Are U.S. Students the Most Heavily Tested on Earth?

Richard P. Phelps
American Institutes for Research, Pelavin Research Center, Washington, DC

How should amount of testing be defined? On average, how many hours does a U.S. student spend on testing? How does this compare with testing time in other countries? How do the type and purpose of testing vary from U.S. to other countries?

I first heard the assertion that U.S. students are the most heavily tested in the world from George Madaus at a 1991 AERA-sponsored panel session in Washington, DC. Five opponents of a proposed national examination system were stating their arguments against such a system and an alleged, relatively large current amount of U.S. testing supported one of Madaus's arguments.[1] Then, I read the claim, written by Monte Neill of the advocacy group FairTest, that the U.S. ranks first in its amount of standardized testing.[2] I have heard that others have made similar claims that U.S. students encounter relatively more testing, standardized testing, norm-referenced testing, or some other kind of testing than students in other countries. But, I have not seen empirical evidence associated with these claims or heard reference to a data source that would support these assertions.

Whether or not U.S. students are the most heavily tested in the world is an important policy issue. This assertion has been used as an argument against the adoption of a national examination system. It could also be used as an argument against any increase in testing.

The purpose of this article is twofold. First, I attempt to determine whether or not the assertion that U.S. students are the most heavily tested in the world can be verified *with the data available from several large-scale national and international surveys.* Second, I examine the data in order to compare the character of U.S. testing to that of student testing in other countries.

The data will be used to answer these questions:

(a) How much systemwide testing[3] is there in the United States, and what is its general character?

(b) How much systemwide testing is there in other countries, and what is its general character?

(c) How does systemwide testing in the U.S. compare to that in other countries?

(d) How do countries compare in their amounts of classroom testing?

Data Sources

Data for this study were compiled from four survey studies:

(a) The U.S. Congress' General Accounting Office (GAO) conducted a survey of state and local school district testing practices in the academic year 1990–1991;

(b) In 1990–1991, the Organization for Economic Cooperation and Development (OECD) studied testing practices in many advanced industrial countries, using a survey developed by James Guthrie, Marilyn Binkley, and Gary Phillips, under the auspices of the U.S. Education Department's National Center for Education Statistics (NCES);

(c) The International Association for the Evaluation of Educational Achievement (IEA) administered questionnaires about classroom practices in conjunction with the International Reading Literacy Study in 1991 and with the Second International Mathematics and Science Study (SIMSS) in 1982; and

(d) The International Assessment of Educational Progress (IAEP) administered questionnaires about classroom practices in conjunction with its mathematics and science tests in 1991.

The OECD study provides information from many European countries and New Zealand, while the GAO study provides information from the U.S. states and local school districts. Both of these studies referred to the school year ending 1991. The OECD survey, mailed out in the Fall of 1990, asked about tests in use that academic year, 1990–1991,[4] while the GAO survey, completed in the Fall of 1991, asked retrospectively about tests given in the same (1990–1991) school year.

Two international tests were also administered during that same academic year, in the Spring of 1991. The IAEP tested 13-year-old students in math and science, and it questioned them about the frequency of tests or quizzes in their mathematics and science courses. The IEA questioned teachers in classrooms participating in the International Reading Literacy Study about their frequency of classroom testing with multiple-choice or other formats. Together, these surveys

Richard P. Phelps is Senior Research Analyst, American Institutes for Research, Pelavin Research Center, 1000 Thomas Jefferson St., NW, Suite 400, Washington, DC 20007. His specializations are finance and international education indicators.

provide a picture of the relative frequency of classroom testing across countries in 1991. The IEA posed similar questions to teachers in the 1982 SIMSS.

Just as there exist a variety of test types, *counting* tests can be done in a variety of ways. Here, I will count systemwide (or standardized) tests in two general ways: by their number and their duration.

Counting tests by their number can, in turn, be done using three methods.

(a) Simple number of tests: An organization develops a test, gives it a name, and it is administered in the schools. The test may be administered one or more times in a year and in one or more grade levels, but it is still identified as a single test if its character, purpose, design, and format are constant even while the curricular content may vary across different grade-level versions.

(b) Grade levels tested: The absolute number of grade levels in which a systemwide test is administered.

(c) Number of separate test administrations (over the course of the average student's primary–secondary school career).

