TESTING . . . 2, 3, 4, . . .
11, 12, 12_+; G_-; G_+
CAP, CAT, CTBS, . . . , ITBS, . . .
ACT, SAT; GRE, LSAT, MCAT; NTE

Cathy Kessel*

60604(a) The Superintendent of Public Instruction shall design and implement . . . a statewide pupil assessment program. . . . That program shall include all of the following:

> (1) A plan for producing valid, reliable, and comparable individual pupil scores in grades 2 to 11, inclusive, and a comprehensive analysis of these scores . . . pursuant to the Standardized Testing and Reporting (STAR) Program. . . . (California Education Code)

Standardized examinations before age 16 *have all but disappeared from the EC countries*. (Feuer & Fulton, 1994, p. 36, authors' emphasis)

[S]tandardized tests play nowhere near the role in Japanese elementary schooling that they do in the United States, where school funding, real-estate prices, and legal sanctions may all hinge on standardized test scores. (Lewis, 1995, pp. 200–201)

---

* An earlier version of this article appeared in the summer 1999 issue of the *Mathematicians and Educational Reform Newsletter*. The Appendix was created in 2006.

The United States is unique in the extent of its use of standardized tests for young children (Office of Technology Assessment [OTA], 1992, p. 31), and testing is on the increase. Between 1960 and 1989, revenues from standardized tests for grades K–12 (in constant dollars) increased about 150% while student enrollment increased about 18%. State spending on tests is projected to increase from $165 million in 1996 to $330 million in 2000 (Fox, 1999). Yet, the rest of the world does not find such extensive testing necessary.[1] History and culture help to explain why the U.S. tests more. Educational research answers, "Why not less?" These answers to "Why?" and "Why not?" suggest that a more important question is not the amount of testing, but the fit between test, school, teachers, students, and curriculum.

Why? (An historical explanation)

Testing in the United States began about 150 years ago as a means of monitoring the school system. The United States was among the first countries to promote extensive public education. Public schooling began under local control, and has continued that way. With local control arose the issue of fairness: Were all students from urban school to rural one-room schoolhouse receiving comparable educations? One way to find out was to test the students. It seemed that the fairest way to do so was to give the same tests to all students at each grade level and grade those tests in the same way—standardized testing.

Changes in industry suggested a different way to think about schooling and added another reason for testing. The efficiency movement began in the late 1800s with Frederick

---

[1]The extent of testing in the U. S. is sometimes explained by the high percentage of students in school. However, the United States has recently lost its status as the country with the highest secondary enrollment rate (see the Organization of Economic Cooperation and Development, www.oecd.org/els/edu/EAG98/index.htm).

Taylor's time-motion studies in industry. Under the assumption that what worked for industry ought to work for education (a line of thought that continues today), Taylorism was applied to schools. The metaphor was that schools were factories, and that knowledge was a product to be acquired by students and delivered piecemeal by teachers. Students needed to be sorted according to grade and "ability" so that teachers could more efficiently deliver instruction tailored to their needs—and tests were the mechanisms by which this sorting was done.[2] Testing itself became more efficient with the advent of multiple choice tests in 1915. By 1930 they were firmly entrenched in schools (OTA, 1992, pp. 117, 124).

Taylorism may have been considered particularly well-suited to schools for another reason. Rather than being a profession as it was in Europe, teaching in the United States was a low-status job with high turnover. Teachers' wages were equivalent to those of unskilled laborers (Conway, 1985). Sorting by means of tests devised by psychometricians may have been an effort to make the classroom into an assembly line for workers who were not expected to stay long enough to learn how to make their own educational judgments.[3]

Research in psychology suggested that tests could sort students by "aptitude" and "ability." In 1904, the French minister of education commissioned the psychologist Alfred Binet to create tests that distinguished between public school children who were "normal" and those who were in need of special education. Binet found that a key relationship was

---

[2]A host of assumptions about teaching and learning underlie this metaphor, for instance the assumption that complex skills can be broken into simple ones and taught piecemeal (ee e.g., OTA, 1992, pp. 45–48, 51–53).

