

Standardized Tests for Dummies

by Mark Holmes

What is a standardized test?

A strict definition would exclude many provincial tests. The general everyday usage in Canada is adopted here, i.e., an external test administered to individuals, groups or populations in such a manner that scores of those tested are statistically calculated to be directly comparable within defined limits of reliability and validity. Reliability is a measure of the extent to which the same results are found if equivalent test items are given to the same individuals under the same conditions. Validity is the extent to which test items reflect the factors the test is supposed to measure; reliability is thus an important component of validity: a test can be reliable without being valid (one administered to the blind that depends on sight), but not satisfactorily valid without being reliable.

Familiar standardized tests include: IQ tests designed to fit the normal curve, e.g., with a mean of 100 and a standard deviation of 15; American college entrance boards, graduate admission and analogy tests; professional tests (e.g., for entry into law school); the CTBS and CAT tests of academic achievement; and provincial tests of achievement.

Typically, the standardized test is composed mainly or entirely of multiple choice items. The reasons for this are: generally speaking, multiple choice is the most valid, reliable and economical procedure. The common accusation that multiple choice measures only knowledge is totally false. In fact, high school essay examinations are more likely to be scored mainly on the basis of knowledge, even though the questions may be written in a form that appears to demand higher level skills. Standardized tests are generally the most effective way to test: knowledge, understanding, inference, interpretation, some skills, and analysis. They can, with difficulty and severe limitations, be used to measure synthesis, evaluation, and aspects of written expression. They are not useful measures of oral, artistic, moral, physical or creative achievement – some of which goals are challenging for any objective form of assessment. Standardized tests should not normally be used alone to make important educational decisions about an individual's future.

What is their purpose?

The use of standardized tests overlaps with that of other (notably teachers') tests. The purposes include: the measurement and diagnosis of students' achievement, aptitude, skills and measured intelligence; assessment of the passing of a specific academic hurdle; the certification of students, e.g., as high school graduates; the evaluation of schools and teachers; moderating movement from high-school to post-secondary education and accession of entry to professions and other employment;

Society for Quality Education, February 2007

and, perhaps most important, as incentives to teachers and students for improvement and excellence. Teachers who prepare and mark their own exams have less incentive than those whose students are subject to external assessment – which explains the continuous grade inflation in Ontario secondary schools. The choices among standardized tests and other assessment measures (e.g., teachers' marks of homework and projects, school examinations, and teachers' subjective assessments) depend on: validity and reliability of the measure; the kind of attributes being assessed; and the values and beliefs of those making the choice. In general, more than one means of assessment is important when decisions are being made about people's future lives.

Is there a good standardized test that parents can use at home to test their own children?

Essentially, the answer is no – a standardized test would not really be a standardized test if it were publicly available. Its reliability depends on its being used in secure circumstances. There is obviously no assurance that a parent would not teach to the test beforehand, give hints during the test, or do other things that would render the result meaningless. Although a parent might simply want to administer the test in good faith for his own knowledge, there would be nothing to stop that parent from subsequently claiming that his child had performed at x level on the test – or even lending the test to a friend whose child was going to take the test at school in a few weeks time. No test-maker will make his test available to the general public. The only solution is to pay to have one's child tested in various ways by a professional.

What kinds of standardized tests are there and how are they used?

A distinction is usually made between external criterion-referenced and norm-referenced tests. The former are based on the attainment of specified objectives, the latter on normal achievement levels. Unfortunately, 'criteria' have become synonymous with 'good tests' (or less bad), norms with 'bad'.

Educators sometimes say, "I like tests that measure students in terms of themselves, not those that compare them with others". That is absurd. Virtually all tests have both criteria and norms – and they all necessitate explicit or implicit comparison with others. The Canadian Achievement Tests are norm-referenced. That is to say, scores are provided by (national) percentile rank, grade level and stanine. In all three cases, the scores are arranged around a mean (the 50th. percentile, the grade level corresponding with that of the students taking the test, or stanine 5), along a normal curve. (The stanine represents the curve most

clearly – with many more students being a 5 than a 1 or a 9). However, the test items are still constructed on the basis of criteria; the criteria being the learning objectives of the various provinces. Test revisions ensure that the mean always represents the national average - irrespective of whether achievement is gradually increasing or (more likely over the last 40 years) declining. (Sometimes, there have been common items held over from one edition to the next so that one can show that the mean level of achievement has changed, a feature not beloved by school systems and provincial governments). Such tests, with their clearly defined interpretation, are unpopular with most educators, who also oppose high-stakes tests. High stakes have less to do with the qualities of the test than the use to which they are put. If achievement tests are used to evaluate teachers or to determine admission to universities they are high-stakes; if they are closely guarded in a drawer and never mentioned, they are low-stakes.

