Assessment and Accountability in Education: Improvement or

The 1990s have been a decade of economic, social and political uncertainty during which virtually every public institution has been criticized for failing to live up to expectations. In education, these criticisms have ranged from questions of educational purpose, practices and reform to questions of accessibility and justification of expenditures.

In uncertain times, education inevitably comes under scrutiny. When people feel anxious and concerned about the future for themselves and their children, they look to schools and teachers for reassurance, and they worry about whether schools are fulfilling their responsibilities. Educators are under pressure to show the public that what they are doing is working, and governments everywhere have seized on education as a cornerstone for their political agendas. Government response has been fairly consistent across countries: more centralized curricula and formal student testing.

A number of dilemmas are embedded in this blend of accountability and large-scale assessment — dilemmas that are often unexamined and sometimes unacknowledged.

A Brief History of Large-Scale Assessment in Education

External tests and examinations have always existed in schools, with a clear and singular purpose:making decisions about the educational status of individual students. They have been seen as a fair way to identify the best candidates for scarce resources, and they have been the vehicle for directing students into various programs or into the world of work. Most Canadian provinces have had provincial examinations over the years. In other countries, centralized assessment programs have meant examinations at key junctures or decision points for students, like the General Certificate of Secondary Education (G.C.S.E.) exams and A-levels in England, the Baccalaureate in France, and the public examinations in Hong Kong.

Beginning in the 1970s most states in the U.S. and a few districts in Canada expanded their standardized testing programs in elementary and secondary schools to help teachers make decisions about individual children, particularly where teacher judgement might be questioned, like entrance to special education programs or academic scholarships. However, although provincial or national assessments get the "sound bites" in the media, daily classroom assessments have always had the greatest influence in Canadian schools. Teachers have designed their own techniques for assessing student achievement based on provincial curriculum, made judgements about the quality of student work, determined student placement, promotion and program, and explained their decisions to parents.

A Change Of Purpose

Recently, large-scale assessment has become the vehicle of choice for accountability around the world, and testing has changed from an instrument for decision-making about students to a lever for holding schools accountable.[1] In most Canadian provinces, only three grades (usually 3,6,9 or 4,7,10) are tested, and in a number of provinces or territories the testing is done with a provincial sample of students (e.g., Saskatchewan, British Columbia). Several provinces have graduation examinations, as well. Some, like Alberta, use multiple choice items exclusively; others, like Ontario, Manitoba, New Brunswick, and Quebec, include a mixture of multiple choice, open-ended items, and tasks that require the student to produce a response, often to *real life* problems that require higher-order thinking and problem-solving.

Most provinces also participate in the School Achievement Indicators Program, a cyclical national assessment program in language, mathematics and science, operated under the auspices of the

By Lorna M. Earl, Ph.D.

Surveillance?

Council of Ministers of Education, Canada. Several provinces have also opted to participate in international assessment programs like the Third International Mathematics and Science Study, where they are included separately, as if they were countries.

So, what is educational accountability all about in 1999? Why has large-scale assessment of student achievement come to dominate educational reform policy? A close look at the field of accountability and assessment reveals a tangle of anxiety, enthusiasm, politics, values, truths, half-truths and misconceptions.

Is Education in Crisis?

After many years in the comfort of general public trust, education has come under scrutiny, and everyone has an opinion about how to reform it. Think tanks, royal commissions, business forums and government reviews have decried the state of public education and prophesied grave futures unless dramatic change is undertaken. Schools are viewed as failing to produce the kinds of learning that students will need for the world that awaits them. The rhetoric describing this perceived crisis portrays a sense of urgency. The political realm has been swift to answer with a plethora of education reforms.

An alternate view holds that public unhappiness with education has been "manufactured" by politicians, the media, and the business roundtable to strengthen their political agendas. Berliner and Biddle are blunt about the results of their investigations, going so far as to claim that: "Organized malevolence might actually be underway… Claims attacking the conduct and achievements of America's public schools are contradicted by evidence".[2]

As is often the case, the reality probably lies somewhere in between. Nevertheless, public and political "eyes" are on education and on the findings of large-scale assessments.

Improvement or Surveillance?

In theory, accountability sounds wonderful. In practice, it raises a host of thorny issues, not the least of which is philosophical. What does accountability mean? No blueprint defines accountability, and a number of very different understandings prevail.

For some, schools are like businesses, with accountability reflected in the bottom line. How good is the "product"? Which schools are best? Education is like a horse race with winners and losers. Accountability rewards the winners and exposes the losers. Assessments sort the schools (or students) into categories just as quarterly profit reports describe the financial status of one company in relation to others.

For others, schools are accountable for the learning and progress of *all* students, and assessment is part of the learning process — a tool to provide the detailed information educators and policy makers need to make good decisions and to identify areas for future actions. The relative position of schools is irrelevant. Instead, the focus is on continuous improvement in all schools, and the nature of the assessments is as important as the results, because they are a starting point for discussions about how to enhance learning.