[3]This division of labor continues. Test development is relegated to the private sector, and U.S. teachers, unlike those in other countries, are rarely responsible for scoring external exams (Feuer & Fulton, 1994).

the age at which a child could perform various tasks on what came to be called the Binet scales. After Binet's death in 1911, his two-valued "normal" and "needing special education" was transformed into the multi-valued "IQ" (100 times mental age as determined by testing, divided by chronological age). Binet's view of intelligence as malleable and subject to environmental influences was ignored in the United States, but his scales were transformed into tests like the Stanford-Binet (developed by Stanford psychologist Lewis Terman), first used to diagnose children in need of special education, later for system-wide tracking or grouping by ability (OTA, p. 118).[4] The use of IQ and aptitude tests in schools continues. The best-known aptitude test is perhaps the SAT. It began life in 1926 as a college examination (the Scholastic Aptitude Test) modeled on the World War I Army Alpha intelligence tests developed by Terman and his colleagues, and has since changed its name to SAT[5] and acquired a variety of other uses.

Students could also be sorted according to "achievement"—how much they had learned or "value added to the raw material . . . during the course of the year" (OTA, 1992, p. 110). Achievement tests were different from IQ and aptitude tests, they were supposed to measure specific knowledge as determined by test designers and used to determine student placement, promotion, retention, or graduation.

Why: Some cultural reasons

Other countries have not found school monitoring and various forms of student sorting necessary or have fulfilled those functions in other ways. For example, ability grouping is

---

[4]For example, a 1925 survey found that 90% of elementary schools and 65% of high schools grouped students by ability (OTA, 1992, p. 122).

[5]A College Board statement says, "Please note that SAT is not an initialism; it does not stand for anything" (Applebome, 1997).

not used in Japan, so no sorting by "ability" is necessary. Students are rarely skipped ahead or retained in grade,[6] and measures other than teachers' tests and observations are not considered necessary for such decisions (Lewis, 1995, p. 15). Instead, another kind of sorting occurs. Students take an exam at the end of compulsory schooling (ninth grade) that determines the upper secondary school each will attend.[7]

Countries that consider school monitoring necessary sometimes use school inspections rather than exams.[8] And when exams are used, they are not used at every grade level. For example, the 1988 reform in England called for external testing at four grade levels—making British students the most tested in Europe (OTA, 1992, p. 161). Monitoring may also occur indirectly, for instance via the "lesson study" of Chinese and Japanese teachers in which groups of teachers design, teach, observe, discuss, and revise particular lessons (Lewis & Tsuchida, 1998). In these countries teaching professionalism replaces outside monitoring. No such activity occurs in the United States.

History and culture suggest why standardized testing is endemic within the United States. It may also be an epidemic. Educational research documents a variety of reasons:

---

[6]Much of Europe does not retain students in grade. Danish educators have said they consider it a barbaric practice, something that would be done only by a primitive culture that didn't really like its children (Bracey, 1999).

[7]There's no evidence that the pressure of exams causes a large percentage of youth suicides in Japan as is often assumed. Moreover, suicide rates are considerably lower in Japan than in the U. S. In the early 1990s the number of suicides per 100,000 for 15- to 19-year-olds was 3.8 in Japan and 11.1 in the United States. For 20- to 24-year-olds, it was 10.4 in Japan and 15.1 in the United States (Lewis, 1995).

[8]Some private schools in the United States also use inspections rather than exams for monitoring.

deleterious side effects on students and teachers, the context of testing, and misinterpretation and misuse of test scores.

Why not: Effects on students and teachers

Stanley Erlwanger studied an extreme case of instruction tailored (or Taylorized) to individuals by testing. In 1973 he published an article about a sixth grader who he called Benny. Since second grade Benny had been enrolled in self-paced mathematics courses in which Individually Prescribed Instruction (IPI) course materials were used by students independently of their teacher. The goal of IPI was to be "maximally adaptive to the individual." Student progress was monitored by multiple choice tests administered by a teacher's aide, thus freeing the teacher to help individual students. Benny was succeeding in this system. In sixth grade his IPI test scores were 80% or higher, and his teacher considered him one of her best pupils.