Criterion-referenced tests

Ontario's achievement tests at grades 3 and 6 are nominally criterion referenced. They were set up to be closely associated with the Ontario curricula, taking account not only of the often obscure learning objectives, but also the required teaching methodologies being used (anything except direct instruction). Hence the achievement of those criteria would be the central point, regardless of the average level of achievement. By criterion-referenced theory, one could have 100% of students at levels 3 or 4 (whereas with norm-referenced tests, the median student is always at the 50th percentile). However, in developing test items for use in schools, consideration is inevitably influenced by what the test-makers believe third- and sixth-grade students are able to do, in other words norms.

Back in 1965, I was made principal of a large junior high school. My predecessor was a progressive who did not believe in tests or examinations. I introduced examinations, which were developed by the teachers of the subjects at the three grade levels. The exams were in effect criterion-referenced (but not standardized) – based on what the teachers had taught during the term. The results were disastrous - teachers, parents and students were equally appalled. The simple fact is that students do not always learn what teachers have taught. That is arguably much more the case where there is no direct instruction, compounded by a lack of clear, measurable objectives and the absence of frequent formative testing, an integral part of direct instruction.

So it was not surprising that the first results of the Ontario tests were equally appalling, so much so that an informal zero had to be added to the four levels, and a revised claim was hastily made that 2 was really

equivalent to a pass mark (a claim that later had to be abandoned in the face of parental derision). Quite apart from the severe deficiencies of the tests (more in terms of their design than in their construction, e.g., a 4 was for those who had gone beyond the curriculum, the test thereby becoming a criterion test with invisible criteria), the fact is clear that there was an intended implicit norm; the professional educators believed that would be the score for a plurality of students. Like the teachers in junior high school, an entire province's educators (with noble exceptions) believed that because something had been "taught" (or more precisely, because students had been given the opportunity to learn that something), the students must mostly have learned it. Efforts have since been made to turn 3 into but an all but written norm, partly by changing the tests.

There are more difficulties in the design of criterion-referenced tests than in norm-referenced, whose idea and design is very straightforward. Ontario learned nothing from its first experience, and suffered the same disasters with its high school literacy and basic math tests – with criteria related to the sometimes-bizarre curricula rather than to the basic literacy and numeracy intended if they were to be associated with a reasonable standard for graduation.

At one time, I was involved in the writing of items for standardized provincial high school tests (not in Ontario). The tests were supposed to be criterion-referenced (but obviously the province did not want too many failing to meet the criteria). The province decided that the language test should be more "authentic", a buzz word at the time, close to the "real" lives of students. I first tried passages from local newspapers, one an article containing interviews with adolescent girls about why they smoked, another based on the proposed establishment of provincial casinos. The first was rejected for sexism (boys smoke too), the second because it was a political issue and the government wanted casinos. I quit at that point. The province accepted instead a nineteenth-century letter from a British soldier stationed in Canada, with questions based on the improbable assumption that his opinions at the time were based on fact. There are numerous testing problems here. The crucial objective of inference in reading was undermined during the search for "authenticity", as the difficulty of items comes to be sidelined by external factors unrelated to the educational objectives ostensibly being tested.

In another case, I was asked by an independent school to develop entrance tests for the school – there were two entry levels. The school, I was told, wanted to accept only the most able applicants, and they needed a way to discriminate accurately. I developed criterion-referenced tests, standardized by trials in public schools with gifted students. The initial response to the testing of applicants was very positive – they were engaged by the challenging tests and the tests discriminated

well. I was later told there was a problem. The school had applications from children of school graduates and benefactors who had not met the test criteria (i.e., who had performed less well than numerous others without those connections). Where could the school draw a line between those who could survive in the school, but without flourishing, and those who would not survive at all? I could not help – there were very few items at that lower level, insufficient to make a fair judgment.

This is not to say that criterion-referenced tests are bad tests, only that their development and use raise complicated issues and the apparent precision and singularity of their purpose mask their drawbacks. Alberta uses criterion-referenced tests (called “performance” tests) to measure mastery, with criteria related to the mastery of the essential objectives of the curriculum. The advantage is that it focuses students and teachers on the essentials. I am not familiar with Alberta’s schools and do not know to what extent that emphasis on mastery has been either helpful or modified over time, but there are possible problems. Should everyone who lacks “mastery” repeat the course? Are there incentives (in the form of higher-level objectives) to challenge those who easily achieve mastery and require a higher bar, or does the bar become a ceiling?