Blaming or Capacity Building?

There is general agreement that large-scale assessment should have an impact on schools and on changing education. There are, however, two quite different views about how these changes might occur. Linda Darling Hammond describes them this way:

Policymakers often try to appeal to both camps by embracing common standards and individual variation, numerical comparability and descriptive sensitivity, assessment to improve student learning and to placate demands for system-wide accountability.

We live in a culture that has come to value and depend on statistical information to inform our decisions. At the same time, we are likely to misunderstand and misuse those statistics because we are "statistically illiterate" and consequently have no idea what the numbers mean.[5]

ENBREF

Les gouvernements ont réagi à la présumée crise en éducation en introduisant des programmes d'études plus centralisés et des systèmes d'examens plus structurés afin de rendre les écoles plus imputables. Or, pour que ces examens puissent s'avérer des moyens efficaces d'améliorer les résultats des élèves, ils doivent certes être considérés comme importants, mais non comme des instruments de jugement public. En tant que moyens permettant d'élaborer des plans d'action et d'améliorer la structure scolaire, l'instruction, le perfectionnement du personnel et l'engagement de la collectivité, ils peuvent ajouter de la valeur à la vie des élèves et garantir que chacun d'eux reçoive une éducation de qualité.

One view seeks to induce change through extrinsic rewards and sanctions for both schools and students, on the assumption that the fundamental problem is a lack of will to change on the part of educators. The other view seeks to induce change by building knowledge among school practitioners and parents about alternative methods and by stimulating organisational rethinking through opportunities to work together on the design of teaching and schooling and to experiment with new approaches. This view assumes that the fundamental problem is a lack of knowledge about the possibilities for teaching and learning, combined with lack of organisational capacity for change.[3]

This dichotomy is evident in many reform agendas and the large-scale assessment that goes with them. According to one view, teachers have both the capacity and the ability to act differently, but are unfocused, lazy and recalcitrant. Those holding this view use testing as the impetus for change, to "name and blame" offending teachers or to reward successful ones. According to the other view, educational change is an internal process requiring time, learning and reflection. Its proponents advocate for creating opportunities for teachers to rethink their assessment and teaching practices and learn new ones. Policymakers often try to appeal to both camps by embracing common standards *and* individual variation, numerical comparability *and* descriptive sensitivity, assessment to improve student learning *and* to placate demands for system-wide accountability.[4]

Statistical Illiteracy

Using assessment results and other indicators of quality has moved education into the world of statistics, resulting in misuse and misinterpretation because of the deceptive simplicity of numbers.

Statistics and assessments do not have a life of their own. They are tools, designed to provide consistent measurements. But unlike a metre stick, they measure things that are invisible and not easily checked. Tests and statistical procedures have been developed to provide *estimates* of invisible human qualities like learning and achievement, and there are extremely important conventions and rules for the measurement of student achievement, especially when the results are being used to make significant decisions. Too often, the symbolic representations of quality have been accepted as objective and unassailable descriptors of student achievement or of school or school system quality. Statistics and test scores may give the illusion of accuracy and objectivity, but the numbers are only as good as the way in which the test was developed and the results interpreted. Most people fail to recognize that some margin of uncertainty in the scores is an inevitable part of measuring any human characteristic that we can't see directly.

So, tests provide an estimate of student achievement, but never give a perfect measurement. When that uncertainty is taken into account, many — sometimes most— differences in raw scores between schools or districts disappear. With small schools, the uncertainty can be very large. And yet, raw score differences continue to be treated as if they were real and used to form opinions and make decisions about schools, even to reward or punish them. Educators have a responsibility to become statistically literate and to use statistics appropriately, so that their interpretations of assessment results are not misguided or misleading.

Consequences: Anticipated and Unanticipated

Experience has shown both expected and unexpected, both positive and negative consequences of large-scale assessment. While the misinterpretation and misuse of test results is sometimes due to a limited understanding of statistical concepts, it is sometimes due to the "high stakes" attached to them. When they are very important to individuals and institutions, or when they are associated with rewards or sanctions, test results are very susceptible to manipulation. This is less a testing issue than a political or moral issue. Any test can be corrupted.

When teachers are held responsible for their students' scores, test scores may

...continued on page 47

go up, but often learning doesn't change.[6] Fewer curricular activities are undertaken while instructional time is spent preparing for the test. Teaching methods become more test-like, often at the expense of good instructional practice (e.g., multiple choice identification of misspelled words rather than spelling correctly from dictation or in composition). Other areas of the curriculum are neglected and instruction is focused on memorization at the expense of thinking.[7] Some studies have even found cheating.

