Valid Uses of Student Testing as Part of Authentic and Comprehensive Student Assessment, School Reports, and School System Accountability

A Statement of Concern from the International Confederation of Principals

by William J. Grobe and Douglas McCall

Introduction

In recent years, governments worldwide have increased their use of standardized tests and other forms of external tests to measure and report on student achievement in selected academic subjects. That trend, responding to public pressure on decision-makers to make school systems more transparent and accountable, is often colored by criticisms of the effectiveness of public schooling (Beldon, Russonello, and Stewart 1998).

Schools and school districts often publish the results of such large-scale tests with little apparent regard for their limited use in improving student performance and system monitoring (Bartley and Lawson 1999). Such tests are often not situated within a coherent policy and accountability framework based on learning and overall assessment of student achievement. In addition, the results of the tests are often not correlated or analyzed by context, student and family characteristics, or other factors that determine school or student success (CISA 2001). The tests often provide no information that helps students and educators improve their practices. Moreover, invalid uses of large-scale testing have been exacerbated by the news media, narrowly focused interest groups, and elected officials (CASA 1995b).

This paper has been prepared because school-based administrators are increasingly concerned that current policies and practices on student testing are causing negative consequences (CASA 1995a). They include

- unfair and invalid assessment of student, school, and system achievement
- a secretive or unintended shift of priorities to focus on a narrow range of student knowledge and literacy-numeracy skills despite the continuing stated and unstated pressures on schools to achieve broad mandates related to other knowledge, vocational preparation, and social development
- ill-informed and nonproductive public debates about schooling that are based on single test results and rankings without obviously required contextual information that is readily available
- few decisions that lead to reallocated or increased resources

Indeed, school leaders report that already some schools are returning to "teaching to the test" and other poor pedagogical practices (Elmore 2002). A recent study in the United States (Amrein and Berliner 2002) shows several unintended consequences emerging from the overuse of large-scale tests. For example, extensive high-stakes testing correlated with declines in results on other standardized tests that have been developed with higher degrees of objectivity, such as the SAT.

School leaders around the world believe that many current student testing policies and related accountability practices contain flaws that detract from our primary goal of improving schools (AASA n.d.). The apparent preoccupation of many jurisdictions with providing comparative statistics on a narrow range of student abilities presents an incomplete and misleading picture to the public. Although our organization, the International Confederation of Principals (ICP), does not reject out of hand the need for culminating, comprehensive experiences that assess academic achievement, we believe that school leaders must, however, differentiate between examinations set by governmental agencies that have a political purpose of their own and those created by educators to substantiate growth. To that end, the ICP recommends that *testing should be based on what we know about learning.* (See "Learning Principles" sidebar, p. 134.)

> Promoting children's learning is a principal aim of schools. Assessment lies at the heart of this process. It can provide a framework in which educational objectives may be set and pupils' progress charted and expressed. It can yield a basis for planning the next steps in response to children's needs. . . . [I]t

> should be an integral part of the educational process, continually providing both “feedback” and ”feed-forward.” It therefore needs to be incorporated systematically in teaching strategies and practices at all levels. (TGAT 1998)
> [See “Principles of Assessment for Learning” sidebar, p. 139.]

Large-Scale Tests: Uses and Abuses

As reported by A. C. Porter (2002), many principals have expressed concern that large-scale tests, by measuring only the cognitive domain and often using only multiple-choice or limited-response questions, reduce the learning being measured to its lowest form and fail to reflect the wide range of skills, knowledge, and attitudes that schools are required to foster among students.

The narrow range of the tests likewise narrows the type of education being offered. Testing should reflect the broad goals of education, not learning in its lowest form and more than simple memorization skills within the cognitive domain. It should measure student cognition such as performing required procedures, communicating their understanding, and solving routine and non-routine problems. Testing should provide an opportunity for students to generalize, prove, and make educated conjectures about the content.