At one time, a colleague and I developed standardized criterion-referenced tests for the Ontario universities. Some of them wanted to have a test for first-year students to identify those needing remedial help in language (which tells us something about the absence of standardized tests when students move from school to post-secondary education). We developed four tests – language usage, reading, essay writing, and the structure of essays and reports. An interesting finding was the high correlation among the four tests, sufficient to permit the exclusion of any one of the four with little loss of explanatory power. One is reminded of the important function of incentives in testing; the temptation would be to exclude the essay-writing test on the grounds of expense and inconvenience. (The others were all multiple choice). Essay writing is after all a major goal. The research finding did confirm the fact that, with rigorous oversight and procedures, essay tests can be almost as reliable as other test formats. However, the much more complex question of validity remains open. Some research has suggested, for example, that the length of the essay turns out to be a strong correlate of other language measures. Is that what scorers sometimes subconsciously measure?

Norm-Referenced tests

Ontario parents and educators have little knowledge or experience of norm-referenced tests, although that does not prevent some educators from making ignorant and thoughtless comments: “Norm-referenced tests compare students with other students, criterion-referenced tests compare them with themselves”; “Norm-referenced tests measure things

that have not been taught”; “Norm-referenced tests are used to rank individual students, classes, teachers and schools”.

Almost every test, including medical tests, compares individuals or groups with an assumed norm. The issue should be: Is the comparison informative and useful? For example, as mentioned, the separation of levels 3 and 4, particularly in the original Ontario tests, was farcical. On some items students were expected to go beyond the “expectations”, while on others students were penalized for giving irrelevant information.

The idea that one should be compared only with oneself is absurd. Is it professional for an oncologist to tell a patient simply that her cancer is improving, when the rate of cell growth means death in two weeks instead of one, or for a teacher simply to tell a seventh-grade student only that his reading (over a grade below level) is steadily improving when he has advanced less than half a grade level in the year? Is it professional for teachers to substitute a vague and comforting parental and student interview, chaired by the student, for any testing at all? If one wishes to take computer science at the University of Waterloo, one needs to know how one is doing compared with others with that ambition, not whether one is doing vaguely better than last year. That point is so obvious that it is hard to believe so many “experts” remain in denial.

One great value of standardized norm-referenced tests is that they can be used to make valid comparisons among students, schools, provinces and countries. If Ashley comes home with an ‘A’ in grade eight science, a typical parent asks, “What did the rest of the class get?” or “What mark did Hank, Jill and Annette get?” – those three being in competition with Ashley. Even if universities accepted students according to an unchanging standard (in fact they cynically change their “standards” according to the required number of available bodies), parents would still need to compare their child’s achievement with that of others in order to give sensible career and educational advice. I have a grandson whose ambition is to be a helicopter pilot in the armed forces; he has the range of potential skills to be a good one, but he will have to work very hard to overcome the academic hurdles. If he happens to be in a school with unusually high standards, he will be disadvantaged in gaining acceptance to the program he wants. Ironically, in Ontario and other provinces where there are no meaningful standardized tests or examinations at the end of high school, it pays to attend a school with low standards, i.e., one that gives higher marks for a lower level of achievement. At an extreme the difference between schools’ average marks (for the same level of achievement), may be 10 percentage points; a variation of 2 or 3 is normal, small but still crucial. Canny students have been known to transfer from a good high school to a bad one for their graduation year.

The unfairness to students is not only in acceptance to programs - scholarships are also typically awarded on the basis of raw school marks. A highly professional and successful teacher in one of the best public high schools in Toronto once asked me, "What can I do? To give my students a fair chance I have to keep lowering my standards and giving higher marks for poorer work?" I had no answer.

Another advantage of norm-referenced achievement tests is their economy. They are easily administered and quickly scored (in most cases) by machine. This means that they can be easily taken annually, giving parents a quick warning if things are going astray. As a principal who administered tests annually (together with the now banned, politically-incorrect tests of ability), I used them to identify less advantaged children with unrecognized potential. Primary teachers frequently confuse readiness, maturity, and appropriate attitudes with ability and achievement. Thus highly-intelligent disadvantaged children, lacking promotion and teaching from home, may fall behind before they are recognized, if they ever are. After five years of schooling at the child's "own rate of progress", such children are likely to be irremediably lost by the end of third grade, quite possibly becoming troublemakers. In addition, annual testing is useful as a way to identify the quality of teaching. Teachers understandably object that single shot tests every few years reflect their students' background culture more than their teaching. But if the same class achieves an average gain of 1.2 grade levels a year for four years and only 0.8 of a grade level in the fifth year, there is a problem. The problem may or may not be the teacher, but it requires recognition and investigation. The absence of annual tests also makes it easier for educators to deceive parents when they flatly refuse to consider allowing an outstanding student to skip a grade.

