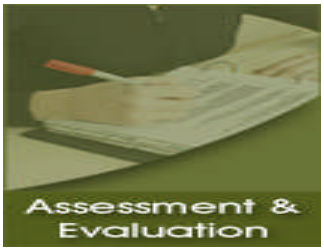
Assessment and Evaluation Branch

Information Bulletin

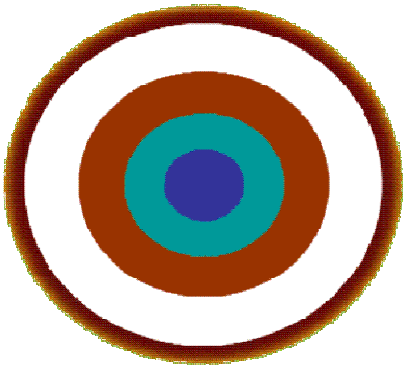


Assessment *of* Learning **September 2008**

The past two Information Bulletins focused on some of the research and methodology behind assessment; in particular, formative assessment practices or assessment *for* learning. Tackling the issue of large-scale assessment is a challenging endeavour as all stakeholders in education have their particular viewpoints. James Popham, a leading voice in the education community, observes that, "Today, more than ever, education assessment plays a pivotal role in the education of students. That's why educators— and everyone else who has an interest in education—needs a dose of assessment literacy" (*Educational Leadership*, March 2006).This bulletin attempts to address some of the issues around large-scale assessments as well as help to promote the 'assessment literacy' of educators in our province. Before we dive into the difficult questions, let's begin with some definitions:

Validity

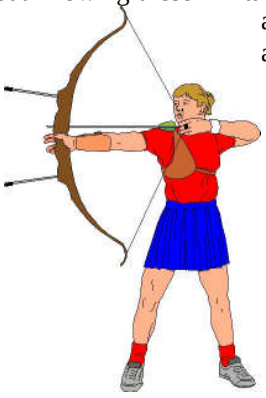
Validity refers to the degree in which a test measures what it intends to measure. There are a variety of issues and considerations involved when discussing validity. For example, content validity refers to the ability of the test to represent the content of a particular subject. Measuring content validity is a difficult task. To use intelligence as an example, an intelligence test would have to measure ability across a variety of aptitudes such as reasoning ability, analytical aptitude, etc. If an intelligence test measures only comprehension, reasoning, and analytical skill, for example, the test is therefore not measuring other intelligences such as practical or social/emotional abilities. Therefore, an intelligence test can only attest to measuring the type of intelligence tested. Consequently, there continues to be a debate over a single test's ability to measure intelligence.



Reliability

Test reliability refers to the replicability or consistency of the test. Again, there is much to take into account when discussing reliability. To begin with, we must consider that there will always be a measure of error in students' responses, whether those mistakes are random mistakes or systemic mistakes (in which case, comparative data is necessary to determine if similar mistakes are made across a population). Random mistakes, due to a student having a bad day, not having breakfast, or having an argument before the assessment are all factors that contribute to errors in measurement.

Both assessment validity and assessment reliability have many more variables than these examples provide. Designing a test with high measures of reliability and validity may be difficult, but knowing these limitations helps test designers ensure they are consistently measuring outcomes as accurately as possible.



Norm-Referenced Tests compare an individual student's performance with that of his or her peers. When these tests are developed, the questions are given to a large population or sample and norms are established using statistical measurement techniques. For example, when establishing norms for the Canadian Achievement Tests, psychometricians sample 44 000 Canadian students and ensure there is a representative sample of the provinces and genders as well as urban, aboriginal and minority populations.

Criterion-Referenced Tests assess student mastery of specific goals or objectives. These are less global than norm-referenced and are closely linked to curriculum or a set of competencies. For example, specialist teachers and ministry staff develop the questions on the provincial assessments in mathematics to align with specific curriculum outcomes and competencies, as outlined in the Atlantic Canada Mathematics Curriculum.

Definitions from: Shay Schumm, Jeanne (2006). *Reading assessment and instruction for all learners*. New York: Guilford Press.

