

The Role and Importance of Standardized Testing in the World of Teaching and Training
Paper presented at the 15th Congress of the World Association for Educational Research
Cadi Ayyad University, Marrakesh, Morocco
June 3, 2008

Richard P. PHELPS

Standardized testing has been called the greatest single social contribution of modern psychology, and it may be the most useful evaluation method available for human resource-intensive endeavors. For most of their history, however, standardized tests have been developed and administered on a large scale and large, typically politically-sensitive organizations have controlled their use.

In the United States, standardized tests' political exposure has sometimes compromised their use despite the intrepid efforts of psychometricians to maintain their integrity. Some of you may recall the infamous "Lake Wobegon" scandal of the 1980s when a medical doctor, John J. Cannell, discovered that every U.S. state claimed an average student score on nationally-normed tests that was above the national average (Phelps, 2005b). Less well known, perhaps, are the persistent efforts of many powerful groups of professional educators to either eliminate the use of standardized tests or limit their use to the most unreliable types (Phelps, 2003).

With powerful forces opposed to the use (or to the proper use) of a beneficial technology that is typically provided by large, politically-sensitive organizations, perhaps it is time to consider alternative methods of providing that beneficial technology. One such alternative method is the topic of today's session.

Why standardized testing?

Standardized tests are not perfect evaluation tools. Used validly and reliably, however, standardized tests provide decision-makers useful information that no other evaluation method can provide.

Many research studies on educational testing dating back to the early part of the 19th century have compared different teachers' evaluations of identical student work or compared the consistency of teachers' marks to those of standardized test results over time. Not surprisingly,

researchers found wide variance from teacher to teacher in grading identical student work or over time with the same teacher.

In the 1910s, for example, researchers Starch and Elliott (1912) made copies of two actual English examination papers and sent them to teachers to grade and return. The marks ranged from 50 to 98 percent. One paper, graded by 142 teachers, received fourteen marks below 80 percent and fourteen above 94 percent. “That is, a paper which was considered too poor for a passing grade by some teachers was rated as excellent by others.”

Starch and Elliot repeated the procedure with duplicate Geometry tests (1913). Teachers’ marks on the 116 returned papers ranged from 28 to 92 percent, with twenty grades below 60 percent and nine of 85 percent and above. According to Lincoln and Workman (1936, 7):

This type of experiment has been repeated many times by investigators and always with similar results. Therefore there is abundant evidence that teachers’ marks are a very unreliable means of measurement.

Without standardized tests (or standardized grading protocols) in education, we would increase our reliance on individual teacher grading and testing. Are teacher evaluations free of standardized testing’s alleged failings? No. Individual teachers can *narrow the curriculum* to that which they prefer. Grades are susceptible to *inflation* with ordinary teachers, as students get to know a teacher better and learn his idiosyncrasies. A teacher’s (or school’s) grades and test scores are far less likely to be generalizable than any standardized tests’ (See, for example, Gullickson & Ellwein, 1985; Impara & Plake, 1996; Stiggins, Frisbee, & Griswold, 1989; Woodruff & Ziomek, 2004a, 2004b). (In Phelps, 2008, Table 1 lists some common fallacies proffered by testing opponents, along with citations to responsible refutations.)

According to the research on the topic, many U.S teachers consider “nearly everything” when assigning marks, including student class participation, perceived effort, progress over the period of the course, and comportment, according to one researcher. Actual achievement vis-à-vis the subject matter is just one factor. One study of teacher grading practices discovered that 66 percent of teachers felt that their perception of a student’s ability should be taken into consideration in awarding the final grade (Frary, Cross, & Weber 1993).

When individual teachers, or individual employers for that matter, are given the responsibility to make judgments unanchored by common standards or rules, those judgments tend to float freely in the currents of time, fitting first one context, then another, and then another. Being idiosyncratic to each particular, temporary context, each free-floating evaluation result is not generalizable to any permanent context. It is a judgment that makes sense only to a particular teacher or employer at a particular point in time and space.

