quality assurance supervision by school administrators, both of which would contribute to ethically robust assessment procedures. Irrespective of what influenced teachers' practice of examining LSA items, this finding raises concerns regarding item security during teachers' administration and scoring of LSAs.

Conclusion

The outcome of this study is intended to provide background information to guide decisions related to the validity of this practice in other jurisdictions. Based on the findings in this study, teachers did not necessarily agree with the accountability nature of LSAs but, for the most part, they were abiding by the 10% guideline. Echoing what was found in a previous study (i.e., van Barneveld & Brinson, 2011), more direction and guidance surrounding the practice of teachers' grading of LSAs is needed to reduce variability in the practice from one teacher to the next. It is important to note that although this study found teachers were following the Department of Education guideline calling for allocating a percentage of students' LSA scores to determine their final grades, this is not an indicator that the curriculum is being covered and outcomes are being met. Rather, it suggests teachers have accepted the practice and the notion of holding students accountable for their learning, believing that students will put forth more effort the more an assessment is worth. More research is needed to explore student perceptions of this belief. For example, how would a weight of 20% compared to a weight of 10% affect students' preparation for and effort on the LSA? Although students have been drawn into the accountability framework of LSAs, more research is needed to understand the impact on their learning and achievement.

References

Alberta Education. (2007). Reporting provincial achievement test results in June pilot evaluation 2005 – 2006. Retrieved from http://education.alberta.ca/media/615086/junereport.pdf

British Columbia Ministry of Education. (2004). *Policy document: Large-scale assessment*. Retrieved from www.auditor.on.ca/en/reports_en/en09/304en09.pdf

Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessment that supports learning: What will it take? *Theory Into Practice, 42*(1), 75–83.

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record, 106*(6), 1047–1085.

Decker, D., & Bolt, S. E. (2008). Challenges and opportunities for promoting student achievement through large-scale assessment results: Research, reflections, and future directions. *Assessment for Effective Intervention, 34*(1), 43–51.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: Bradford Books/MIT Press.

Fushell, M. (2011, April). *Teachers' determination of final grades using large-scale assessments: Policy implications.* Paper presented at American Education Research Association, New Orleans, LA.

Gambell, T., & Hunter, D. (2004). Teacher scoring of large-scale assessment: Professional development or debilitation? *Journal of Curriculum Studies, 36*(6), 697–724.

Horne, P. (2008). *Opening address.* Prince Edward Island Teachers' Federation.

Howell, D. (2007). The treatment of missing data. In Outhwaite, W., & Turner, S. (Eds.), *Handbook of Social Science Methodology* (pp. 208–224). London: Sage.

Klinger, D., DeLuca, C., & Miller, T. (2008). The evolving culture of large-scale assessments in Canadian education. *Canadian Journal of Educational Administration and Policy, 76*, 1–34.

Klinger, D., Rogers, W. T., Miller, T., & DeLuca, C. (2008). Large-scale assessments in Canada. *NCME Newsletter, 16*(3), 9–14.

Koch, M. J. (2011a). *One test, two scores: Dilemmas emerging from teachers' use of a large-scale mathematics assessment as part of students' grades.* Paper presented at the American Educational Researchers Association, New Orleans, LA.

Koch, M. J. (2011b). *Teachers' use of large-scale assessment as part of students' grades: Case studies of teacher practices in Ontario schools*. Paper presented at CSSE Fredericton.

Kohn, A. (2000). *The case against standardized testing.* Portsmouth, NH: Heinemann.

Kuriel, R. (2005). *Excellence in education: A challenge for Prince Edward Island. Final report of the Task Force on Student Achievement.* Retrieved from www.gov.pe.ca/photos/original/task_force_edu.pdf

Lam, T. C. M., & Klockars, A. J. (1982). Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurement, 19*(4), 317–322.

Leckrone, M. J., & Griffith, B. (2006). Retention realities and education. *Children and Schools, 28*(1), 53–58.

Lenth, R. V. (2001). Some practical guidelines for effective sample-size determination. *The American Statistician, 55*(3), 187–193. Retrieved from http://socrates.berkeley.edu/~maccoun/PP279_Lenth.pdf

Miller, T. (2010, May). *Provincial assessments in Prince Edward Island: Three years of Grade 9 mathematics assessment.* Paper presented at the Canadian Society for Studies in Education, Montreal, QC.