To understand how the different counting methods work, consider the testing program of my own local school district, which administers the CTBS once a year, in 10 grades, and no other systemwide test. The simple number of tests is one, the number of grade levels tested is 10, and the number of separate test administrations that the average student will encounter is 10 (over 10 years). Some readers might question whether the CTBS should be counted as just one test, given that it is administered in 10 grades and no two grade levels see exactly the same version of the test. This would be an understandable reaction. Nonetheless, it is true that many educators would refer to the CTBS's administration in my school district as just one test. Those who filled out the OECD and GAO questionnaires followed this same pattern of counting.

One can also count tests by their duration, counting the length of time allotted to complete each of them, then summing their durations over the course of a year or over the course of a student's career in school. Personally, I believe counting tests by their duration to be the superior method because I believe there to be a correlation between test duration and test importance and effect.

Furthermore, I believe duration to be the more reliable measure, because the boundaries of a test unit are somewhat ambiguous. For example, a school district that administers the Metropolitan Achievement Test (MAT), including a multiple-choice achievement battery over five subject areas and an open-ended written composition, could be said to administer one test (the MAT), two tests (multiple-choice and open-ended), or, in the most extreme interpretation, six tests (in six subject areas). But, the duration would always be the same, whether it was counted as the length of time for one test or as the cumulative duration of two to six tests.

How Much Are Students Tested in the United States?

In the Summer of 1991, the country was debating the proposition that the United States adopt a national examination for elementary and secondary school students. Several proposals with some measure of detail were proposed by various policy-oriented groups. Early in the debate over national testing, however, decision makers saw that they lacked some key information. What was the current extent and cost (in both time and dollars) of testing in the schools, and how much would a national examination cost? To obtain its own estimates, the Congress asked the GAO to examine the current extent and cost of testing in the United States (see U.S. General Accounting Office, 1993).

The GAO research project restricted the domain of tests to include only *systemwide* tests—that is, those tests administered to every student, to almost every student, or to a representative sample of all students in at least one grade level in a local school district or a state. Since the GAO intended to use questionnaires as its primary source of data, they realized it was impossible to ask about all tests, or even all standardized tests, because the reporting burden would have been too great and the response rate decreased in consequence.

The domain of systemwide tests, nonetheless, included about 86% of all standardized academic tests. It included all standardized tests except those administered to special populations, such as special education and gifted and talented students; optional tests, such as college entry exams; and some tests used for Title I evaluation.

Amount of Time Devoted to Testing and Test-Related Activity

In analyzing the study data, the GAO discovered that the average U.S. student took 2.5 systemwide tests during 12 years of school, in 5.8 grade levels, and in 11.6 separate test administrations. On average, that student spent less than 4 hours per year taking systemwide tests (less than 0.5 percent of a school year).[5] Counting all the time devoted to test-related activity, such as learning test-taking skills or listening to test instructions or results, the mean time still averaged less than 7 hours a year (the median was less than 6 hours). (See Table 1.)

Mandated and high-stakes testing. Only some of these systemwide tests were administered by states or mandated by states, however.[6] The average U.S. student spent about 2.1 hours per year taking state-mandated tests (and, thus, 1.3 hours per year taking non-state-mandated systemwide tests). Counting all the time devoted to state-mandated test-related activity, such as learning test-taking skills or listening to instructions or results, the duration still averaged less that 4 hours per year. (See Table 1.)

Furthermore, only some of these systemwide tests were taken for *high stakes*. In the GAO study, tests were judged to be for high stakes if it was reported that a purpose of the test was "student-level accountability, assessment used to determine promotion, retention, or graduation." The average U.S. student spent about 1 hour per year taking high-stakes tests (and, thus, about 2.4 hours per year taking low- or no-stakes systemwide tests), most of which, but not necessarily all of which, were state-mandated. Counting all the time devoted to high-stakes test-related activity, such as learning test-taking skills or listening to test instructions or results,

Table 1

U.S. Students' Time in Systemwide Testing Per Year, By Type of Testing and Activity: 1991

Activity (in the original wording)[1]	Mean number of minutes (hours) per year in *systemwide* testing[2]	Mean number of minutes (hours) per year in systemwide testing that was also *state-mandated*	Mean number of minutes (hours) in systemwide *high-stakes* testing
Minutes taking the test	202 (3.4)	125 (2.1)	55 (0.9)
Minutes in other test-related activities[3]	215 (3.6)	99 (1.7)	58 (1.0)
Total	417 (7.0)	224 (3.8)	113 (1.9)

[1]The question in the GAO questionnaire was written thus: "For this test, how many minutes did *students* spend in each of the following activities, measured in number of minutes spent *per tested student?* If the test was given in more than one form, use averages."