When I was young, standardized tests were often called “objective tests,” which implied that teacher-made tests were “subjective.” Standardized tests’ clear separation from the influence of local decision-makers, be they classroom teachers or personnel managers responsible for hiring

new employees, remains one of their most beneficial features. The adoption of standardized university admission testing in the United States in the mid-twentieth century, for example, helped to pave the way for minorities who lacked the familial connections and social pedigree of wealthy WASPs (i.e., White, Anglo-Saxon Protestants).

According to Professor Stephen G. Sireci (2005, 113), the bad reputation of standardized tests portrayed by some critics “is an undeserved one.” He continues

People accuse standardized tests of being unfair, biased and discriminatory. Believe it or not, standardized tests are actually designed to promote test fairness. Standardized simply means that the test content is equivalent across administrations and that the conditions under which the test is administered are the same for all test takers. ...Standardized tests are used to provide objective information. For example, employment tests are used to avoid unethical hiring practices (e.g., nepotism, ethnic discrimination, etc.). If an assessment system uses tests that are not standardized, the system is likely to be unfair to many candidates.

There is more to subjectivity in decision-making than ethnic, racial, gender, or class bias, however. The fact is that true objectivity requires too much time to be practical in making everyday decisions. Double-blind controlled experiments or program evaluations with random assignment require time, money, and trained professional observation to monitor their progress. In our daily lives, we make judgments and decisions continuously. We cannot set up a controlled experiment, and wait for the results, every time we must choose which laundry detergent to purchase, where to go on vacation or, for that matter, whom to hire for a job or whom to admit to the last available place at university.

The time-saving decision-making technique we typically use to get on with our lives, apparently, is Bayesian reasoning, named for the early 18th-century statistician Thomas Bayes. In Bayesian reasoning, we employ what relevant prior knowledge we have to each decision. We calculate the “subjective probabilities,” which are not, in the strictest meaning of the term really “subjective.” More accurately, they are incomplete probabilities that incorporate the information we have accumulated that is relevant to the matter at hand. That information may be reliable or not, verified or not, true or not. Nonetheless, until we discover a Fountain of Youth to provide us everlasting life, we must rely on Bayesian reasoning as a time-saving heuristic to negotiate our lives in the short time allotted to each of us (“Bayes Rules,” 2006).

Thus, a standardized test is more than an antidote to biased judgment. We need standardized tests because each of us is a prisoner of our own limited experiences and observations. Standardized tests provide an opportunity to make decisions about individuals that are free of subjectivity, be that subjectivity due to bias or Bayesian shortcuts. In developing standardized tests, trained professionals collect empirical data, apply statistical benchmarks, and make detached, objective evaluations.

Standardized testing: The long view

Standardized tests have provided information for making important decisions at least since the first administration of the Chinese civil service examination many centuries ago (Zeng, 1999, 8). The “scientific” standardized test (with statistically-calibrated score scales), however, is just a century old (Phelps, 2007b, chapter 2). The innovators responsible for the development of the scientific standardized test—e.g., Binet, Simon, Rice, Thorndike—though, likely would be amazed by the improvements made in testing technology within the relatively brief period since—e.g., computer-adaptive testing or open-source, Web-based platforms, such as the Examination Assessment Management System (ExAMS).

It would seem that testing technology has improved over time exponentially. Test developers have increased the complexity and technical sophistication of their product in response to market and regulatory demands. Today’s standardized tests are better in most every way than their progenitors. They provide more information for the price, and they are more reliable, fair, and valid (when used as they are designed to be used).

But, the exponential rate of improvement carries some risk. At the same time standardized tests have improved in quality and convenience, they have become more difficult for the average person or policymaker to understand. Most standardized tests administered a century ago were simply larger-scale, standardized versions of an ordinary classroom teacher’s examination. In all apparent aspects, they looked familiar to the average examinee.

Some of today’s standardized tests might seem to the average citizen or policymaker as different in character from their 100-year-old ancestors as today’s airplanes or automobiles do from their 100-year-old antecedents. Any of you who have tried in plain language to explain to policy makers the concepts of item response theory, differential item functioning, computer-adaptive testing, or point-biserial correlation will know what I mean.

