Characteristics of an Effective Student Testing System

by Richard P. Phelps

The U.S. public has consistently favored standardized testing in the schools, preferably with consequences (or "stakes") riding on the results, ever since the first polls taken on the topic several decades ago. Depending on how the question is framed, those in favor of high-stakes standardized testing outnumber those opposed at ratios as high as twelve to one. Parents are stronger supporters of high-stakes testing than are nonparents, and that support does not budge when they consider the possibility of their own progeny failing.

Results from different polls approaching the topic in different ways suggest that nearly all Americans would like to see high-stakes tests administered at least once at every grade level. In twelve years of elementary and secondary school, however, the typical U.S. school district offers just one or two standardized tests with high stakes for students.

With only a few exceptions, U.S. educational testing programs fall short of what the public wants, and short of what most industrialized countries have.[1]

Comprehensive examination systems are multi-level and multi-targeted

A comprehensive testing system captures the complete benefits of standardized testing—and for all students, not just some. Those benefits include:

- *Information* that can be used for diagnosis (of individual students or teachers, of schools, of school programs)
- *Efficiencies* from alignment, when the tests are matched to curricular standards and teachers teach to those standards (and yes, teach "to the test," as they are supposed to do with standards-based tests)
- *Motivation* to study and to attain goals

The best testing regimes, such as one finds in many European and Asian countries, capture those benefits through multi-level and multi-target systems.

Multi-level means administering high-stakes tests at more than just one educational level (i.e., primary, intermediate, lower and upper secondary). European and Asian students typically face high-stakes tests at the beginning or the end (or both) of one educational level, and often for more than one educational level (e.g., the end of primary school, the beginning and end of lower- and upper-secondary school, the beginning of postsecondary education, etc.).

Multi-target means that every student, regardless of achievement or choice of curriculum, faces a high-stakes test that, ideally, offers a challenging but attainable goal. In some systems, tests are set at multiple levels of difficulty with multiple levels of certification (e.g., a "regular" diploma and an "honors" diploma). In other systems, different tests cover different subject matter.

In the United States, high-stakes student tests are uncommon at any but the upper-secondary level (i.e., high school). Moreover, with few exceptions, they are single-target tests—meaning that every student, no matter what level of achievement or ability, course selection, or curricular preference, must meet only one common standard of performance.[2]

Ironically, largely socialist Europe, with its relatively smaller socioeconomic (and academic achievement) disparities, acknowledges children's differences by offering a range of academic options and multiple achievement targets. The more libertarian United States, with its relatively large socioeconomic (and academic achievement) disparities, nonetheless typically provides all children the same curriculum—often called the "college track"—and sets a single academic-achievement target.

A single academic-achievement target must by necessity be low: otherwise, politically unacceptable numbers of students will fail to reach the targets. School systems that set low targets typically focus on bringing the lowest-achieving students up to that target. Unfortunately, they may also neglect average- and higher-achieving students (or, in the most perverse cases, deliberately hold them back). Schools judged as a whole on student performance can increase their average scores, for instance, by retaining high-achieving students with their age-level peers rather than letting them advance a grade or by making them take courses in basic subject matter they have already mastered.

The differential effect may help explain why some minority-rights advocates support minimum-competency testing, while parents of "gifted and talented" children often oppose it. From the perspective of

The "High Stakes Cause Test-Score Inflation" Myth

John J. Cannell's late-1980s "Lake Wobegon" reports suggested widespread, deliberate educator manipulation of norm-referenced standardized test administrations and scores, resulting in artificial test-score gains—such that every U.S. state had "above average" test scores. The Cannell studies have consistently been referenced in education research since, usually as evidence that high stakes, not cheating or lax test security, cause test-score inflation—this despite the fact that only one of the dozens of Cannell's score-inflated tests had stakes attached.

In fact, a careful reading of Cannell's reports shows that low, not high, stakes are associated with test-score inflation. Low-stakes tests make cheating possible because those tests are often administered with lax or no security.

Conversely, high-stakes tests are more likely to produce reliable test results because those tests are typically administered with tighter security. Given current law and practice, the typical high-stakes test is virtually certain to be accompanied by item rotation, sealed packets, monitoring by external proctors, and the other test-security measures itemized as necessary by Cannell in his late-1980s appeal to clean up the rampant corruption in educational testing and reporting.

Other test-security enhancements that also tend to accompany high-stakes tests include a high public profile, media attention, and voluntary insider (be it student, parent, or educator) surveillance and reporting of cheating. Do a Web search of stories about test cheating, and one finds that, in many cases, cheating teachers were turned in by colleagues, students, or parents. Public attention does not induce otherwise honest educators to cheat; it enables otherwise successful cheaters to be caught.