[2]The set of systemwide tests includes all state-mandated and district-mandated tests.

[3]These activites include: minutes of instruction in test-taking skills, of taking practice tests, or in motivational activities geared to the test; minutes receiving directions for taking the test; minutes listening to or reading a report of the results; minutes in any other way pertaining to the test.

Source: U.S. General Accounting Office, 1993.

the mean time still averaged less than 2 hours per year. (See Table 1).

Table 2 contains a recalculation of U.S. students' test-taking time so that it can more easily be used comparatively. Instead of average test-taking time per year, the durations in Table 2 represent the number of hours that the average U.S. public-school student could expect to spend taking systemwide, state-mandated, or high-stakes tests in his or her primary–secondary school career. Counted this way, the average student sat for 40.8 hours of systemwide tests. Of this total, 25.2 hours were for state-mandated tests, whereas 15.6 hours were for districtwide tests that were not state-mandated. Moreover, 10.8 hours of testing were for high-stakes, whereas 30.0 hours were not for high-stakes.

Systemwide Testing in Other Countries

The OECD survey, conducted by James Guthrie, Marilyn Binkley, and Gary Phillips in 1990–1991, was entitled "First International Survey of National and Intra-National Educational Outcome Assessment Practices," (hereafter, the OECD survey). It requested of country-level education officials detailed information about: (a) "systems of outcome measurement from which one could generalize results to the country as a whole" and (b) "systems of outcome measurement used to describe or measure student performance on a smaller scale, be it the province, state, district, school or student level" (Guthrie et al., 1990).

The OECD survey sought "relevant information about each system of assessment or examinations used within a country that has the potential to describe the performance of major portions of the student population." The instructions directed the respondents to include test information "even if its assessment consists only of examinations administered to students but not aggregated to assess school programs."

The 28-page questionnaire that Guthrie, Binkley, and Phillips (1990) designed asked for considerable detail about the character of each assessment. Each assessment required a separate questionnaire. For the purposes of this study, the most important information gathered includes the scope of each assessment, the frequency, the duration, the stakes, and the mandate. Some countries included international tests, but I did not count them.[7]

Because the OECD's aim was to study the *character* of national testing systems, however, it was not of great concern to the researchers to compile information on each and every systemwide test (Guthrie et

Table 2

U.S. Students' Hours Spent Taking Systemwide Tests During Their Primary- and Secondary-School Careers, By Type of Test and Activity: 1991

Type of test	Mean number of of hours taking tests
Systemwide (i.e., districtwide) tests*:	40.8
State-mandated systemwide tests	25.2
Non-state-mandated systemwide tests	15.6
High-stakes systemwide tests	10.8
Low- or no-stakes systemwide tests	30.0

*The set of systemwide tests includes all state-mandated and district-mandated tests.

Source: U.S. General Accounting Office, 1993.

al., 1991). Most countries, it appears, turned in complete sets of information about their *national* tests. But, because the surveys were completed by national education ministries, one would expect the information on national tests to be more complete than the information on tests at the province, state, or district level.

So, the returned set of OECD questionnaires probably represents an undercount of the extent of systemwide testing in the participating countries in 1991. In particular, local systemwide tests—tests administered to all students in at least one grade level in a region or local district—were generally not included. Some of the completed OECD surveys made reference to such tests when they were developed or scored by the national education ministry, but their extent of use can only be guessed at.

Comparing Systemwide Testing in the U.S. to That in Other Countries

Because the OECD questionnaires were so long (28 pages) and asked for much detail (with essay responses), it was fairly easy to match up the OECD study domain to the GAO study domain according to certain test characteristics, such as test duration, stakes, mandate, or *referencing* (criterion- or norm-referenced). Tests in the OECD study were judged to be for *high-stakes* if they were used "to determine promotion, retention, or graduation"—the definition used in the GAO study. Several tests in the OECD study were used in "moderation" to determine "blended marks." That is, the test score might be used in conjunction with other considerations, such as a teacher's judgment of classroom performance and homework, to determine promotion, retention, or graduation. Probably, respondents to the GAO survey would have classified such tests as being for "student-level accountability; assessment used to determine promotion, retention, or graduations." But, because I cannot be certain of that, I biased my accounting in favor of the null hypothesis and classified such tests in the OECD study as "low-stakes." They were, then, not counted in the high-stakes totals.

