An Alternative Vision for Large-scale Assessment in Canada

Louis Volante
Brock University

Abstract

Concern over the quality of education has prompted virtually every Canadian province and territory to develop large-scale assessment programs to measure student achievement. The approach of individual provinces and territories varies according to the grades tested, sample size, test format, and frequency of administration. Many provinces also participate in national and international testing programs. This paper provides a general overview of the various large-scale assessment programs across Canada and outlines central arguments for and against student achievement testing. Research documenting the impact of large-scale testing on students and teachers is also reported. The discussion proposes an alternative vision for large-scale assessment aimed at supporting teacher's instructional practices and student learning. A set of key considerations within this vision serve as a basis for assessment policy reform.

Introduction

Large-scale assessment continues to provoke intense debate both within and outside of Canada. This is not surprising as policymakers, educators, and the public are increasingly looking to these measures to serve a variety of important purposes such as gauging student learning, signaling worthy goals for students and teachers to work toward, providing useful feedback for instruction, and ultimately holding education systems accountable (Chudowsky & Pellegrino, 2003). It is perhaps the last of these reasons—educational accountability—that has caused such a backlash against international, national, and provincial testing programs. Educators continue to resist these tests as measures of educational quality and provide a number of compelling reasons to dismiss these test results. Policymakers remain unmoved and view external testing as essential for evaluating the efficacy of learning programs and school systems (Heubert, 2003). This paper examines arguments for and against testing and describes an alternative vision for large-scale assessment geared to supporting teachers' instructional practices and student learning. The proposed vision is meant to serve as a model for assessment policy reform both within and outside Canada.

A survey of Ministry and Department of Education websites reveals that every province and territory administers some form of large-scale student assessment. The approach of individual provinces and territories varies according to the grades tested, sample size, test format, frequency of administration, and

degree of participation in national and international testing programs. In addition to provincial and territorial assessment programs, some Canadian provinces participate in the Program for International Student Assessment (PISA), Progress in International Reading and Literacy Study (PIRLS), and/or the Trends in International Mathematics and Science Study (TIMSS, formerly known as the Third International Mathematics and Science Study). These international assessments are in addition to the School Achievement Indicators Program (SAIP) that is conducted at the national level.

Large-Scale Assessment in Canada

International programs

PISA. This assessment program is conducted by the Organisation for Economic Co-operation and Development (OECD). PISA is designed to provide policy-oriented international indicators of the skills and knowledge of 15-year-old students in reading, mathematics, and science. The program involves more than 40 countries and tested approximately 265,000 students in the most recent administration. Three administrations of PISA are planned. In each cycle, two-thirds of the testing time is devoted to a major literacy domain: Reading in 2000, Mathematics in 2003, and Science in 2006. The survey includes a two-hour written test to assess students' knowledge in reading, mathematics, science and problem solving and a 30-minute questionnaire that collects background information on the student. Although the last sample in most countries was between 4,000 and 10,000 students, in Canada, more than 28,000 students were tested in more than 1,000 schools. This sample was intended to provide information from all ten provinces and to allow for estimates for both official language groups (cf., Government of Canada: <http://www.pisa.gc.ca/>).

PIRLS. This assessment program is conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA). PIRLS assesses the reading literacy of 9-year-old students with respect to two major reading purposes: literary and informational. Reading achievement results present each country with an opportunity to examine educational policies and practices against a globally-defined benchmark. PISA also provides information about children's early literacy experiences and reading instruction. Approximately 150,000 students were tested in 40 countries during the first administration in 2001. Alberta, British Columbia, Nova Scotia, Ontario, and Quebec were represented within this sample. The next administration will take place in 2006 and will measure trends in reading literacy achievement. Last, PIRLS is intended to complement the IEA's Trends in International and Mathematics Study (TIMSS), which also assesses 9-year-old students (cf., IEA: <http://www.iea.nl/pirls2001.html>).

TIMSS. This assessment program provides data on the mathematics and science achievement of nine- and thirteen-year-old students in more than 50 countries. Approximately 500,000 students have participated in this study during the various testing sessions in 1995, 1999, and 2003. From 2003, the study has placed more emphasis on questions and tasks that assess students' analytical, problem-solving, and inquiry skills and capabilities. Students, teachers, and

school principals are also asked to complete questionnaires concerning the context for learning mathematics and science. This information is meant to provide a resource for interpreting achievement results and to track changes in instructional practices. Both Ontario and Quebec were represented within the most recent administration. The next round of TIMSS data collection will take place in 2007 and will include Alberta, British Columbia, Ontario, and Quebec (cf., IEA: <http://www.iea.nl/timss2003.html>).