European examinations (including the International Baccalaureate) at the end of high school are generally hybrids of criterion- and norm-referenced tests. Their criteria are more closely related to curricula than are American college boards (e.g., SATs) and have traditionally themselves set the criteria (the program is what the exam tests), as was the case in Britain, or closely followed a national curriculum, as in France. However, as academic success has become more competitive, the problems of varying scores from year to year attributable to varying difficulty in the examinations have led to increasingly-acknowledged statistical norming.

A complication with Canadian (and, today, many European) final-year exit tests and examinations is the problem of combining two very different assessments – the teachers' and the tests'. Teachers and some parents argue that the teacher knows the student best, that the teacher's mark, based on ten months' experience, is more valid than the score on a

test taking a couple of hours. On the other hand, there is the problem of teachers' differing standards and of their essentially marking their own success; not many teachers like to think they have failed with a large portion of the class. The answer in most jurisdictions has been to give "equal weight" to each.

That sounds fair. Teachers tend to value attitudes and behaviour such as effort, diligence, memory (of what the teacher has said), and compliance (one can't grade an assignment not done). School assignments are sometimes influenced by the knowledge and diligence of a parent. Tests place greater value on accuracy, inference, logic, knowledge, analysis and skill. Both sets are important; teachers' marks tend to correlate better with future teachers' (professors') and test results better with other external measures such as tests of aptitude. In some fields of work, diligence and manners are crucial; in others it helps more to be smart. There is of course considerable overlap between the two measures.

British Columbia, for example, simply adds the two scores (teacher and test) together. This procedure does little to resolve the inherent unfairness of the teacher or school score described above. If two scores are added, one objective, one subjective, the result is subjective. Quebec has the answer. A school's marks are transformed to the metric of the same student's performance on the external exam. The students are then scored in the combined results in terms of their collective performance on the external exam, but individually are ranked with half the weight given to their school marks. School mark inflation or deflation has no effect on the combined score. The effect of a complex formula is to provide a student's final score based on the same scale across the province, but whose rank, compared with other students in her school, is based equally on his or her scores on the two measures.

One of the common objections to external tests is that "they do not test what I teach". Superficially, that seems reasonable; it is not fair to test a student in advanced physics who has not taken the relevant course. In practice, commercial achievement tests are based on the common objectives of provincial curricula – and, at least in the basic skills, they are not nearly as different as provincial bureaucrats like to imagine. Canadians are a mobile people, but students have more difficulty with social than academic adaptation.

What should be our priorities for standardized testing?

It should be clear that the debate should not be between norm-referenced and criterion-referenced tests. In practice, tests are often hybrids – even teachers' tests and examinations, not standardized, have implicit norms as well as (one hopes) explicit criteria. Sometimes those

informal norms defy common sense when one looks at how teachers' marks are used. For example, why should English marks in grade 12 usually fall between 50 and 90, math marks between 20 and 100? Why should it be easier to both get a scholarship and fail in math and science than in English and history?

The weaknesses of teacher and school marks are most evident in terms of graduation and movement to post-secondary education. Graduation should be based on clear standards – as it is in most jurisdictions except Ontario. Students throughout the province should be judged on the same basis in order to win a scholarship or gain entry to a highly competitive program – just as they would to win a hockey scholarship or win a prize in gymnastics. In neither of the latter cases would they win simply on the basis of their own coach's assessment. This is the first priority – and one which nearly all developing countries recognize. Of lesser importance is the choice between criterion-referenced and norm-referenced tests.

In the case of grade 12 exit assessment, my preference would be a hybrid, along the lines of the International Baccalaureate, but at two levels, both lower than the I.B. itself which would remain an option for the most advanced students. The lower level would essentially consist of tests of what is needed for good citizenship, with emphasis on literacy, numeracy, and basic knowledge and understanding in science, finance, health, etc. Test results should be combined with teachers' assessments of work over the year and final exams, notably in areas not well assessed by objective tests.

In the lower grades, annual tests are necessary – too much can go wrong in three years and accountability is important throughout the system. Norm-referenced tests are the most efficient and economical and serve the broadest band of purposes.

Criterion-referenced tests are most important in areas such as physical fitness and the arts. There are basic standards of fitness which should be required of all students. Similarly in music, there are basic standards in knowledge, appreciation as well as choral and instrumental performance that are valuable to all.

None of these things is particularly difficult or expensive. The smaller provinces could form a consortium to produce and implement common measures. The idea that New Brunswick students need different tests from those in Nova Scotia and Manitoba is simply absurd in the twenty-first century. It is unrealistic to imagine that Ontario would ever condescend to join such a consortium. Its problem is not financial. The evaluation deficit in Canada is one of will, not money.

(Dr. Holmes is Professor Emeritus at the Ontario Institute for Studies in Education. He lives in Port Hope, ON.)