—R.P.P.

See Phelps, R. P. 2005. "The Source of Lake Wobegon." *Third Education Group Review* 1(2). <http://www.thirdeducationgroup.org/Review/Articles/v1n2.htm>.

the former, the tests pull achievement levels up. From the perspective of the latter, the tests pull achievement levels down.

The single-target problem has two solutions, one passive and one active. The passive solution, currently used in many U.S. states, essentially involves letting individual students take the minimum-competency test early in their school careers; once they pass it, they are allowed to move on. If the test is high stakes only for individual students, no one has an incentive to hold higher-achieving students back, that is, to prevent them from taking accelerated course work from then on.

The active solution to the single-target problem, and the solution that promises greater overall benefits, is to offer multiple targets. New York has stood out historically as the one U.S. state that employs a multiple-target examination system, with its Regents "Competency" exams and Regents "Honors" exams. The former was required for high school graduation with a "regular" diploma, while the latter was required for graduation with an "honors" diploma.

European and Asian examination systems exist in a variety that reflects the educational programs offered. Students are differentiated by curricular emphasis and ability level, and so are their high-stakes examinations. The differentiation, which starts at the lower-secondary (i.e., middle school) level in many countries, exists in virtually all of them by the upper-secondary level. Students attend schools with vastly different occupational orientations: advanced academic schools to prepare for university; general schools, for the working world or for advanced technical training; and vocational-technical schools, for direct entry into a skilled trade. Typically, all three types of school require an exit examination for a diploma. Some of those exams are very tough.

Supporters of the one-size-fits-all U.S. system often label the European system as "elitist" and our system as a more "democratic," "second chance" system. That contrast may have been valid forty years ago, but no longer. It is easier to enter upper academic levels in the current European systems, and most countries now offer bridge programs for, say, a dissatisfied vocational-track graduate to enter university or advanced technical programs. Typically, bridge programs are free of charge.

If the U.S. system is neither less elitist nor more conducive to "second chances," how is it superior? It is not, really. In the typical European system, multiple programs and multiple tracks offer multiple opportunities for students to attain high achievement levels. A Swiss, German, Danish, or Austrian student who enters a vocational-technical track at the lower-secondary level and finishes by passing the industry-guild certification examination as a machinist enters an elite of the world's most skilled (and best-paid) craftspersons. By contrast, a vocational-technical student in the United States may be stigmatized by a curriculum with a

reputation as a "dumping ground" and receive only low-quality training, with out-of-date equipment, for low-level jobs.

Fair high-stakes tests are aligned to standards

Most high-stakes student examinations are aligned to common standards; it is simply not fair to attach stakes to a test containing content to which students have not been exposed. What is more, no standards-based test, regardless of the care and effort put into writing it, can salvage poor curricular standards.

Profound disputes over curriculum and instruction are the major reasons high-stakes state tests can vary so widely in character. Take the neighboring states of Maryland and Virginia, for example. Several years ago, Maryland's School Performance Assessment Program (MSPAP) incorporated test administrations at three different grade levels, and performance carried consequences for schools. The entirely "performance-based" test had no multiple-choice test items and even included group work and "hands-on" demonstrations. It emphasized "process" over content. By contrast, Virginia's traditionally administered, content-oriented Standards of Learning examinations (SOLs) contained a large proportion of multiple-choice items, completed in their entirety by individual students.

Different theories about what should be taught and how it should be taught underlay the development of the Maryland and Virginia examination programs. To be sure, different theories of assessment were also involved, but they were inextricably tied to curricular and instructional preferences. Only the most extreme testing opponents decried both the Maryland and Virginia tests. Many "progressive" educators liked Maryland's, whereas many "traditionalists" preferred Virginia's.

Examination systems require careful implementation

Even if one assumes that, say, the French examination system is worth emulating, a U.S. state with no testing program cannot replicate it overnight. The French system is supported by a relatively uniform curriculum-development system, which is managed by university subject-area experts. This developed curriculum buttresses several (or many, in vocational areas) curricular tracks that students can follow. Students are provided multiple opportunities to pass the examination of their choice, and they receive substantial help, such as further classes and tutoring, to pass those examinations. But they must pass before they can go on to university, polytechnic, or a specialized trade. Although given every reasonable aid to succeed, in the end they must know the basic subject matter of their chosen path, or they will not be allowed to proceed at taxpayers' expense.

Two general issues are involved in building an examination system: sequencing and structuring.

Sequencing is the more straightforward of the two. Implementing a standards-based test requires time and care. The standards must first be in place, and taught to, before students can be tested fairly. New tests then should be field tested to address problems and set baselines for performance. The process takes at least a year, and more commonly two or three.