National Programs

SAIP. This is a program of pan-Canadian assessments of student achievement in reading, writing, mathematics, and science. The program tests a random sample of more than 35,000 13- and 16-years-old students. SAIP is conducted by the Council of Ministers of Education, Canada (CMEC). The first cycle of assessments began in 1993 with the administration of the mathematics assessment. The reading and writing assessment followed in 1994 and the science assessment was administered in 1996. The second cycle of assessments was administered from 1997 to 1999. A third cycle began with the administration of the mathematics assessment in 2001. The program is designed to complement existing assessments in each province and territory and to provide each minister of education with a basis for examining the curriculum and other aspects of the school system. Reporting occurs at the provincial and national levels only. A streamlined assessment initiative known as the Pan Canadian Assessment Program (PCAP) will replace SAIP in 2007. PCAP will assess 13-year-old students in reading, mathematics, and science (cf., CMEC: <http://www.cmec.ca/saip/>).

Provincial Programs

Alberta. In Alberta, all students are assessed in grades 3, 6, 9, and 12. Grade 3 students write mathematics and language arts tests; Grades 6 and 9 students write tests in English language arts, mathematics, science, and social studies; Grade 12 students write diploma examinations in biology, chemistry, English language arts, French, pure and applied mathematics, physics, science, and social science. Provincial examinations are worth 50% of the students' final grade. In addition to school and jurisdiction reports, Alberta Learning prepares individual profiles, showing each student's performance in relation to provincial standards (cf., Government of Alberta, Alberta Education, 2005: <http://www.education.gov.ab.ca>).

British Columbia. In British Columbia, students participate in a variety of assessment programs such as the Foundation Skills Assessment (FSA), Provincial Learning Assessment, and the Grades 10-12 Provincial Examinations. FSA is administered annually to all BC students in Grades 4 and 7. Students are tested in reading comprehension, writing, and numeracy. The Provincial Learning Assessment Program is also administered to Grades 4 and 7 students, but on a sample basis as needed. It covers the subjects or cross-curricular areas not covered by FSA. Lastly, the Provincial Exam Program certifies the performance of BC students graduating from Grade 12. Exam marks are worth

40% of the students' final grade (cf., Government of British Columbia, Ministry of Education, 2005: <http://www.bced.gov.bc.ca>).

Manitoba. In Manitoba, students are assessed at the beginning of Grade 3 and in each semester of Grade 12. Grade 3 students are tested in reading and numeracy; Grade 12 students are tested in English language arts, applied and consumer mathematics, pre-calculus, and French. Senior provincial assessments are worth 30% of the students' final grade. Test results are distributed in various types of reports to district offices and schools. The Instruction, Curriculum and Assessment Branch is currently engaged in broad consultations with the goal of developing a provincial assessment system that draws from classroom-based assessment and supports student learning in the middle years (cf., Government of Manitoba, Ministry of Education, Citizenship and Youth, 2005: <http://www.edu.gov.mb.ca>).

New Brunswick. In New Brunswick, students are assessed in both elementary and secondary schools. At the elementary level, Grade 2 and 7 students are tested in reading and writing whereas Grade 5 and 8 students are tested in mathematics. In secondary schools, Grade 9 students complete an English Language Proficiency Test, which is a requirement for graduation. Grade 11 students are also tested in mathematics and English, with these examinations counting for 30% of their final grade (cf., Government of New Brunswick, Department of Education, 2005: <http://www.gnb.ca/0000/index-e.asp>).

Newfoundland and Labrador. In Newfoundland and Labrador, students are assessed in Grades 3, 6, 9, and 12. Grade 3 and 6 students write language arts tests; Grade 9 students complete tests in mathematics, science, and language arts; Grade 12 students write mandatory public examinations in each of their courses. Senior examinations are worth 50% of students' final course grades. Reporting of student performance occurs at the provincial, district, and school levels and is managed by the Division of Evaluation and Research (cf., Government of Newfoundland and Labrador, Department of Education, 2005: <http://www.gov.nl.ca/edu>).

Nova Scotia. In Nova Scotia, students are assessed in elementary, junior high, and senior high schools. Students in Grades 5 and 8 are given a mathematics test, while students in Grades 6 and 9 are given a language arts test. In Grade 12, students complete the Nova Scotia Examinations (NSE) in chemistry, physics, mathematics, and English. NSE examinations are worth 30% of students' final course grades. The Evaluation Services Division is responsible for reporting provincial assessment results (cf., Government of Nova Scotia, Department of Education and Culture, 2005: <http://www.ednet.ns.ca/>).