On a global scale, teachers, parents, school administrators, district staff, provincial ministries, and teaching unions continue a conversation as to the potential effects of large-scale, external, 'summative' testing. Before launching into the issue, it is important to establish that the above groups and individuals are engaged in this debate for the purpose of ensuring the best interest of students. It is also important to note that when engaged in this discussion, focusing on the New Brunswick context is an important perspective to maintain due to the particularities of different educational contexts. In New Brunswick, the purpose of the external assessment program is to gather data and track student progress so as to inform programming. These are very different purposes from other jurisdictions where 'high-stakes' is truly high-stakes for teachers, where incentives and penalties are issued for test performance, and for schools and districts where school closures are possible for underperformance. With the best interest of students as the touchstone, and New Brunswick as the context, some of the issues with regard to our provincial assessments can be addressed with research as follows:

Are provincial assessments formative or summative?

Assessment serves a variety of purposes: as a measure of understanding of a unit of study (summative assessment), or as an ongoing process of teaching and learning (formative assessment). Often defined as assessment *for* learning, formative assessment is an ongoing, dynamic process and takes place in the classroom when both the teacher and the student can make decisions towards promoting further learning. Defined as assessment *of* learning, summative assessment is used to make a judgement, such as what grade a student will receive on an assignment, or to determine whether a particular program was effective.

It is important to appreciate the complementary nature of the different assessment purposes. The philosophy behind both forms of assessment is to inform teaching and learning. Large scale assessments are a mere snapshot of student achievement and the value is in the scale of data. Teaching and learning in the classroom involves collecting information from a variety of sources on an ongoing basis, both gauging and stimulating learning. Rick Stiggins suggests that formative assessment **supports** learning and summative assessment **verifies** learning.

If large-scale provincial assessments aren't formative, then why do them?

The implementation of a provincial-scale student achievement measure is critical to support learning. Without a large-scale picture, it would be very difficult to accurately, and fairly, inform programming and identify schools or areas of the curriculum that have specific needs.

In the 1960's, James Samuel Coleman produced a comprehensive report for the American government. In this report, *The Equality of Educational Opportunity*, Coleman found that economics, in large part, determined the success of the student. Schools were thought to have no more than a marginal effect on a child's development (Rutter, 2002).

In 1979, while researching reading difficulties and emotional and behavioural problems, Michael Rutter discovered that pupil achievement in low SES students was indeed affected by the school. After further research, they found direct correlates of school quality affecting student achievement despite economic status (Rutter, et.al., 1979). Rutter's work in the U.K. and later Lezotte's similar work on school effectiveness served to shift the responsibility of student achievement from the student to the schools, and school management shifted to a more research-based data-driven focus with implementations such as curriculum standards, an increase in teacher training, standards, and tools focussing on results (Anderson, AERA, 2001).

Around the same time, technological innovations created the ability to collect and analyze student information never previously possible due to the bulk and limitations of paper-and-pencil record keeping. Tracking student data based on economics, gender, or minority group and correlating that data against large sets of testing data helped identify trends, outliers, and issues. For example, Anderson (2001) sites one research study where they were able to mine provincial data to find a reliance on specific behaviour programs based on gender and minority group. Once this trend became visible through utilizing new data systems, they could use the information to implement intervention programs. Of note, one of the correlates of school effectiveness found in both Rutter's and Lezotte's research is the use of pupil achievement as the basis for program evaluation. Assessment data is necessary as a diagnostic tool in whatever form or scale the data is obtained. The challenge is to use that information to improve teaching and learning.

The benefits of the information obtained are clear; however, educational stakeholders also have to be aware of the potential pitfalls associated with large-scale assessments. Promoting unfair competition, teaching to the test, and narrowing curriculum for 'testable skills' are noteworthy complaints and will be addressed in the next Information Bulletin. To contend with these potential pitfalls, James Popham recommends that all parties involved in education, in particular our students, need to develop assessment literacy (*Educational Leadership*, 2008).