However, tests are often based on recently developed curricula, for which the intended learning outcomes have been poorly communicated. Teachers often have inadequate opportunities to learn the curricula and adapt their teaching practices. The content of the curricula required as the basis of the test is rarely reviewed and prioritized to ensure that the stated outcomes can be achieved within the time available for instruction. The learning outcomes for the curricula that will be tested are not always stated clearly and in measurable, specific terms.

Studies conducted in other jurisdictions indicate that large-scale testing can cause more qualified and experienced teachers to leave schools that repeatedly score low in the publicized rankings. Such schools often serve disadvantaged students or those with special needs, so those students are increasingly underserved by the public school system.

The purposes and rationale for examining the knowledge, attitudes, and skills that are to be measured by a test should be presented and discussed before developing a test. That rationale should explicitly describe the purposes of the test, how it will help students, teachers, and parents to improve learning, and how the results will be used in decision-making. The rationale should also describe in detail the consultative processes and independent expert reviews that will be used to develop, implement, interpret, and announce the results of the tests. The tests used by education authorities should be developed, administered,

Learning Principles

1. Learning needs to be activity-based.
2. Learning needs to include cooperative learning opportunities.
3. Learning is dependent on situations that are meaningful to the child.
4. Learning needs to address attitudes and values.
5. Learning needs to encompass the use of literature.
6. Learning needs to develop critical thinking skills.
7. Learning is affected by developmental stages.
8. Learning is affected by evaluation strategies.
9. Learning is dependent on developing communication skills.
10. Learning is reinforced through integrated experiences.
11. Learning needs to be promoted without gender bias.

scored, interpreted, reported, and acted upon in explicitly defensible ways that are based on solid research evidence (ETS n.d.).

A clear distinction between tests used for accountability and monitoring purposes and those used for the improvement of student learning is not always maintained. For example, state- or province-wide tests involving all students are unnecessary if the purpose is to monitor program effectiveness. A random sample of students would suffice. Because of the number of students, the marking of such tests often takes months and feedback is provided several months after the test, making such tests meaningless to students, teachers, and parents. Further, the participation of disadvantaged groups, which typically are defined in terms of cultural differences, disabilities, language, access to community and family resources, etc., is typically not sufficiently taken in to account in the preparation, delivery, and interpretation of the test.

In addition, the administration and interpretation of the tests should respect well-defined professional practices and standards (CPA 1996).

Educators are finding that large-scale testing is placing a disproportionate burden on schools, and that the balance between program development and program evaluation is not being maintained. For example, one jurisdiction is spending one dollar per pupil on curriculum development while spending three dollars per pupil on large-scale student testing.

Large-scale tests developed by governments are often not reviewed by independent experts and others who reflect the diversity of the test takers, parents, educators, and other constituencies involved with the curricula, program, and schools. The results are often announced and prematurely presented to the public as being reliable. The research indicates that such tests require a minimum of three years' use before they should be considered reliable. Nor are large-scale tests often subjected to scrutiny through independent surveys to determine if they are considered credible by representative samples of teachers, parents, and students.

Often estimates of reliability and standard errors of measurement for the tests are not well understood by the news media and the public, leading to misinterpretation and false claims. Confidence intervals should be provided as well as the procedures used to obtain samples, and the nature of the populations being studied should be described. Those concepts and cautions should be clearly articulated in all written documents, emphasized in all announcements, and explained clearly at every opportunity to parents, teachers, and the news media.

Baselines and benchmarks (interpretations of "high" standards and levels) are often undertaken after the test is completed, and some reports indicate that education authorities have sometimes manipulated such standards after the test results are compiled. To prevent such abuses, benchmarks, or standards for achievement, should be based on scientific, published evidence that they are achievable and that achieving those benchmarks will have a long-term impact on outcomes later in the student's life and career. Data sets and tables should be freely available to qualified independent researchers so that they can conduct secondary analyses of the data without interference or control by education authorities or governments.

Yet other abuses of large-scale testing abound. For instance, the reported results of large-scale tests often ignore the number of times the participating students have taken such tests. For example, students in Alberta province are often among the highest scorers on international and national tests. The students in Alberta are also the students who most often take such tests. Similarly, students in Asian countries are tested more often than students in other countries because they tend to score well in large-scale tests.