But Erlwanger's interviews with Benny suggested that Benny's mathematical knowledge had some serious flaws. Benny was aware that although an answer could be expressed in different ways, the IPI test format only allowed one of those ways to be marked as correct. For example, he said that $\frac{1}{2}$ is the same as $\frac{2}{4}$, but that if he answered $\frac{2}{4}$, his answer would be marked wrong because the aide and teacher "have to go by the [answer] key . . . what the key says . . . I don't care what the key says." Because of Benny's disregard for the answer key, wrong answers did not constrain his views of fractions and decimals—he distinguished what was "right" from what was "acceptable" according to the answer key. Indicating that different answers were perfectly acceptable to him, Benny said:

> If I ever had this one [i.e. 2 + .8] . . . actually, if I put $2\frac{8}{10}$, I get it wrong. . . . if I had this example [i.e. $2\frac{8}{10}$] and I put 1.0, I get it wrong. But really they're the same, no matter what the key says. . . .
>
> If I did 2 + .3, that will give me a decimal; that will be .5. If I did it in pictures [i.e. physical models] that will give me 2.3. If I did it in fractions like this [2 + $\frac{3}{10}$], that will give me $2\frac{3}{10}$. (Erlwanger, 1973, p. 15)

Erlwanger summarized some of Benny's views about mathematics:

> He regards operations as merely rules; for example, to add 2 + .8, he says: "I look at it like this: 2 + 8 is 10; put my 10 down; put my decimal in front of the zero." However, rules are necessary in mathematics, "because if all we did was to put any answer down, [we would get] 100 every time. We must have rules to get the answer right." (p. 17)

The practice of continually testing students may have deleterious effects on their beliefs about tests. Paris et al. (1991) studied students in grades 2 to 11. Elementary students tended to agree that "Test scores show how intelligent you are" and "Test scores help identify which teachers do the best job," but the contrary held for older students who were also *less* likely to agree that "I gave my best effort on the test we took" and "I have good strategies for taking tests" and *more* likely to report they cheated, got nervous, had difficulty concentrating, or guessed on tests.

Testing may also have demoralizing effects on teachers. Smith (1991) studied a group of elementary teachers whose students' progress was judged by their performance on the Iowa Test of Basic Skills (ITBS). Some of the students exhibited emotional distress

during the exam by crying, fighting, giving up, marking answer sheets randomly, vomiting, etc. Including time for preparation (drilling on test preparation materials), administration, and allowing students to recover, tests cost about 100 hours of instructional time per class. The teachers believed that the test scores were worthless because of the mismatch between their curriculum and the test, the psychometric inadequacies of the test, and the emotional status of their students at the time of the test.

Why Not: Problems of Administration, Interpretation, and Use

Tests need to be carefully administered. For example, the research of Claude Steele and his colleagues (e.g., Steele, 1997) has shown that members of groups such as females or African Americans who are sometimes stereotyped as not being able mathematically, may be at a disadvantage if their consciousness of their female or African American status is triggered before or during a mathematics test. For example, in one of Steele's experiments, the subjects, undergraduates with good calculus grades, were given examinations consisting of mathematics questions from the GRE. The experimenters told one group of subjects that they expected gender differences in scores, but did not mention these "expectations" to a control group. The results: For easy exams the treatment had no effect. Average scores for males and for females were the same in both experimental and control groups. For difficult exams the experimental group had a gap in the average scores of males and females that favored males.