Ontario. In Ontario, an arm's-length testing agency known as the Education Quality and Accountability Office (EQAO) is responsible for assessing students in Grades 3, 6, 9, and 10. Grades 3 and 6 students complete tests in reading, writing, and mathematics; Grade 9 students complete a mathematics test; Grade 10 students complete the Ontario Secondary School Literacy Test (OSSLT). OSSLT assesses students in reading and writing and is a graduation requirement. Provincial assessment results are reported at the provincial, district, and school levels. Individual profiles are also provided that show each student's performance in relation to provincial standards (cf., EQAO, 2005: <http://www.eqao.com>).

Prince Edward Island. PEI does not administer a large-scale student assessment program that is distinct from other provinces. Currently, they participate in SAIP and the Atlantic Provinces Education Foundation Indicators (APEF) program that includes New Brunswick, Newfoundland, and Nova Scotia. APEF is in the process of developing an assessment resource to assist teachers in identifying the strengths and needs of students. The resource is intended for teachers as they work with students to develop an understanding of the mathematics concepts and skills prescribed by the Atlantic curriculum Grades K-3 (cf., Government of Prince Edward Island, Department of Education, 2005: <http://www.gov.pe.ca/education/>)

Quebec. In Quebec, the Ministry of Education administers uniform examinations in French, English, physical science and history. Students are tested in Secondary IV and Secondary V (the equivalent of Grades 10 and 11, respectively). In order to graduate, students must pass the two levels of languages (language of instruction and their second language—French or English). Provincial examinations are compulsory prerequisites for certain college programs and account for 50% of the students' final grade (cf., Government of Quebec, Ministry of Education, 2005: <http://www.mels.gouv.qc.ca/>).

Saskatchewan. In Saskatchewan, the Assessment for Learning and System Information and Accountability Unit monitors students' learning from kindergarten to Grade 12. Students in Grades 5, 8, and 11 are tested in language arts, mathematics, technological literacy, and critical and creative thinking. Grade 12 students complete diploma examinations that are worth a minimum of 25% of the student's final grade. At the discretion of the teacher, these test scores can count for up to 50 percent of the student's final grade. Students and teachers also complete questionnaires that provide contextual information related to student learning. Provincial assessment results are released in a year-end document (cf., Government of Saskatchewan, Saskatchewan Learning, 2005: <http://www.sasked.gov.sk.ca/>).

Territorial Programs

Northwest Territories. The Northwest Territories (NWT) does not administer a large-scale student assessment program that is distinct from other provinces or territories. Students within this province complete the achievement tests and diploma examinations prescribed by Alberta (cf., Government of North West Territories, Department of Education, Culture, and Employment, 2005: <http://siksik.learnnet.nt.ca/>).

Nunavut. As in the NWT, students within this territory complete the Alberta diploma examinations. This new territory is also set to participate in the initial administration of PCAP in 2007 (cf., Government of Nunavut, Department of Education, 2005; <http://www.gov.nu.ca/education/eng/>).

Yukon. In Yukon, students in Grades 3, 6, and 9 complete the Yukon Achievement Tests (YAT). YAT is based on the Western Canada Protocol Common Curriculum Frameworks in language arts and mathematics. Senior students in Grades 11 and 12 are administered a language proficiency index that is typically utilized by post-secondary institutions. Students in Grade 12 are also required to write the British Columbia provincial examinations in English,

mathematics, and science (cf., Government of Yukon, Department of Education, 2005: <http://www.education.gov.yk.ca/>).

Differences Across North America

It is important to point out that Canadian teachers are responsible for the development and grading of provincial and territorial large-scale criterion-referenced assessments. The national SAIP instruments are also developed and graded by classroom teachers. These activities are coordinated with respective ministries or departments of education. This is in stark contrast to the United States, where large-scale assessments are typically norm-referenced and are run primarily by commercial organizations outside of the education system. This distinction is important and may account for the fact that Canadian testing programs tend to account for greater linkages with classroom practice than their American counterparts (Gambell & Hunter, 2004). The reliance on criterion-referenced testing also suggests that Canadian testing programs tend to be more aligned with mandated curricula. The latter is a frequent criticism of norm-referenced achievement testing and may account for the fact that many American states (i.e., Missouri, New York, and Utah) have recently switched to criterion-referenced assessment programs.

Despite the important difference between norm- and criterion-referenced assessment, both forms of testing have served high-stakes purposes across North America. Indeed, the term “high-stakes achievement tests” has been defined as including either mandated norm- or criterion-referenced tests where the aggregated results are used for summative purposes such as student promotion, graduation, and/or judging the effectiveness of schools (Burger & Krueger, 2003). Conversely, low-stakes assessments aggregate scores for provincial, national, or international trends and profiles, but not individual marks for promotion or graduation purposes (Gambell & Hunter, 2004). Using the previous descriptions of high- and low-stakes assessment, it is easy to see how virtually every province administers high-stakes tests because they comprise a significant percentage of students’ final grades (i.e., typically between 30 and 50%) and in some cases are a graduation or compulsory requirement for postsecondary attendance (i.e., New Brunswick, Ontario, and Quebec).