Table 3 shows clearly that a blanket assertion that U.S. students are "the most heavily tested on earth" has some validity problems. Table 3 lists durations for systemwide, state-mandated, and high-stakes tests for 13 countries and states returning OECD surveys, along with averages for U.S. school districts derived from the GAO survey (see Tables 1 and 2). The reader may observe that, on some types of tests, not only were U.S. students not the most heavily tested on earth, in certain ways, they were *the least heavily tested* in this group of 14 countries and states. In this group, U.S. students ranked second to last in the amount of time they spent taking state-mandated tests—well below the country average. U.S. students ranked *dead last* in the amount of time they spent taking high-stakes tests—far below the country average. U.S. students spent slightly more than one quarter the amount of time taking high-stakes tests as the country *average* for these 14 countries and states in 1991.

The contrast between the United States and other countries in the amount of high-stakes testing provides the most startling difference to be found in these data. Students in France, Italy, Denmark, and Belgium spent more than *five times* as much time taking high-stakes tests than did U.S. students. It appears that when other countries took on the expense and difficulty of developing and administering standardized tests, they were likely to make tests that counted—tests that were required and had serious consequences.

It is also fairly easy to make comparisons between the United States and other countries on their relative quantity of testing based on another characteristic—whether tests were norm- or criterion-referenced (i.e., curriculum-based). In the other countries and states, virtually all systemwide tests in the OECD study were criterion-referenced. The few exceptions were some of the national sample system monitoring exams which, all told, took up very small amounts of students' time (as measured by their "expected durations" for each student).[8] The testing experience of the average U.S. student in 1991 tells an entirely different story. Less than one third of systemwide tests taken in the United States were criterion-referenced; almost two thirds were norm-referenced.

Local Systemwide Tests

Neither the category of high-stakes tests nor the category of state-mandated tests encompasses all systemwide tests. As shown in Table 2, U.S. students face an average of 40.8 hours of systemwide tests in their primary–secondary school career. Only some of these tests are state-mandated (and, only some of them are for high stakes). Because state-mandated tests account for 25.2 test hours in a student's career, the other 15.6 hours of systemwide testing are made up of districtwide tests that are not state-mandated. (Districtwide tests that are not state-mandated may or may not be taken for high stakes.)

Suppose we include these non-state-mandated tests in the U.S. total systemwide test duration; would the U.S. total then exceed the test duration totals we see for other countries and states? No. Even counting just the state-mandated tests for the other countries and states included here, students in 6 other countries and states face more systemwide testing than do U.S. students. Indeed, all the systemwide testing that U.S. students face doesn't even add up to the country *average* for state-mandated tests, or even the country average for high-stakes testing.

U.S. students are clearly not the most heavily tested on earth if one compares systemwide tests according to their durations.

Counting Tests By Their Number

What if one measures the extent of systemwide testing, instead, by counting the simple number of tests? Such a count reveals that 10 of the 13 other countries or states had more systemwide tests than the U.S. average for systemwide tests of 2.5. So, the U.S. students do not seem to be the most heavily tested on earth according to simple counts of the number of systemwide tests each student faces in his or her school career.

But, there are still other ways to count tests. An individual test can be given more than once during the school year and at more than one

Table 3

Quantity of Systemwide Testing Encountered By Average Student During a Primary- and Secondary-School Career, By Counting Method and Country or State: 1991

Country	Student hours spent taking systemwide tests, by type of test			Number of tests, by counting method		
	All systemwide tests	Mandated tests	High-stakes tests	Tests	Grade levels tested	Individual test administrations
Belgium (French)	>50.5	50.5	50.0	3	4	5
Denmark	>170.0	100.0	170.0	3	5	9
England and Wales	>33.0	33.0	33.0	3	13	>3
Finland	31.3	31.3	30.0	2	6	>2
France[1]	61.5	61.5	51.5	4	5	5
Germany[2]	125.7	15.7	15.7	2	11	45
Italy	55.0	55.0	55.0	3	3	3
New Zealand	42.6	42.6	33.0	3	9	10
Norway	39.0	35.0	30.0	3	3	4
Scotland	>39.0	39.0	37.9	6	6	>12
Sweden	34.1	34.1	16.8	3	7	11
Switzerland (Aargau)	33.3	33.3	33.3	2	2	2
Switzerland (Geneva)	65.0	65.0	20.0	2	4	7
United States (average)[3]	40.8	25.2	10.8	2.5	5.8	11.6
Country average	>58.6	44.4	41.9	3.1	6.0	>9.3

[1]Incomplete survey data supplemented with 1995 testing data (see *Ambassade de France,* 1995a, 1995b).
[2]Data represent only the two states using the "centralized" examination.
[3]Data represent an average for all U.S. school districts.