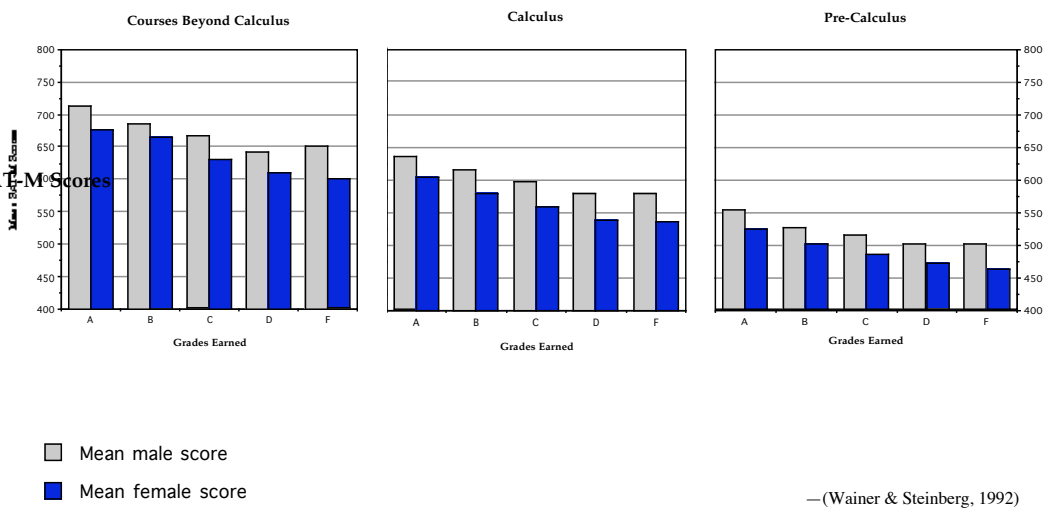
Test scores also need to be carefully interpreted. In 1987, for example, *no state* scored below the national norm on nationally normed elementary tests such as the ITBS (Cannell, 1988)! This situation was dubbed the Lake Wobegone effect, after Garrison Keillor's fictional small town in Minnesota where "all the women are strong, all the men are good-looking, and all the children are above average." Several explanations have been given for the Lake Wobegone effect: poor test security, "teaching to the test," and

inaccurate or out-of-date test norms. Tests may be not only intrusive, but also not serve their intended purpose.

Scores should be appropriately used. The SAT has been designed and validated for use with high school grades in prediction of first-year college grades, and no other uses are officially sanctioned. However, the SAT, like other standardized tests, has acquired a variety of other uses. These have included placing students in college mathematics courses, determining fellowship awards, measuring "mathematical ability" in educational research, and measuring educational progress via the SAT wall charts.

Figure 1 displays statistics (from Wainer and Steinberg, 1992) that suggest why using SAT scores for course placement might not be a good idea. Fortunately, the wall charts showing SAT score averages by state have been discontinued. Two major reasons why this use was inappropriate are that percentages of SAT-takers in different states vary considerably and the SAT is not designed as an achievement test. Unfortunately, some journalists still consider average state SAT scores as a good indicator of educational progress or regress.

Mean SAT-M scores for males and females earning the same grade in college mathematics courses



—(Wainer & Steinberg, 1992)

Conclusion

Standardized tests are not straightforward either in their results or in their effects. They upset students, alienate teachers, cost taxpayers money, cost students and teachers time for testing, and sometimes considerable time for test preparation—sometimes weeks of preparation unrelated to ongoing curriculum, but are seen as necessary because of the "high stakes" nature of the tests. Standardized tests may convey misleading views of mathematics and their results can be misused or misinterpreted. For all these reasons, decisions to test, selection and administration of tests, and interpretation of test results should be undertaken with caution.

A larger question is the role of standardized tests in the educational system of the United States. Attempts to cure educational ills with even more tests may simply create iatrogenic diseases in the U. S. educational system. The experience of other countries suggests that the functions for which the U.S. prescribes tests—system monitoring, student sorting, and student certification—might be accomplished by a reorganization of schooling in which tests—carefully compounded and taken in moderation—are just what the doctor ordered.