Nevertheless, it is important to point out that in Canada large-scale assessment results tend to have high-stakes consequences for students, not teachers or administrators. Although educators may feel undue pressure to raise test scores, particularly when results are published, they are rarely (if ever) officially sanctioned for poor student performance. In the United States, however, test results tend to have significant consequences for both groups. For example, schools that fail to meet Adequate Yearly Progress (AYP), as reflected in mandated improvements in test scores, are labeled as “failing” and are inevitably taken over by the state if scores do not improve. Simner (2000) notes that an unintended consequence of this increased pressure are cases of improper teacher and administrator behaviour in states such as New York, Texas, Massachusetts, Maryland, Ohio and Connecticut. Examples of improper behaviour included encouraging low achieving students to stay home the day of the test, sending low achieving students away on school trips, and even providing students with the correct answers during the testing session.

Large-Scale Assessment Debate

The following section identifies sets of arguments for and against large-scale assessment. These lists are not meant to be exhaustive. Rather, they represent some of the most frequently cited problems and benefits to utilizing large-scale measures in the academic literature.

Critics of large-scale assessment

Key arguments against large-scale assessment are that testing:

- supports policy-decisions that have not been adequately scrutinized and are based on measures with important psychometric limitations (Kane, 2002; Popham, 2001),
- acts as a form of educational and social control that subverts the professional autonomy of educators (Canadian Teachers Federation, 1999),
- holds teachers responsible for results with inequitable resources and contributes to feelings of lower self-efficacy (American Educational Research Association, 2000; Darling-Hammond, 2004),
- widens the gap between minorities and white students by ignoring key factors known to affect student performance such as socio-economic status, language of origin, and the students' physical and/or emotional health (McNeil, 2000; Valencia & Villarreal, 2003),
- undermines the quality of education by penalizing divergent thinking, creativity, and intellectual work in general (Anderson & Postl, 2001; Glaser, Linn, & Bohrnstedt, 1997; Hess & Brigham, 2000; Wideen, O'Shea, Pye, & Ivany, 1997),
- narrows and distorts the curriculum by encouraging "teaching to the test" techniques which take valuable time away from non-tested subjects, particularly when high-stakes are attached to results (Kohn, 2002; Smith & Fey, 2000; Volante, 2004),
- provides information that is of relatively little use to improving classroom practice (Stiggins, 2002; Falk, 1998), and
- damages children's self-concept and leads to student disengagement (Meaghan & Casas, 2001; Miller, & Tovey, 1996).

The last argument is perhaps the most damaging as schools are expected to promote the physical, cognitive, and social-emotional health of their students.

Advocates of large-scale assessment

Advocates of large-scale assessment argue that critics have overstated their case and failed to provide empirical data to support many of their assertions (Covaleskie, 2002). They point to research that has documented positive outcomes for students, teachers, and schools. For example, research has shown that testing can:

- make Canadian students work harder (Anderson, 1990) and perform better on international assessments such as the TIMSS (Bishop, 2001),

- identify the most successful teaching practices and proficient teachers (Cizek, 2001; Sanders & Horn, 1998),
- enhance teachers' reflective and critical thinking when planning instruction (Goldberg & Roswell, 2000),
- lead to positive increases in educators' knowledge about testing and of testing issues (Earl & Torrance, 2000; Runte, 1998),
- improve teachers' assessment and instructional practices, particularly when they are involved in marking these assessments (Gambell & Hunter, 2004; Green, 1998),
- stimulate action research that serves as a basis for school improvement (Wideman, 2002),
- provide schools with valuable information about the consequences of their past practices and program effectiveness (Anderson & Postl, 2001; Taylor & Tubianosa, 2001).

Collectively, these points underscore the central argument that large-scale assessment programs provide valuable information for provinces and territories to judge their standards and the quality of their programs (McEwen, 1999).

The main limitation with much of the previous literature is that little attention has been paid to the different environments in which large-scale assessment programs operate. For example, the long-standing criticism that testing promotes 'teaching to the test' techniques may appear to be a more pressing concern and egregious practice in jurisdictions that rely exclusively on norm-referenced measures. Levinson (2000) argues that although this phrase "teaching to the test" is used disparagingly in the United States, other countries (i.e., Australia, Canada, Czech Republic, England, Japan) view testing as a way to promote the curriculum, as the same agency that develops the test also develops the curriculum. Although this statement is not uniformly supported in the international literature, it does underscore the importance of considering test design and other contextually related issues before making broad statements on the advantages and disadvantages of large-scale testing.