In another variation, results of tests are often released without prior knowledge and consultations with the key stakeholders. That practice often leads to public debates and widely varying interpretation of the results, sowing confusion among parents and the public. The lack of a protocol and process enables interest groups to interpret the data from anomalous perspectives. Although all forms of debate and analysis are welcome, it behooves the leaders and decision makers within the system to analyze jointly the results with a view to identifying common interpretations and plans of action to improve the results and learning process.

Often the results of one test are equated with similar tests, with earlier versions of the test, or with similar tests at different grade levels or ages. That practice is inappropriate unless such comparisons have been specifically planned in the development of the test. Often, such comparisons employ "item response theory" and are therefore the product of arcane mathematical and statistical interpretations. In effect, apples are compared to oranges and the results the test suggests are of no value.

Media treatment of the test most often adds to the confusion. News media representatives are often not fully briefed about the interpretation of the results and meetings are often held without media editorial staff to explain the results.

Most large-scale tests currently in use do not measure a variety of forms of intelligence and learning styles (Rudner and Plake 1989). Instead of covering all required subjects and courses, they examine only a selected few subjects. Testing and publishing test results create pressure on schools to focus on the tested subjects, to the possible neglect of many subjects currently accepted as part of a well-rounded, required school curriculum. If subjects—from literature, math, science, and history to family studies, physical education, computer skills, career education, and the arts—are less important than literacy, numeracy, and the ability to reason, they should be declared non-compulsory (Gordon 2002).

A Fair Assessment: Assessment *for* Learning

Testing should respect principles of fair assessment. Not only should assessment and evaluation be continuous, but they should be an integral part of the teaching-learning process as well. Assessment and evaluation should also take into account a child's learning profile (defined as including the child's cognitive, affective, and psychomotor domains as well as development level and learning style), and they should be designed specifically to assess particular and clearly stated instructional goals, objectives, and educational outcomes.

Process skills as well as content knowledge should be assessed and evaluated. The methods used should be both valid and free from language, gender, cultural, and racial bias. Reading and writing should be viewed as processes during assessment and evaluation. Finally, students and their parents or guardians should be active participants in assessment and evaluation, and evaluation procedures and results should be fair and expressed to them in clear language.

Assessments *of* learning and assessments *for* learning are both important. We already have many assessments of learning in place; therefore, if we are to balance the two we must make a much larger investment in assessment for learning.

It is tempting to equate the idea of assessment for learning with our more common term "formative assessment," but they are not the same. Assessment for learning is about far more than testing more frequently or providing teachers with evidence so they can revise instruction, although those steps are part of it. In addition, educators now agree that assessment for learning must involve students in the process.

When teachers assess for learning, they use the classroom-assessment process and the continuous flow of information about student achievement that it provides in order to advance, not merely check on, student learning (Cimbriez 2002). To do so, teachers undertake several steps:

1. understand and articulate the achievement targets that their students are to hit;
2. inform their students about those goals, in terms that students understand, from the beginning of the teaching and learning process;
3. transform their expectations into assessment exercises and scoring procedures that accurately reflect student achievement;
4. use classroom assessments to build students' confidence in themselves as learners and help them take responsibility for their own learning, in order to lay a foundation for lifelong learning;
5. translate classroom assessment results into frequent, descriptive (as opposed to judgmental) insights to help students improve;

6. continuously adjust instruction based on the results of classroom assessments;
7. engage students in regular self-assessment, with standards held constant so that students can watch themselves grow over time and thus feel in charge of their own success; and
8. actively involve students in communicating with their teachers and their families about their achievement status and improvement.

As it plays out in the classroom, the effect of assessment for learning is that students keep learning and remain confident that they can continue to learn if they keep trying to learn. In other words, students don't give up in frustration or hopelessness.

Recommendations

Our school administrators have identified several actions that, if implemented, could alleviate the abuses associated with large-scale testing.