The combination of technical complexity and the widespread use of testing for public purposes should elicit a clear, measured, and open public discussion on testing policy. And, I hope that it does where you live. In the United States, unfortunately, the public and policymakers are generally showered with obfuscation, misinformation, and disinformation.

The testing policy debate in the United States: The sound of one hand clapping

Standardized testing in the United States is an enigma. Arguably, the country hosts much of the world’s most advanced technical research and innovation. Yet, debates on testing policy remain primitive and one-sided.

The late economist Mancur Olsen (1965, 1982) developed a theory to explain the political power of “special interests” in democratic societies. Individuals join groups that provide private benefits, such as protection against market competition, disruptive technologies, or other challenges to the familiarity and security of the *status quo* like those portended by externally-imposed evaluations of performance, such as standardized tests. While the benefits to members

of the group (e.g., a professional association of educators) can be large (e.g., the absence of standardized testing programs) the costs (e.g., lowered student achievement, a less efficient education or employment system) tend to be diffused over society at large and may not even be noticed by those who bear them. Special interests accrete more and more private benefits (and political power) over time, however, until they become “vested” interests—wealthy, powerful, and entrenched.

Olson’s theory is particularly applicable to education in the United States, because its governance is so widely dispersed. Each of the 50 states is constitutionally responsible for public education and, in 49 states, some governance and taxing authority is further deferred to local school districts, which are typically governed separately from other local units of government. Some national associations of educators maintain substantial memberships in each and every local school district, state legislative district, U.S. congressional district, and television, radio, and newspaper media market. They can saturate the country with the policy-related information they prefer and block out the information dissemination efforts of less powerful individuals or groups that offer contrary points of view.

In the United States, society’s understanding of standardized testing may be shrinking. The technical psychometric research literature would seem to be safe. But, the research literature related to testing *policy* (i.e., its administration, program structure, use, extent, effects, cost, benefits, public opinion, research dissemination) is diminishing. There are simply too few who cite the research literature in any substantial depth or breadth, and too many willing to declare it barren.

The most common debating tactic of testing opponents is to avoid debate (Phelps, 2007a). Whereas scientists seek the scrutiny of their peers in order to confirm (or deny) the value of their work, *advocates* tend to avoid scrutiny, especially when selling falsehoods. Scientists do not circumvent the research literature, but engage it. They respond to rival hypotheses with counterevidence. They confront conflicting scientific results. Advocates, however, simply ignore them. The easiest way to win a debate is by not inviting an opponent. Testing critics rightly fear an open, fair scientific contest.

Indeed, it has become quite common for testing opponents to declare nonexistent an enormous research literature that contradicts their claims. With the help of the fourth estate (Lieberman, 2007, chapter 11), they have been fairly successful in eradicating from the collective memory thousands of studies conducted by earnest researchers over the course of a century.

In one effort of mine—accumulating studies on the effects of standardized testing—I started out thinking that there were a dozen or so. A few years ago I knew that there were hundreds. Now I know that their number exceeds a thousand. (In Phelps, 2008, Table 2 provides a brief synopsis of the research literature.)

In the end, however, it will not matter for society's sake if we find ten thousand studies. There will remain other education researchers, prominent and with hugely abundant resources at their disposal—researchers whose work is frequently covered by U.S. education journalists—who will continue to insist that *no* such studies ever existed. It is U.S. education research's dirty big secret: research that generates results that are unpopular among the vested interests can be successfully—and easily—censored and suppressed (see, for example, Phelps, 1999; 2000; 2003, Preface & chapter 7; 2005a, chapter 3).

Wildlife conservationists tell us that a biological species cannot survive when mating individuals cannot find each other. When numbers decline to such an extent that predators (or hunters) can more easily find members of the species than can potential mates, the species crosses a demographic threshold and heads toward its inevitable extinction. Those who work with endangered species call this the “extinction vortex.”