Because of these unanticipated consequences, researchers, educators and measurement specialists have worked hard to align the testing process more closely with classroom activities. Many jurisdictions have adopted assessment strategies that make "teaching to the test" desirable.For example, when the test items and tasks change for each administration and are designed to push students beyond recall of facts and algorithms to higher order thinking and problem solving, the best strategy for preparing the students is good teaching of all of the curriculum.

Educational assessment is an emotional issue. In some jurisdictions (largely in the United States), individuals and groups have challenged test fairness, test validity, test use, standards, or the accountability of the testing organizations in the courts. Competency tests have been challenged when opportunities to learn have been denied or there is evidence of bias. Some cases have addressed the use of assessment results for purposes that were not intended when the assessment was developed; others have arisen as a challenge to the process and the result of setting a cut-score or a standard of performance.[8]

Accountability and assessment are also very political. The national assessment in England originally combined teacher-assigned and externally-set assessment tasks, reported as student profiles.Its central purpose was to strengthen pupils' learning and teachers' professional role, while at the same time satisfying legitimate demands for public accountability.[9] Over the years, it has become the measurement of a standard product for consumers; its purpose has shifted from influencing school practice to providing a currency for accountability. In Ontario, there is evidence of major changes in pedagogy, assessment and curricular focus as a result of the Grade 3 assessment.[10]

Surveying the Landscape

Clearly, there are no simple answers in the realm of accountability and assessment. There are, however, some fairly clear issues that warrant attention– socio-political issues and technical issues.

Socio-Political Issues

What is the purpose of large-scale assessment? Is it a gate on the road, providing access or privilege for some and punishment or blame for others, or is it a road map for planning the future? This is a real choice. It is possible to serve both purposes, but very tricky because of an interesting paradox in the world of assessment. When the stakes are not high, large-scale assessments, in conjunction with other measures, can be reasonably accurate indicators of learning and provide clues about avenues to improve learning. When the stakes are high, however, assessment results become less accurate and sometimes downright invalid. Why? Because when the scores matter in life-changing ways, people will invariably move their focus away from enhancing learning towards increasing scores.So, a precondition of using assessments for improvement is that they be seen as important, but not as instruments of public judgement. Rather, they should be vehicles for developing action plans to improve school organization, instruction, staff development, resources and community engagement — as ways of adding value to students' lives and ensuring that all students receive a high quality education.

Technical Issues

If educators or politicians are going to rely on sophisticated measurement techniques and use them to make serious decisions about people or policy, it is their business to understand what the numbers mean and to ensure that they are being used and interpreted appropriately. There is a whole industry to analyze, interpret, and monitor the financial world around us. Why rely exclusively on newspaper reporters and politicians to assess schools? The technical requirements in educational measurement need to be monitored, challenged, and developed. If we ignore them, we run the risk of making important decisions using the Mad Hatter's logic, and justifying them by pointing at the Emperor's new clothes.

1 W. Firestone, D. Mayrowetzand J. Fairman, "Performance-based assessment and instructional change: the effects of testing in Maine and Maryland," *Educational Evaluation and Policy Analysis, (20), no. 2 (Summer 1998): 95-113.*

2 D. Berliner and B. Biddle The Manufactured Crisis: Myths, Fraud and the Attack on American Public Schools, (Reading, Mass. Addison Wesley, 1998).

3 L. Darling Hammond, "Performance-based assessment and educational equity," *Harvard Educational Review,* 64 (1994):23.

4 A. Hargreaves, L. Earl, and M. Schmidt, *Four perspectives on classroom assessment,* forthcoming.

5 L. Earl, "Assessment and accountability in Ontario, *"Canadian Journal of Education,* 20 (1)(1995): 45-55.

6 L.Shepard, "Psychometricians' beliefs about laerning," *Educational Researcher* (October, 1991):2-16.

7 L. Darling-Hammond, J. Ancess and B. Falk, *Authentic assessment in action: Studies of schools and Students at work* (New York: Teachers College Press, 1994).

8 W. Mehrens and J. Popham. "How to evaluate the legal defensibility of high stakes tests," *Applied Measurement in Education* 5 no.3 (1992): 265-283.

9 P. Black, "Performance assessment and accountability: The experience in England and Wales," *Educational Evaluation and Policy Analysis* 16 no.2 (1994): 191-204.

10 L. Earl and N. Torrance, *Impact of EQAO Assessments On School and Classroom Practices,* forthcoming.

Dr. Lorna Earl is an Associate Professor in the Theory and Policy Studies Department and Associate Director of the International Centre for Educational Change at the Ontario Institute for Studies in Education/University of Toronto. Her primary interest is the wise application of research, assessment and evaluation knowledge to the realities of schools and classrooms. In 1994, she was named a "Distinguished Educator" by O.I.S.E. in recognition of contributions which served to stimulate and enrich education in Ontario.