1. *There should be public decisions based on test results and other assessment results that involve the allocation or reallocation of human, administrative, and financial resources* (Bond 1995).
   - Remedial and support programs should be readily available for students who fail the test.
   - The specific test results should be reviewed in sufficient detail to make necessary adjustments in the specified learning outcomes, curricula, teaching and learning materials, program, school organization, and teaching practices.
   - A timetable for formal, public review and decision-making based on the results must be published when the test is being administered.

2. *The results of student tests should be compiled into comprehensive, contextual school or community profiles composed of data from a variety of sources and made available to educators and parents for planning purposes*

   Such data could include the level of education, level of income, and first languages of people in the community. Descriptive and administrative data could include the number and nature of programs offered in the school covering several aspects of the school environment: e.g., the number and nature of students participating in extracurricular and community service activities; the qualifications of teachers assigned to teach the subjects or assignments; the number of parent volunteers; and the accessibility and service levels of student health, social service, youth, justice,

Principles of Assessment for Learning

Principle 1: Assessment for learning should be part of effective planning of teaching and learning.

Principle 2: Assessment for learning should focus on how students learn.

Principle 3: Assessment for learning should be recognized as central to classroom practice.

Principle 4: Assessment for learning should be regarded as a key professional skill for teachers.

Principle 5: Assessment for learning should be sensitive and constructive because any assessment has an emotional impact.

Principle 6: Assessment should take into account the importance of learner motivation.

Principle 7: Assessment for learning should promote commitment to learning goals and a shared understanding of the criteria by which they are assessed.

Principle 8: Learners should receive constructive guidance about how to improve.

Principle 9: Assessment for learning should develop learners' capacity for self-assessment so that they can become reflective and self-managing.

Principle 10: Assessment for learning should recognize the full range of achievements of all learners.

and employment workers from the community who are coordinated with the school programs.

The data should be grouped and accessible so that parents, educators, and others can make reasonable comparisons of their own schools' trends over several years with those of similar schools in other jurisdictions.

3. *The results of student tests should be included in an Indicators system that accurately monitors all the relevant factors that affect learning*:

   School systems should benefit from well-developed and -implemented Indicators or reporting systems (Lashway 2001). Unfortunately, Indicators systems have been badly misused by education authorities. Serious errors and invalid uses of Indicators include:

   - monitoring only student outputs and not reporting on context (student characteristics), inputs (financial and human resources), processes (program status and implementation), and long-term outcomes (relevance to post-graduation life and career achievement)
   - poor consultation procedures
   - reluctance to publish results by province or state unless the authorities control the data and the reporting mechanism

The International Confederation of Principals suggests that to implement these principles and practices, testing for monitoring and accountability purposes should be clearly separated from testing for student improvement and progress. For example, to monitor the effectiveness of a system or program, it is neither necessary nor cost effective to test all students when a random sample can provide the same information.

We also recommend that the learning outcomes of all curricula should be achievable, and consensus-seeking during the curriculum development process should not result in inflated or unachievable outcomes. We must ensure that all learning outcomes within curricula be stated clearly and ensure that stated learning outcomes are based on scientific evidence wherever available.

Educators should have access to the means to create authentic alternative assessment tools such as scoring rubrics, student portfolios, and locally developed tests and quizzes. We must ensure that secondary analyses of the large-scale tests are undertaken to measure the impact of factors such as socioeconomic status, family and student characteristics, community resources, and program resources such as school facilities, equipment, teacher qualifications, and school organization. At least three years of test development and piloting should elapse before test results are announced and before they are considered reliable.

In addition, independent confirmation and evidence are needed to determine that current tests are not biased for certain groups of students or toward certain learning styles. We must also seek independent confirmation that current tests are appropriate for their stated purposes and that the results lead to meaningful analysis, policy-making, program development, and professional development. Independent assessors should validate large-scale tests to ensure that they meet the standards defined by professional authorities and experts. Independent confirmation should also establish that such tests are not leading to unintended consequences such as “teaching to the test.”