Similarly, the censorship and suppression of the research literature on the effects of educational achievement testing has become so successful that it has become difficult to find its progenitors. For example, I may have spent more time than anyone combing the research literature. Nonetheless, I was a few years into my effort before I discovered the work of Frank Dempster (1991, 1997), one of the world's foremost authorities, or that of Jim Haynie who works in career and technical education. Why did it take me so long to find their work? Their work is not popular among the vested interests in education—they find the benefits of testing to be strong and persistent—thus it is not widely advertised.

One hundred years of research and experience left behind

Indeed, the No Child Left Behind (NCLB) Act, passed by the U.S. Congress in 2002, could have been informed by a cornucopia of research and experience. Instead, it was informed by virtually none. Prior research and experience would have told policymakers that most of the motivational benefits of standardized tests required consequences for the students and not just for the schools. Those stakes needn't be very high to be effective, but there must be some. As NCLB imposes stakes on schools, but not on students, who knows if the students even try to perform well.

Prior research and experience would have informed policymakers that educators are intelligent people who respond to incentives, and who will game a system if they are given an opportunity to do so (see, for example, Cannell, 1987, 1989). The NCLB Act left many aspects of the test administration process that profoundly affect scores (e.g., incentives and motivation, cut scores, degree of curricular alignment) up for grabs and open to manipulation by local and state officials.

Prior research and experience would have informed policymakers that different tests get different results and one should not expect average scores from different tests to rise and fall in unison over time (as some interpreters of the NCLB Act seem to expect with the National Assessment of Educational Progress [NAEP] benchmark) (Phelps, 2005b).

Prior research and experience would have informed policymakers that the public was *not* in favor of punishing poorly-performing schools (as NCLB does), but *was* in favor of applying consequences to poorly-performing students and teachers (which NCLB does not) (see, for example, Phelps 2005a, chapter 1).

What are the effects of test-based accountability? Table 3 in chapter 3 of the forthcoming *Correcting fallacies about educational and psychological testing* (Phelps, 2008) lists just a small sample of useful, insightful, relevant studies that effectively answered this question, could have informed the design of NCLB, and have been declared by prominent educators to not exist.

Had the policymakers and planners involved in designing the NCLB Act simply read the freely-available research literature instead of funding expensive new studies and waiting for their few results, they would have received more value for their money, gotten more and better information, and gotten it earlier when they actually needed it.

With the single exception of the federal mandate, there was no aspect of the NCLB accountability initiative that had not been tried and studied before. Every one of the NCLB Act's failings was perfectly predictable, based on decades of prior experience and research. Moreover, there were better alternatives for every characteristic of the program that had also been tried and studied thoroughly by researchers in psychology, education, and program evaluation. Yet, policymakers were made aware of none of them.

The resulting scantily-informed public policy includes a national testing program that would hardly be recognizable anywhere outside of North America. The standardized testing component of NCLB includes no consequences for the students. This sends the subliminal message to the students that they need not work very hard and the testing's largest potential benefit—motivation—is not even accrued.

By contrast, schools are held accountable for students' test performance; they are held responsible for the behavior of other human beings over whom they have little control. Moreover, the most important potential supporters of testing programs—classroom teachers and school administrators—are alienated, put into the demeaning position of cajoling students to cooperate.

Taking testing directly to the people

I interpret the highly successful censorship and suppression of a century's research literature on the effects of standardized testing to be evidence that the vested interests in U.S. education now control the testing policy debate. This is lamentable; but what does it have to do with today's topic?

Quite a lot, as it turns out. When the forces of censorship and suppression gain effective control of the main routes of information dissemination, the only way for others to reach the

public is via alternative routes. Where entrenched interests attempt to impede communication between testing producers and testing consumers, it is only natural that the two interested parties should try to communicate directly. And, that is where the innovation of Web-based open-source platforms fits in.

Hopefully, Web-based open-source testing platforms will not only facilitate the spread of high-quality testing use but also its understanding. As more and more test users learn how to use the technology they simultaneously become better informed citizens not only about a technology but about public policies related to testing.