Most U.S. states building new examination systems have started with tests designed to measure minimum levels of knowledge and skill deemed adequate for earning a high school diploma. Passing a minimum-competency examination can frequently imply nothing more than a sixth-, seventh-, or eighth-grade level of achievement.

A minimum-competency exam brings a state only to the edge of the French examination system, however. There, minimum-competency examinations are given at the end of lower-secondary school; passage is required before a student can move on to upper-secondary or specialized vocational schools. The students who advance through this next level of education in France choose a curricular track and then face tough exit examinations of a type and level of difficulty that scarcely exist in the United States.[3]

Examination difficulty, in fact, introduces another aspect of sequencing. Some U.S. states have constructed high-quality examinations aligned to their standards only to discover, during field tests, that few students could pass them. Any state in which the majority of students fails a test required for graduation—only a year after the untested students of the previous class all graduated—will face a public uproar.

Aside from merely easing the difficulty of the examination, the problem has at least two feasible solutions. One can start easy and gradually ratchet up the difficulty level of the examination, or one can provide students extra assistance to pass the examination. The latter strategy was adopted with great success in Massachusetts. The state's elected officials, however, absorbed considerable invective from testing opponents (including one state teachers union) while they stood firm on the standards. The Massachusetts strategy is not for the faint of heart.

Structuring is more varied and complicated than sequencing. In countries with well-developed testing systems, two general types are distinguishable by relative degree of curricular specialization, or "splintering."

- The French example above describes considerable curricular splintering or tracking, common to the European "continental" system. Starting at the beginning of lower-secondary school, or perhaps even earlier, students are tracked into different types of

schools according to ability level and personal and parental choice of curricular focus.

- The traditional "two-tiered" British system represents another examination system: general curricula well into high school, but at two levels of difficulty—the "O," or ordinary level, and "A," or advanced level. The former two-level Regents examination system in New York State also represented this model.

Whatever the testing-system model employed, it should make sense as an integrated whole. To be fair to all students, a testing system should offer opportunities and incentives to all students, and students are not all the same.

Consequences of eliminating standardized testing

Standardized testing has often been measured against utopian perfection rather than what is likely to take place in its absence. It is true that neither standardized tests nor the manner in which they are administered will ever be perfect, but the consequences of abandoning standardized testing are far from perfect, too.

One likely consequence of eliminating standardized testing is a system of social promotion with many levels of nominally the same subject matter, ranging from classes for the self-motivated kids to those for the kids who quit trying years before, kids the system has ignored ever since. Too often, the result is a system that graduates functional illiterates. In schools where students are routinely passed whether or not they earn it, teachers brave enough to assign failing grades may well have their marks erased and changed by school administrators, thus allowing the failing students to graduate and avoiding controversy. In schools where some students pass courses and graduate despite doing little work, other students, and their parents, will assume that they, too, can pressure school administrators for easy credentials. Behind-the-scenes prerogatives become the implicit academic standards.

Another likely consequence of eliminating high-stakes standardized testing is the large-scale institution of remedial programs in colleges to compensate for any deficiencies of instruction in elementary and secondary schools.

A third likely consequence of eliminating high-stakes standardized testing is a blackout of reliable information on student performance anywhere outside a student's own school district. Eliminating *high-stakes* standardized testing would increase schools' reliance on teacher grading and testing, which are far more likely to prove idiosyncratic and nongeneralizable than any standardized test. Individual teachers can narrow the curriculum to what they personally prefer. Grades are susceptible to

inflation as students learn teachers' idiosyncrasies and how to manipulate their opinions. According to research on the topic, many teachers, when assigning marks, tend to consider noncognitive outcomes, including student class participation, perceived effort, progress over the period of the course, and comportment. Actual subject-matter mastery is just one among many factors. Moreover, given most teachers' relatively brief training in testing and measurement, it is not clear that their testing and grading practices would be superior even if they focused only on subject-matter mastery.

If the curriculum is not tested, it is difficult to know if any of it works. Without standardized tests, reliably gauging student progress becomes problematic for anyone outside the classroom. One must accept whatever each teacher says, and without standardized tests, points of comparison for different classrooms become progressively rarer.

Without either common standards or high-stakes standardized tests, there may be no effective way to monitor systemwide performance *at all*. Some U.S. teachers may be doing a wonderful job in their totally customized classes, but some may be doing an awful job. How is one to know or tell which? One must hope that teachers will face down their own natural inclinations as well as those of students, parents, and schools to avoid accountability and hold themselves and their students to high standards of performance regardless. One must also hope that teachers will know how.