Alternative Vision

The literature is replete with instances of large-scale assessment programs being imposed by policymakers with significant repercussions attached to poor performance. Not surprisingly, this imposition often alienates a significant portion of students and teachers and, generally speaking, has a harmful effect of the education system (AERA, 2000; Darling-Hammond, 2004). It seems clear that no single test, given at one point in time, should override a teacher's professional judgment—particularly for high-stakes purposes like student promotion or graduation. Nevertheless, educators cannot simply ignore the potential of large-scale assessments for improving teacher practice and student learning. This position is untenable and will only marginalize educators to the point of irrelevance (Covaleskie, 2002). The challenge, then, is how to utilize large-scale assessments in a manner that can be embraced by a range of primary stakeholders. In order to accomplish this lofty feat, large-scale assessment programs need to be developed more effectively, utilized for educational improvement, reported in a manner that encourages professional collegiality, and supported by appropriate professional development. These four considerations

serve as the cornerstone for an alternative vision that is meant to balance some of the concerns of educators and policymakers. The following offer further explanation of these four factors.

Technical quality

If one accepts the premise that tests drive curriculum and instruction, perhaps the easiest way to reform instruction and improve educational quality would be to construct better tests (Yeh, 2001). Unfortunately, many large-scale assessment programs have been highly susceptible to inappropriate test preparation strategies, like “teaching to the test” techniques, that narrow the curriculum and focus instruction on simulated test activities. These preparation strategies do little to promote authentic learning and often provide artificially inflated scores (Messick, 1995; Smith & Fey, 2000). Testing can be improved by changing test items and tasks for each subsequent administration and by utilizing test items that are designed to push students beyond the recall of facts and algorithms to higher order thinking and problem solving. When such a strategy is utilized, it is good teaching of the entire curriculum that serves as the best test preparation strategy (Earl, 1999). Essentially, “teaching to the test” becomes a desirable objective as the measure itself encourages preparation strategies that focus on critical thinking and broad curriculum coverage.

Large-scale assessment programs also need to explore innovative ways to assess skills that do not readily lend themselves to traditional paper-and-pencil tests. Currently, provincial and territorial assessment programs are unable to measure a host of skills that are essential for future success such as speaking clearly, managing time effectively, or designing and carrying out a project. These performance-based tasks can also serve policy purposes of assessment if there is some standardization (e.g., tasks are centrally determined to some degree with consistent rules for scoring) and flexibility for schools, teachers, and students (e.g., some choice around which tasks from a broader set and when students respond to them) (Chudowsky & Pellegrino, 2003). Although there are likely to be challenges to developing and utilizing performance-based measures in large-scale contexts, these types of authentic assessments provide the basis for a more detailed analysis of the students’ knowledge and skills. Fortunately, research is emerging in the United States, England, and Australia where traditional and performance-based forms of assessment have been successfully integrated for accountability purposes (Wilson, 2004).

Assessment for improvement

The ultimate value of large-scale assessment programs depends on their ability to foster, and not impede student and teacher learning. To do this, large-scale assessments must be seen as important, but not as measures for high-stakes decisions. Ontario serves as a compelling example of the distinction between high- and low-stakes testing. Although research has been able to demonstrate positive outcomes in the low-stakes elementary assessments, such as improved planning and classroom practice (Earl & Torrance, 2000; Wideman 2002), no corresponding literature can be identified for the high-stakes OSSLT. All indications suggest that utilizing the OSSLT as a graduation requirement has

negatively impacted the high school completion rate for secondary students. Although the completion rate was steady in the mid-90s to 2001 at 78%, there was a sharp drop in 2001 to 71 percent, which has remained relatively unchanged (People for Education, 2006). Alberta has seen a similar trend, with only 43% of their high school students entering postsecondary institutions. This is the lowest percentage amongst Canadian provinces and is an unintended result of their accountability system's continuing over-emphasis on high test scores (Alberta Teachers Association, 2005). Provincial and territorial testing programs need to be wary of pushing their students out of school rather than pulling them into contexts of improved teaching and learning.

This push-pull distinction is a direct result of utilizing large-scale assessments for decisions that far outstretch their power on matters such as promotion, graduation, or utilizing single test scores for a significant portion of a student's final grade (i.e., more than 30%). Given the brief nature of their administration, these large-scale assessment measures are often unable to faithfully sample the entire breadth and depth of the curriculum (Messick, 1995; Popham, 2003). If student scores are required in particular subject areas, broader coverage can be achieved by extracting evidence from classroom assessments (often referred to as curriculum-embedded assessment). This can be used to supplement the information collected from on-demand large-scale assessments (Chudowsky & Pellegrino, 2003). This mixing of large- and classroom-based assessment does not render testing data meaningless. Rather, it positions the large-scale assessment results to support (or challenge) the inferences that follow from teachers' daily classroom assessments. Another advantage of this utilization approach is that it leads to improved student retention rates. For example, American states that rely on multiple measures (i.e., portfolios, performance assessment, grades in courses tied to state standards, and student exhibitions of learning) tend to maintain higher and steadier rates of graduation as opposed to test-only graduation systems (Darling-Hammond, Rustique-Forrester, & Pecheone, 2005). Thus, it seems imperative that Canadian provinces also consider the benefits of a multiple measures approach—particularly in jurisdictions where testing has profound consequences for students.