References

- Bayes Rules. (2006, January 5). *The Economist*. Retrieved April 27, 2008 from http://www.economist.com/science/displaystory.cfm?story_id=E1_VPVQJG
- Cannell, J.J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all fifty state are above the national average*. (2nd Ed.), Daniels, WV, USA: Friends for Education.
- Cannell, J.J. (1989). *How public educators cheat on standardized achievement tests*. Albuquerque, NM, USA: Friends for Education.
- Dempster, F. N. (1991, April). Synthesis of research on reviews and tests, *Educational Leadership*, 71–76.
- Dempster, F. N. (1997). Using tests to promote classroom learning. (pp. 332–346). In R. F. Dillon, (Ed.). *Handbook on testing*. Westport, CT, USA: Greenwood Press.
- Frary, R. B., Cross, L. H., & Weber, L. J. (1993). Testing and grading practices and opinions of secondary school teachers of academic subjects: Implications for instruction in measurement, *Educational Measurement: Issues and Practice*, 12(3), 23+.
- Gullickson, A.R. & Ellwein, M. C. (1985). Post-hoc analysis of teacher-made tests: The goodness of fit between prescription and practice. *Educational Measurement: Issues and Practice*, 4(1), 15–18.
- Impara, J. C. & Plake, B. S. (1996). Professional development in student assessment for educational administrators. *Educational Measurement: Issues and Practice*, 15(2), 14–20.
- Lieberman, M. (2007). *The educational morass*. Lanham, MD, USA: Rowman & Littlefield.
- Lincoln, E. A., & Workman, L. L. (1936). *Testing and the uses of test results*. New York, NY, USA: Macmillan.
- Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*, Cambridge, MA, USA: Harvard University Press.
- Olson, M. (1982). *The rise and decline of nations: Economic growth, stagflation, and social rigidities*, New Haven, CT, USA: Yale University Press.
- Phelps, R. P., (1999, April). Education establishment bias? A look at the National Research Council's critique of test utility studies, *The Industrial-Organizational Psychologist*, 36(4), 37–49.
- Phelps, R. P. (2000, Winter). Estimating the cost of systemwide student testing in the United States. *Journal of Education Finance*, 343–380.

- Phelps, R. P. (2003). *Kill the messenger: The war on standardized testing*. New Brunswick, NJ, USA: Transaction Publishers.
- Phelps, R. P., Ed. (2005a). *Defending standardized testing*. Mahwah, NJ, USA: Lawrence Erlbaum.
- Phelps, R.P. (2005b). The source of Lake Wobegon. *Nonpartisan Education Review / Articles*, 1(2). Retrievable at <http://nonpartisaneducation.org/Review/Articles/v1n2.htm>
- Phelps, R.P. (2007a). The dissolution of education knowledge. *Educational Horizons*, 85(4), 232–247.
- Phelps, R.P. (2007b). *Standardized testing primer*. New York, NY, USA: Peter Lang.
- Phelps, R.P. (2008). Educational achievement testing fallacies, Chapter 3 in R.P. Phelps (Ed.), *Correcting fallacies about educational and psychological testing*. Washington, DC, USA: American Psychological Association.
- Sireci, S. G. (2005). The most frequently unasked questions about testing. In R. P. Phelps (Ed.). *Defending standardized testing* (111–122). Mahwah, NJ, USA: Lawrence Erlbaum.
- Starch, D., & Elliot, E. C. (1912). Reliability of the grading of high school work in English. *School Review*, 21, 442–457.
- Starch, D., & Elliot, E. C. (1913). Reliability of grading work in mathematics. *School Review*, 22, 254–259.
- Stiggins, R. J., Frisbee, D. A. & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda, *Educational Measurement: Issues and Practice*. 8(2), 5–14.
- Woodruff, D. J., & Ziomek, R. L. (2004a, March). *Differential grading standards among high schools*. ACT Research Report 2004-2, Iowa City, IA, USA: ACT.
- Woodruff, D. J., & Ziomek, R. L. (2004b, March). *High School Grade Inflation from 1991 to 2003*. ACT Research Report 2004-4, Iowa City, IA, USA: ACT.
- Zeng, K. (1999). *Dragon gate: Competitive examinations and their consequences*. London, UK: Cassell.