Schools need the means to collect, analyze, and monitor local community, student, and program data so that comprehensive profiles of school communities can be created, monitored, and used in school planning and coordination with other agencies. Schools should use technology to access and analyze their own data.

Schools should consult stakeholders on how the results of large-scale tests should be released to the public and, in consultation with stakeholders, define a transparent process for the review, analysis, and joint interpretation of the results of large-scale tests. That process should note when and how decisions based on the results will be made. Orientation sessions for journalists held before the release of such data will help ensure that they are aware of the science underlying appropriate interpretations of the data from large-scale tests.

Finally, schools should provide independent experts with easy access to conduct secondary analyses of the results of large-scale tests. Data-collection procedures for international, national, and state or provincial tests and surveys should be consolidated so that the response burden on schools is reduced.

References

AASA (American Association of School Administrators). n.d. *Using Data to Improve Schools*. Arlington, Va.: American Association of School Administrators.

Amrein, A. L., and D. C. Berliner. 2002. *The Impact of High-Stakes Tests on Student Academic Performance*. Tempe: Arizona State University.

Bartley, A. W., and A. Lawson. 1999. *Varieties of Assessment: Issues of Validity and Reliability*. EQAO Research Series No 2. Toronto, Ont.: Education Quality and Accountability, Office of Ontario.

Beldon, Russonello, and Stewart. 1998. *Accountability for Public Schools: Developing School Report Cards*: *Findings of Group Research for Education Week*. Washington, D.C.: A-Plus Communications.

Bond, L. A. 1995. *Rethinking Assessment and its Role in School Reform*. Oak Brook, Ill.: North Central Regional Education Laboratory.

CASA (Canadian Association of School Administrators). 1995a. *CASA Comments: Valid Education Indicators.* Oakville, Ont.: Canadian Association of School Administrators.

———. 1995b. *CASA Comments: Fair Assessment of Student Achievement*. Oakville, Ont.: Canadian Association of School Administrators.

CPA (Canadian Psychological Association). 1996. *Guidelines for Educational and Psychological Testing.* Ottawa, Ont.: Canadian Psychological Association.

Cimbriez, S. 2002. "State-Mandated Testing and Teacher Beliefs and Practice." *Education Policy Analysis Archives* 10 (2) (January 9).

CISA (Commission on Instructionally Supportive Assessment). 2001. *Building Tests to Support Instruction and Accountability: A Guide to Policymakers*. Washington, D.C.: American Association of School Administrators, National Association of Elementary School Principals, National Association of Secondary School Principals, National Education Association, National Middle School Association.

Dietel, R. J., J. L. Herman, and R. A. Knuth. 1991. *What Does the Research Say about Assessment?* Oak Brook, Ill.: North Central Regional Education Laboratory.

Elmore, R. 2002. "Testing Trap." *Harvard Magazine* 10 (1): 35–37.

ETS (Educational Testing Service). n.d. *Basics of Assessment*. Princeton, N.J.: Educational Testing Service.

Gordon, D. T. 2002. "Moving Instruction to Center Stage." *Harvard Education Letter* (September/October).

Lashway, L. 2001. "Educational Indicators." *ERIC Digest*. Eugene, Ore.: ERIC Clearinghouse for Educational Management.

Porter, A. C. 2002. "Measuring the Content of Instruction: Uses in Research and Practice." *Educational Researcher* 31 (7): 3–14.

Rudner, L., J. Conoley, and B. Plake, eds. 1989. *Understanding Achievement Tests*: *A Guide for School Administrators*. Washington, D.C.: ERIC Clearinghouse on Assessment and Evaluation.

TGAT (National Curriculum Task Group on Assessment and Testing). 1998. *A Report*. Department of Education and Professional Studies. London: King's College.

William J. Grobe, Ed.D., is an associate professor of educational administration in the Department of Educational Leadership of the College of Education at East Carolina University. Douglas McCall is a staff professional at the Canadian Association of Principals.