Ontario Ministry of Education. (2009). *Annual report of the Office of the Auditor General of Ontario*. Retrieved from www.auditor.on.ca/en/reports_en/en09/2009AR_en_web_entire.pdf

Pan-Canadian Assessment Program (PCAP). (2007). *PCAP-13 2007: Report on the assessment of 13-year-olds in reading, mathematics, and science.* Retrieved from www.cmec.ca/Publications/Lists/Publications/Attachments/124/PCAP2007-Report.en.pdf

Pan-Canadian Assessment Program (PCAP). (2010). *PCAP-13 2010: Report on the assessment of 13-year-olds in reading, mathematics, and science.* Retrieved from www.cmec.ca/Publications/Lists/Publications/Attachments/274/pcap2010.pdf

Peterson, P., Woessmann, L., Hanushek, E., & Lastra-Anadón, C. (2011). *Globally challenged: Are US children ready to compete?* Retrieved from www.hks.harvard.edu/ pepg/PDF/Papers/PEPG11-03_GloballyChallenged.pdf

Popham, W. J. (2001). Teaching to the test. *Educational Leadership, 58*(6), 16–20.

Popham, W. J. (2004). All about accountability / Why assessment illiteracy is professional suicide. *Educational Leadership, 62*(1), 82–83.

Prince Edward Island Government. (2006, April). *Education budget supports literacy, student achievement, and post-secondary learning*. News release. Retrieved from www.gov.pe.ca/newsroom/index.php?number=news&dept=&newsnumber=4547

Prince Edward Island Government. (2010). *PE demographics and labour force statistics*. Retrieved from www.gov.pe.ca/photos/original/PEDLF.pdf

Prince Edward Island Department of Education and Early Childhood Development. (2009). *Atlantic Canada mathematics curriculum guide: Introduction*. Retrieved from www.gov.pe.ca/photos/original/ed_math_intro.pdf

Prince Edward Island Department of Education and Early Childhood Development. (2013). *Intermediate Mathematics Assessment 2013 Administration Guide*. Retrieved from www.gov.pe.ca/photos/original/eecd_imaguide.pdf

Prince Edward Island Department of Education and Early Childhood Development. (2013). *Information for parents of grade 9 students.* Retrieved from www.gov.pe.ca/photos/original/eecd_parent_let.pdf

Programme for International Student Assessment (PISA). (2003). *Literacy skills for the world of tomorrow: Further results from PISA 2000*. Retrieved from www.oecd.org/edu/school/2960581.pdf

Programme for International Student Assessment (PISA). (2005). *Problem solving for tomorrow's world: First measure of cross-curricular competencies from PISA 2003.* Available from www.oecd-ilibrary.org/content/book/9789264006430-en

Programme for International Student Assessment (PISA). (2006). *Science competencies for tomorrow's world*. Available from www.oecd-ilibrary.org/content/book/9789264040014-en

Sawchuck, S. (2006). Does NCLB create state reliance on multiple-choice? *Education Daily, 39*(30), 3.

Simon, M., van Barneveld, C., King, S., & Nadon, C. (2011). *Having large-scale assessment results count for students' final grades: A matter of policy*. Paper presented at the American Education Research Association, New Orleans, LA.

Smith, M. L., & Fey, P. (2000). Validity and accountability in high-stakes testing. *Journal of Teacher Education, 51*(5), 334–344.

Statistics Canada. (2010). *Visual census. 2006 census*. Retrieved from www12.statcan.gc.ca/census-recensement/2006/dp-pd/fs-fi/index.cfm?Lang=ENG&TOPIC_ID=6&PRCODE=01

Student Achievement Indicators Program. (2001). *Report on Mathematics Assessment III.* Retrieved from www.cmec.ca/Publications/Lists/ Publications/Attachments/8/saip2001math.en.pdf

Student Achievement Indicators Program. (2002). *Report on Writing Assessment III*. Retrieved from www.cmec.ca/Publications/Lists/Publications/ Attachments/7/saip2002.en.pdf

Van Barneveld, C., & Brinson, K. (2011). *The rights and responsibilities of test takers when some parts of a large-scale test count toward class marks.* Paper presented at the American Education Research Association, New Orleans, LA.

Van Barneveld, C., King, S., & Nadon, C. (2011). *Teachers' use of large-scale assessment results in Ontario: Grading issues and polices.* Paper presented at the American Education Research Association, New Orleans, LA.

Volante, L. (2004). Teaching to the test: What every educator and policy maker should know. *Canadian Journal of Educational Administration and Policy*, *35*. Retrieved from www.umanitoba.ca/publications/cjeap/articles/volante.html

Appendix A

Grade 9 Mathematics Assessment
Teacher Questionnaire

Teaching Assignment

1. Please check the **grade**(s) you are teaching during the current school year.

7	8	9	10

2. Did you teach in a French Immersion Program?

☐ Yes ☐ No

3. Considering that the average total family income (before taxes) in Prince Edward Island is about $41,500, how would you describe the average socio-economic level of the community that your school served? *Select ONE response.*

☐ Far above average ☐ Above average ☐ Average

☐ Below average ☐ Far below average

4. Relative to other schools, the academic achievement of students in the school you taught was:

☐ Far above average ☐ Above average ☐ Average

☐ Below average ☐ Far below average

5. Which of the following best describes the school in which you taught?

☐ Urban ☐ Suburban

☐ Semi-rural ☐ Rural

6. Was the school in which you taught identified for any special Ministry initiatives or interventions?

☐ Yes ☐ No

If Yes, please specify: _____

_____

Instructions for answering survey items:

Other people have identified several purposes/uses for the Grade 9 Common Assessment. What we are interested in is . . . How appropriate do you feel each purpose/use is for supporting the education of your students?