Reporting results

Even when a test is designed and utilized effectively, the results must be reported in a manner that promotes professional collegiality. The education community must model effective dissemination of assessment information, whether international, national, or provincial (McEwen, 1999). Results must be scrutinized with attention given to communication strategies that report the delimitations and limitations of the test (Burger & Krueger, 2003). For example, Alberta researchers have shown that social class variables accounted for as much as 45% of student achievement results (Lytton & Pyryt, 1998). Similarly, Willms (2004) found that when factors such as measurement and sampling error, family background, and school contexts were considered, Alberta's ranking on the PISA dropped from first to fifth place in Canada. Despite these important factors, schools, districts, and provinces continue to be ranked causing misinformation and unhealthy competition (Simner, 2000). This competition ultimately pits

teacher against teacher and school against school so that a community of learners is discouraged.

Test results must also be reported to teachers in a way that facilitates their ability to refine classroom practice. EQAO provides schools with individual profiles that clearly explain students' assessment results in relation to provincial standards. These profiles also provide a strategy for teachers to use exemplars to talk to parents and their child about how the assessment information fits with the provincial curriculum expectations as well as other information about the child. Schools and districts are required to prepare reports based on material provided by EQAO along with other material related to important factors such as demographics. These school and district reports include interpretations of the results and an action plan for improving student learning. Collectively, this process seems better able to improve the utilization of large-scale assessment data than simply providing numerical scores on discrete categories. Other educational jurisdictions would be wise to capitalize on some of the reporting advances made by provinces like Ontario.

Professional development

A major reason for the advent of externally controlled, large-scale achievement tests was the belief that teachers could not be trusted to make sound educational decisions about what students know and are able to do (Popham, 2004; Stiggins, 2002). Essentially, these tests were implemented to reduce the subjectivity that is inherent in teacher's classroom assessment strategies. This is somewhat of a moot argument given that large-scale assessments also have numerous validity concerns associated with their content (Messick, 1996). Nevertheless, this accusation does underscore the sad point that most teachers have not been adequately trained in assessment and substantial and ongoing professional development to create reliable and valid tasks (Jones, 2004). This training is an essential step if teachers' assessment data is to stand up to public scrutiny.

The success of educational reforms is inextricably connected to the preparation and professional development of educators (Fullan, 2001). Policymakers and district personnel must provide the time and resources for this important training. This would appear to be a prudent investment, especially when one considers the massive amounts of money spent on large-scale achievement testing every year. Preservice and inservice training would enhance the utilization of large-scale assessment data by classroom teachers, something that has been woefully absent in many large-scale testing contexts (Popham, 2003). Training would also facilitate better teacher practice, an activity that should ultimately lead to improved student outcomes. Emerging research suggests that when large jurisdictions have made significant investments in professional development, such as enhancing teachers' assessment literacy, they have demonstrated improvements in teacher efficacy levels and student performance (Volante & Melahn, 2005). Canadian provinces and territories should seek to emulate these rare instances of large-scale investment in teacher development.

Implications for Educational Policy and Practice

The word assessment comes from the Latin “assidere,” which means “to sit by one’s side” (Wiggins, 1993). To date, most large-scale assessment programs have run ahead of teachers, working to trivialize the assessment work they do on a day-to-day basis. Indeed, current accountability systems appear to hold that on-demand standardized testing is more credible than the diverse range of teacher assessments applied throughout the school year. Education leaders need to rethink some of these assumptions and resist the temptation to over-rely on large-scale assessment results for accountability purposes. Large-scale assessment data is part of an accountability system; it is not the entire system itself (Darling-Hammond, 2004). As Earl (1999) reminds us, policymakers need to promote such measures primarily as vehicles for developing action plans to improve instruction, staff development, school organization, resource and community engagement—as ways of improving students’ lives and ensuring that all children receive a high quality education. Failure to do so will invariably lead to continued disillusionment and outright hostility with the entire testing process.