Source: U.S. Education Department, National Center for Education Statistics, 1991, and U.S. General Accounting Office, 1993.

grade level. It can still be called a single test if the development, content, type, and purpose are similar across grade-level or seasonal administrations. (Respondents to both the GAO and OECD surveys counted tests this way, as single tests that could be administered in several versions in several grade levels. Of course, survey respondents had an incentive to define tests this way; they could then fill in fewer survey forms.) But, it also makes sense to count each grade-level or seasonal administration separately because each represents a separate occasion when a student takes a test during his or her school career.

So, counts of all separate grade-level and seasonal administrations of systemwide tests were also computed. Measured this way, the United States appears to test heavily or, rather, frequently. Only Scotland and Germany among the 13 other countries or states represented had more individual test administrations, and that condition only holds if one counts the national assessment sampling tests (held at three grade levels) in Scotland and the "Written Tests" in Germany, which seem to have been administered with so much local discretion that they barely merit being classified as systemwide or national tests.

It is rather common in U.S. school districts to administer an off-the-shelf, norm-referenced, short duration, low- or no-stakes, multiple-choice test in multiple grades for the purpose of system monitoring or student diagnosis. Indeed, around 10% of U.S. school districts administer common tests at 10 or more grade levels simultaneously.

Four countries reported more systemwide testing than did the average U.S. school district, however, if one simply counts the number of grade levels affected by testing, rather than all the individual test administrations. Sweden achieved that position without qualification (i.e., without counting any national assessment sample tests).

Summary

Based on a comparison of mostly national or state tests in 13 other countries and states to all systemwide tests in the United States in 1991, systemwide testing in the U.S. appears to have been starkly different in character from that in other countries or states. U.S. tests tended to be shorter, often much shorter, in duration. This may be because U.S. tests were more likely to be set in a multiple-choice format and taken for low stakes. It also appears that U.S. school districts were more prone than their foreign counterparts to exploit available scope economies; if a U.S. school district purchased a test from a test developer/publisher, it may have administered it at several grade levels (in slightly varying versions), reducing the unit costs of the test administration. Such arrangements make more sense when a test doesn't count for much.

In the other countries and states, important, high-stakes tests of long duration were set at key transition

points of students' careers. Such tests are not easily replicated at several grade levels. Such tests would not make much sense administered at other grade-levels because they are curriculum-specific and organized around set standards.

Counting only systemwide tests, there is one counting method which implies that U.S. students might be one of the most heavily tested among this group of 14 countries or states. It consists of counting tests by the number of individual administrations throughout students' primary–secondary school career rather than by their duration.

Counting tests in other logical ways does not produce the same rank for the United States. Based on the simple number of separate tests, 10 countries or states ranked higher. Based on the cumulative *duration* of tests, several other countries or states had more state or national testing than the average U.S. school district had of any kind of systemwide testing. Comparing the subcategories of state-mandated tests and high-stakes tests in terms of their cumulative duration, U.S. students saw relatively little testing.

Vocational-Track Tests

The U.S. education system is notable for the weakness of its secondary-education-level vocational curriculum. Whereas many other countries maintain a rather separate vocational educational system that students are either steered toward or away from in their lower secondary years, the typical U.S. high school offers only some perfunctory vocational courses, usually within the context of a general academic curriculum.

Some observers might argue that some of the upper secondary-level tests in other countries pertain only to academic-track students. But, in these countries, a sizeable minority of students, maybe even a majority, attends classes, instead, in vocational-track schools.[9] These observers might advocate reducing the national or state testing counts to account for an absence of vocational-track students.

That would be a mistake, because those students are tested. In some countries with strong upper secondary vocational tracks (such as Belgium, New Zealand, and Scotland), all students do, indeed, take the regular, academic upper secondary exit examination, no matter which track they're in. In other countries with strong upper secondary vocational tracks (such as Germany, Austria, Switzerland, and Korea), the vocational-track students must take skills certificate examinations that are developed by national or state skill boards (often with craft union representation, employer representation, or both); these examinations are usually performance-based and can be rather lengthy. In still other countries with strong upper secondary vocational tracks, vocational-track students take a general upper secondary school leaving exam designed for their track (such as the technical or commercial series *baccalauréat* in France).