This document excerpts from Kill the Messenger: The War against Standardized Testing, *by Richard P. Phelps, published by Transaction Publishers. Copyright © 2003 by Transaction Publishers, Rutgers University, 34 Berrue Circle, Piscataway, NJ 08854. ISBN 0-7658-0178-7 / cloth / 331 pp.*

Richard P. Phelps, the author of Kill the Messenger, *upon which this article draws in part, is also the editor of* Defending Standardized Testing *(Lawrence Erlbaum 2005) and the forthcoming* The Anti-Testing Fallacies *(APA Books 2007); and a member of the Third Education Group (www.thirdeducationgroup.org).*

Tests on Trial

Judging standardized tests against a benchmark of utopian perfection that does not and cannot exist means standardized tests always look bad. How would the criticisms look compared to the actual, available alternatives?

The Case Against Standards and Tests	The Testing Rebuttal	What's More . . .
Teaching to the Test		
Teachers will teach only material that will appear on a standardized test.	If high-stakes tests are kept behind lock and key until the day of test administration, teachers will not know what material will be on the test, except in the most general terms.	In the absence of common standards and tests, the curriculum becomes arbitrary and of uncertain origin. Why is that better than teaching to a required curriculum.
Narrowing the Curriculum		
A common curriculum prescribed by standards has less content than a teacher-made curriculum.	A school year's fixed number of hours and days renders it unlikely that a common curriculum will have less content than a teacher-arbitrary curriculum. I.e., a teacher who drops one topic when standards are introduced has necessarily added another.	What teachers and schools do in the classroom without common standards is not necessarily "broader." In fact, it can often be "narrower"—governed in the absence of other criteria by personal preferences.
Cheating by Students		
High-stakes standardized testing increases students' incentives to cheat.	Cheating is far easier to prevent and detect with standardized tests. Different forms used in the same classroom can make copying unrewarding. Computer programs run after the fact can look for telltale patterns.	Cheating in regular classroom work has become epidemic. The overwhelming majority of students admit to cheating in polls. Teachers and schools are ill equipped to monitor or detect most cheating. Meanwhile, the Internet makes cheating far easier than it used to be.

(continued)

Cheating by Teachers		
Many teachers have been caught cheating on high-stakes standardized tests.	The fact of detection may be evidence of how easily such cheating can be detected.	Social promotion and grade inflation provide evidence that nonstandardized testing and grading are far from infallible. And consider that in surveys, the majority of teachers claims overwhelming pressure to award high grades to undeserving students.
Preferred Instructional Methods		
Classrooms governed by standards are barren, dreary places where only factoids are learned.	A curriculum will always rely on some sort of standard or criteria for inclusion. The question is, Do we want formal, open standards, openly arrived at, or should their origins be more obscure?	Many teachers, especially inexperienced or quick hires in underperforming schools, will rely on the teacher's versions of basal textbooks for course content or, worse, make it up. Is the classroom shorn of standardized testing automatically a wonderful place—rich with innovative curriculum, the joy and magic of learning, and so on? What is the evidence?
Opposition to Norm-Referenced Tests		
Norm-referenced standardized tests are unfair. (I.e., it is unfair to simply rank kids, rather than measure them against standards.)	Norm-referenced tests provide information that cannot be obtained any other way. Many educators find them useful as measurement benchmarks and for curricular diagnosis.	The alternative, grade-point averages, are norm-referenced measures, normed at the school level.

Preference for Teacher-Made Classroom Testing		
Standardized tests are imposed from outside by persons and committees unfamiliar with and perhaps insensitive to the local students and community.	Standardized tests are developed by testing and measurement Ph.D.'s. The most capable measurement experts in the world work in North America developing standardized tests.	Teacher ed programs provide few teachers with more than cursory training in measurement, yet the absence of standardized testing would have us rely exclusively on their measurement decisions.

Notes

1. For a roundup of such polling results, see my "Persistently Positive: Forty Years of Public Opinion on Standardized Testing," chapter 1 in *Defending Standardized Testing,* ed. Richard P. Phelps (Mahwah, N.J.: Lawrence Erlbaum, 2005), 1–22.

2. The federal No Child Left Behind (NCLB) Act set in place what is largely a testing program. NCLB, however, falls far short of a comprehensive multi-level, multi-target high-stakes testing system; it sets only one achievement target (for schools), establishes no stakes for students (little motivation to take the test seriously), and provides curricular alignment that can be less than perfect.

3. An ambitious American student could simulate an equivalent program by taking several Advanced Placement (AP) examinations, except that she will graduate from high school and be accepted by some college no matter how she scores on them. Some states (e.g., Virginia, Michigan) are attempting to build something like this structure by requiring passage of a certain number of end-of-course examinations in high school.