Acknowledgements

My thanks to Iben Maj Christiansen, Maryl Gearhart, Ilana Horn, Betty Levitin, Alan Schoenfeld, Miriam Gamoran Sherin, and Natasha Speer for their comments on previous versions of this article, and to Dawn Davidson for the production of the graphic.

References

Applebome, Peter. (April 2 1997). Insisting it's nothing, creator says SAT, not S. A. T. *New York Times, 146*, A16.

Bracey, Gerald. (1999). Retention in grade fails children. *Education Week, 18*(40), 42.

Cannell, John J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average. *Educational Measurement: Issues and Practice 7*(2), *5–9*.

Conway, Jill Ker. (1985). Politics, pedagogy, and gender. In Jill K. Conway, Susan C. Bourque, & Joan W. Scott (Eds.), *Learning about women: Gender, politics, and power* (pp. 137–152). Ann Arbor: University of Michigan Press.

Erlwanger, Stanley H. (1973). Benny's conception of rules and answers in IPI [Individually Prescribed Instruction] mathematics. *Journal of Children's Mathematical Behavior 1*(2), 7–26.

Feuer, Michael J. & Fulton, Kathleen. (1994). Educational testing abroad and lessons for the United States. *Educational Measurement: Issues and Practice, 13*, 31–39.

Fox, Jonathan. (1999). Grilling our young. *Salon*, http://www.salon.com/mwt/feature/1999/11/08/testing/print.html salon.com.

Lewis, Catherine C. (1995). *Educating hearts and minds: Reflections on Japanese preschool and elementary education*. Cambridge: Cambridge University Press.

Lewis, Catherine C., & Tsuchida, Ineko. (1998). A lesson is like a swifly flowing river. *American Educator 21*(3), 12, 14-17, 50-52.

Paris, Scott, Lawton, Theresa, Turner, Julianne, & Roth, Jodie. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher 20*(5) 12–20.

Smith, Mary L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher 20*(5), 8–11.

Steele, Claude M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist 52*(6), 613-629.

U.S. Congress, Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions*, OTA-SET-519. Washington, DC: U.S. Government Printing Office.

Wainer, Howard & Steinberg, Linda. (1992). Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. *Harvard Educational Review 62*(3), 323-336.

Appendix

Added in 2006, not thoroughly checked yet!

External Mathematics Tests in France

Test	When taken	Format
	Compulsory	
Diagnostic	beginning grade 3	multiple choice, constructed response
Diagnostic	beginning grade 6	multiple choice, constructed response
Diagnostic	beginning grade 7	multiple choice, constructed response
	Non-compulsory	
Brevet des Collèges	end middle school	
Certificat d'Aptitude Professionnelle	end two-year high school vocational program	
Brevet d'Enseignement Professionnel	end two-year high school vocational program	
Baccalauréat	end four-year high school program	constructed response

Note. Except for students with disabilities, all students follow the same curriculum until high school.

External Mathematics Tests in California

Test	When taken	Format
	Compulsory	
STAR	end grade 2	multiple choice
STAR	end grade 3	multiple choice
STAR	end grade 4	multiple choice
STAR	end grade 5	multiple choice
STAR	end grade 6	multiple choice
STAR	end grade 7	multiple choice
STAR	end grade 8	multiple choice
STAR	end grade 9	multiple choice
STAR	end grade 10	multiple choice
STAR	end grade 11	multiple choice
CAHSEE	high school	multiple choice
	Non-compulsory	
NAEP	grade 4, stratified sample	multiple choice, constructed response
NAEP	grade 8, stratified sample	multiple choice, constructed response
NAEP	grade 12, stratified sample	multiple choice, constructed response
SAT or ACT	end of high school, college-intending students	multiple choice
SAT subject tests	end of high school, college-intending students	multiple choice
Advanced Placement tests	end of high school, college-intending students	multiple choice, constructed response

Note. The CAHSEE (a high school exit exam) was put into effect in 2006. Students who do not pass this test are not considered to have graduated from high school. STAR is the Standardized Testing and Reporting program mandated by the California Education Code.