You will use a five-point scale to indicate the degree to which you feel each purpose is appropriate. The scale ranges from: **1** = *not appropriate at all* to **5** = *very appropriate*.

For example, using the five-point scale, how appropriate do you feel the following purpose is for the tests and/or examinations that you are going to talk about?

1. Improve the achievement of all students.

Not approriate.....Very approriate

1 2 3 4 5

Provincial Assessment Beliefs

How appropriate do you believe the following purposes and uses of the **Grade 9 Common Assessment** are for students?

Suggested Purpose/Use:	**Not App. 1**	**2**	**3**	**4**	**Very App. 5**
1. Ensure high academic standards					
2. Measure student achievement					
3. a. Improve the achievement of all students					
b. Reduce the achievement differences among students					
4. Support curriculum implementation					
5. Focus instruction on the provincial curriculum					
6. Improve and enhance teaching					
7. Provide common measures so that teachers can link their own assessments to provincial standards					
8. Inform parents about the performance of their children					
9. Inform parents about the performance of the child's school					
10. Inform parents and the public about the performance of the school system					
11. Determine how well the students are learning the intended curriculum					
12. Evaluate the quality of:					
- students					
- teachers					
- schools					
- school districts					
provincial education programs					

Suggested Purpose/Use:	**Not App. 1**	**2**	**3**	**4**	**Very App. 5**
13. Provide data for data-based decision-making at the					
- student level					
- class level					
- school level					
- school district level					
- provincial level					
14. Identify exemplary					
- students					
- teachers					
- programs					
- schools					
- school districts					
15. Help rank students					
16. Help determine final grades					
17. Motivate students to work hard					
18. Help parents select the school students will attend next year					
19. Provide information for growth of					
- students					
- schools					
- school districts					
20. Identify students in need					
21. Identify schools in need					
22. Increase teachers' assessment knowledge and skills					

Using Provincial Assessment Results

1. As a teacher, do you actually use the results?

   ☐ Yes ☐ No

2. Do you feel that provincial testing programs have any value?

   ☐ Yes ☐ No

Issues Related to Provincial Assessment

How **serious** do **you** feel each issue is for supporting the education of your students?

	Not Iss. 1	**2**	**3**	**4**	**Very Ser. 5**
1. Provincial testing narrows teaching of the curriculum at the					
a. grade levels at which the tests are given.					
b. remaining grade levels.					
2. Provincial testing takes up too much instructional time at the grade level the test is given.					
3. Results of provincial testing are used to publicly rank schools.					
4. Results of provincial testing are used to evaluate teacher effectiveness.					
5. Provincial tests are used as an accountability tool.					
6. Results from the provincial tests are encouraged to be included in the students' final grade.					
7. Provincial test results provide a "snap-shot" of what students know and can do.					
8. Teachers teach towards the test.					

	Not Iss. **1**	**2**	**3**	**4**	**Very Ser.** **5**
9. Classroom activities are limited to the learning expectations assessed on Provincial Assessments.					
10. Classroom assessment instruments reflect item format and content on provincial tests.					
11. Content of the provincial tests is not consistent from year to year.					
12. The May administration is too early to get accurate achievement data.					
13. Marking of constructed response (students provide the answer) test items is not reliable.					
14. Multiple-choice items are over-used.					
15. Performance standards are not consistent across years (scale is adjusted each year to reflect the percentage of who should pass).					
16. Students are excluded from participating in order to improve school results.					
17. The inclusion of special needs (i.e., IEP) and ESL students on provincial tests.					
18. Due to the nature of the reported results, the data cannot be used to support instruction.					
19. Teachers do not know how to interpret assessment results.					
20. Teachers do not know how to use the test results.					
21. Principals do not know how to interpret the test results.					
22. The press ignores the limitations of the results when publishing rankings of schools based on provincial test results.					
23. The public does not know how to interpret the test results.					

Teacher Background

Lastly, we would like a little more information about you.

1. Gender:

☐ Male ☐ Female

2. Highest Level of Education.

☐ Teaching Certificate ☐ Degree (e.g., B.A., B.Sc.) plus B.Ed.

☐ Ph.D./Ed.D. ☐ Bachelor of Education

☐ M.A./M.Sc./ M.Ed ☐ Other

3. How confident are you about your knowledge of the

   a. Common Assessment program?

Not confident				Very Confident
1	2	3	4	5

   b. National (e.g., PCAP) and International Assessments (e.g., TIMMS, PISA)

Not confident				Very Confident
1	2	3	4	5