Nevertheless, developing an improvement orientation for large-scale assessment programs is premised on a number of contentious, and at times, competing considerations. For example, in order to promote improvements in teaching and learning, more sophisticated measures and reporting techniques are required. The latter substantially increases the time and costs associated with test development and administration. Similarly, while testing each and every child provides more detailed and useful feedback—it also permits closer scrutiny that is often used to impose penalties on individual students, teachers, or schools. Although these issues are not easily reconciled, the impact of using sub-standard measures for important high-stakes decisions are even more problematic. Certainly, the demands of the new knowledge economy suggest the necessary monetary investments and utilization safeguards are well worth the trouble.

Conclusion

The knowledge base for assessment is constantly changing as new methods and procedures are being implemented around the world. Assessment-led reforms must follow suit and adapt and change in response to sound empirical research—not what is politically expedient. The previous considerations serve as a foundation for assessment policy reform. Collectively, they provide direction for provincial and territorial programs that position Canada on a path toward excellence and ensure large-scale measures support improved teaching practices and student learning. This should be the ultimate litmus test for any assessment-led reform.

References

- AERA (2000). Cautions issued about high-stakes tests. Available [online]: http://www.education-world.com/a_issues/issues110.shtml.
- Alberta Teachers Association (2005). *Accountability in Education*. Available [online]: <http://www.teachers.ab.ca>.
- Anderson, J. O. (1990). *The impact of provincial examinations on education in British Columbia: General report*. Victoria, BC: British Columbia Department of Education.
- Anderson, B., & Postl, B. (April 2001). *Using large data sets as a basis for school improvement*. Paper presented at the Annual Meeting of the American Educational Research Association,

Seattle, WA.

- Bishop, J. H. (2001). A steeper, better road to graduation. *Education Next*, 1(4), 56-61.
- Burger, J. M., & Krueger, M. (2003). A balanced approach to high-stakes achievement testing: An analysis of the literature with policy implications. *International Electronic Journal for Leadership in Learning*, 7(4). Available [online]: <http://www.ucalgary.ca/~iejll>.
- Canadian Teachers Federation. (1999). *Province-wide assessment programs*. Available [online]: <http://www.ctf-fce.ca/e/what/other/assessment/testing-main.htm>.
- Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessment that supports learning: What will it take? *Theory into Practice*, 42(1), 75-83.
- Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20(4), 19-27.
- Covaleskie, J. F. (2002). Two cheers for standardized testing. *International Electronic Journal for Leadership in Learning*, 6(2). Available [online]: <http://www.ucalgary.ca/~iejll/volume6/covaleskie.html>.
- Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record*, 106(6), 1047-85.
- Darling-Hammond, L., Rustique-Forrester, E., & Pechione, R. L. (2005). *Multiple Measures Approaches to High School Graduation: A Review of State Student Assessment Policies*. Stanford, CA: School Redesign Network. Available [online]: http://www.schoolredesign.net/srn/mm/pdf/multiple_measures.pdf.
- Earl, L. (1999). Assessment and accountability in education: Improvement or surveillance? *Education Canada*, 39(3), 4-6.
- Earl, L., & Torrance, N. (2000). Embedding accountability and improvement into large-scale assessment: What difference does it make? *Peabody Journal of Education*, 75(4), 114-41.
- Falk, B. (1998). Testing the way children learn: Principles for valid literacy assessment. *Language Arts*, 76(1), 57-66.
- Fullan, M. (2001). *The new meaning of educational change* (3rd ed.). New York: Teachers College Press.
- Gambell, T., & Hunter, D. (2004). Teacher scoring of large-scale assessment: Professional development or debilitation? *Journal of Curriculum Studies*, 36(6), 697-724.
- Glaser, R., Linn, R., & Bohrnstedt, G. (1997). *Assessment in transition: Monitoring the nation's educational progress*. Stanford, CA: National Academy of Education (1997).
- Goldberg, G. L., & Roswell, B. S. (2000). From perception to practice: The impact of teachers' scoring experience on performance-based instruction and classroom assessment. *Educational Assessment*, 6(4), 257-90.
- Green, J. M. (1998). Authentic assessment: Constructing the way forward for all students. *Education Canada*, 38(3), 8-12.
- Heubert, J. P. (2003). First, do no harm. *Educational Leadership*, 60(4), 26-30.
- Hess, F. M., & Brigham, F. (2000). None of the above: The promise and peril of high-stakes testing. *American School Board Journal*, 187(1), 26-29.
- Jones, K. (2004). A balanced school accountability model: An alternative to high-stakes testing. *Phi Delta Kappan*, 85(8), 584-90.
- Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), 31-41.
- Kohn, A. (2000). *The case against standardized testing: Raising scores, ruining the schools*. Portsmouth, NH: Heineman.
- Levinson, C. Y. (2000). Student assessment in eight counties. *Educational Leadership*, 57(5), 58-61.
- Lytton, H., & Pyryt, M. C. (1998). Predictors of achievement in basic skills: A Canadian effective school study. *Canadian Journal of Education*, 23(3), 281-301.
- McEwen, N. (1999). Using international test results to improve student performance. *Education Canada*, 39(3), 44-46.
- McNeil, L. (2000). Creating new inequalities: Contradictions of reform. *Phi Delta Kappan*, 81(10), 729-34.
- Meaghan, D. E., & Casas, F. R. (1995). On standardized achievement testing: Response to Freedman and Wilson and a last word. *Interchange*, 26(1), 81-96.
- Messick, S. (1996). *Validity and washback in language testing* (Report No. ETS-RR-96-17). Princeton, NJ: Educational Testing Service.

- Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 10(1), 5-8.
- Miller, E., & Tovey, R. (Eds.). (1996). *Motivation, achievement, and testing*. Boston, MA: Harvard Education Press.
- People for Education (2006). *The Annual Report on Ontario's Public Schools 2006*. Toronto, ON: Author.
- Popham, W. J. (2004). All about accountability / Why assessment illiteracy is professional suicide. *Educational Leadership*, 62(1), 82-83.
- Popham, W. J. (2003). The seductive allure of data. *Educational Leadership*, 60(5), 48-51.
- Popham, W. J. (2001). Teaching to the test. *Educational Leadership*, 58(6), 16-20.
- Popham, W. J. (2001). *The Truth about Testing*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Runte, R. (1998). The impact of centralized examinations on teacher professionalism. *Canadian Journal of Education*, 23(2), 166-81.
- Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, 12(3), 247-255.
- Simner, M. L. (2000). A joint position statement by the Canadian Psychological Association and the Canadian Association of School Psychologists on the Canadian press coverage of the province-wide achievement test results. Available [online]: http://www.cpa.ca/documents/joint_position.html.
- Smith, M. L., & Fey, P. (2000). Validity and accountability of high-stakes testing. *Journal of Teacher Education*, 51(5), 334-44.
- Stiggins, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, 83(10), 758-65.
- Taylor, A. R., & Tubianosa, T. (2001). *Student assessment in Canada: Improving the learning environment through effective evaluation*. Kelowna, BC: Society for the Advancement of Excellence in Education.
- Valencia, R. R., & Villarreal, B. J. (2003). Improving students' reading performance via standards-based reform: A critique. *The Reading Teacher*, 56(7), 612-21.
- Volante, L. (2004). Teaching to the test: What every educator and policy-maker should know. *Canadian Journal of Educational Administration and Policy*, 35. Available [online]: <http://www.umanitoba.ca/publications/cjeap/articles/volante.html>.
- Volante, L., & Melahn, C. (2005). Promoting assessment literacy in teachers: Lessons from the Hawaii School Assessment Liaison Program. *Pacific Educational Research Journal*, 13, 19-34.
- Websites: <http://www.bced.gov.bc.ca/>; <http://www.cmec.ca/saip/>; <http://www.ednet.ns.ca/>; <http://www.education.gov.ab.ca/>; <http://www.education.gov.yk.ca/>; <http://www.edu.gov.mb.ca/>; <http://www.eqao.com/>; <http://www.gnb.ca/0000/index-e.asp>; <http://www.gov.nl.ca/edu/>; <http://www.gov.nu.ca/education/eng/>; <http://www.gov.pe.ca/education/>; <http://www.iea.nl/pirls2001.html>; <http://www.iea.nl/timss2003.html>; <http://www.mels.gouv.qc.ca/>; <http://www.pisa.gc.ca/>; <http://www.sasked.gov.sk.ca/>; <http://siksik.learnnet.nt.ca/>.
- Wideen, M. R., O'Shea, T., Pye, I., & Ivany, G. (1997). High-stakes testing and the teaching of science. *Canadian Journal of Education*, 22(4), 428-44.
- Wideman, R. (2002). Using action research and provincial test results to improve student learning. *International Electronic Journal for Leadership in Learning*, 6(20). Available [online]: <http://www.ucalgary.ca/~iejll/volume6/wideman.html>.
- Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco, CA: Jossey-Bass.
- Willms, D. J. (2004). *Variation in literacy skills among Canadian provinces: Findings from the OECD PISA*. Statistics Canada. Available [online]: <http://www.statcan.ca>.
- Wilson, M. (Ed.). (2004). *Towards coherence between classroom assessment and accountability: 103rd yearbook of the National Society for the Study of Education, Part II*. Chicago, IL: University of Chicago Press.
- Yeh, S. S. (2001). Tests worth teaching to: Constructing state-mandated tests that emphasize critical thinking. *Educational Researcher*, 30(9), 12-17.