Unfortunately for this study, however, only one of the countries in the OECD study—Switzerland—provided information about vocational-track tests. Switzerland's Canton of Aargau reported that 270 professions had certificate standards enforced by the state. The three most popular professions—business, building designer, and auto mechanics—together accounted for about 50% of vocational-track graduates. Each of these professions required satisfactory passage of exit certificate examinations lasting 21.5, 31, and 32 hours, respectively. These were high-stakes, state-mandated tests, and any one of them alone represented more time in testing than all the high-stakes tests taken in 12 years of school by U.S. students. Two of the three exams each alone represented more time in testing than did all state-mandated testing for a U.S. student.

Classroom Testing in the United States and Other Countries

The International Association for the Evaluation of Educational Achievement is a loose-knit organization of national education ministries that occasionally puts together massive, complicated worldwide administrations of student achievement tests in one or two subject areas, each test written in the national or regional language of the students. The resulting test scores are assembled onto a common scale and compared.

Along with its achievement tests, the IEA administers questionnaires to teachers, students, and education ministry officials regarding classroom practices, national education policies, student study habits, and so on. The responses to these questionnaires provide context for the test scores. The IEA's 1991 Reading Literacy Study in particular included a questionnaire for reading teachers that asked them about the frequency of their assessments in reading. Teachers in 31 countries, including the United States, responded.

Were the U.S. students the most heavily tested? No. Responses to the questions about the frequency of classroom assessment were ordinal, ranging from "almost never" to "about once a week or more." According to teachers, U.S. 9- and 14-year-old students were tested with multiple-choice instruments more than the average for all 31 countries. The U.S. was tied for seventh place with Greece and Slovenia for its frequency of use of multiple-choice tests in reading (at a reported frequency of about "once a term"), ranking below Thailand, Botswana, Nigeria, the Netherlands, Cyprus, and the Philippines (Lundberg & Linnakyla, 1993, pp. 77–79).

In the frequency of use of most other types of testing instruments in reading, the U.S. tied with Botswana, Nigeria, and the Netherlands (at a reported frequency of slightly less than "once a month"), below 22 other countries. In the frequency of use of most types of reading tests, then, U.S. students ranked among the least tested in the world.

Data from the IEA's Second International Mathematics and Science Study (SIMSS) in 1982 show results regarding the frequency of classroom testing similar to those of the 1991 Reading Literacy Study. While U.S. teachers reported a greater frequency than the average for other countries in one or another type of testing (in science, it was teacher-made short-answer tests), the U.S. did not rank highest. And, when all types of classroom tests were considered, the U.S. seemed about average (Wolf & MacRury, 1991).

In 1991, the International Assessment of Educational Progress administered mathematics and science tests in 20 countries. The IAEP orga-

nization was also rather loose-knit and also run by a committee of participating education ministries. But, unlike the IEA, which was run in all its aspects as an international collective, the IAEP was assembled by the Educational Testing Service of the United States, which developed the test (modeled on the U.S. National Assessment of Educational Progress) and analyzed and reported its results.

As with the IEA exams, however, the 1991 IAEP mathematics and science tests were accompanied by questionnaires for participating teachers, students, and education ministry officials regarding classroom practices, country education policies, student study habits, and so on. One multiple-choice question asked 13-year-old students how often they took mathematics (or science) tests or quizzes. The first three possible responses were: A - every day, B - several times a week, and C - once a week. If one counts just the first two responses (A + B = at least several times a week), the United States ranked 10th out of 20 countries in its math test frequency and 5th out of 19 countries in its science test frequency. If one counts the first three responses (A + B + C = at least once a week), the United States ranked 3rd out of 20 countries in math and 3rd out of 19 countries in science. Other countries which tested frequently included Taiwan, China, France, the Soviet Union, and Jordan (Educational Testing Service, 1991).

Data from the three aforementioned international surveys—SIMSS, Reading Literacy, and IAEP—do not support the proposition that U.S. students see the most classroom tests. U.S. students may see more multiple-choice or short-answer tests than the average student, and the frequency of testing may be especially high in science, but, apparently, the frequency of testing in reading is especially low. All told, given any subject matter and given any method of counting classroom testing frequency, one can always find other countries that test more.

Conclusion

Are U.S. students the most heavily tested on earth? Data from the OECD and GAO surveys would suggest that one might be able to argue the point either way. But, without doubt, testing in the U.S. appears to be very different in character from that typical in other countries.

U.S. students face:

- fewer hours and fewer numbers of high-stakes standardized tests than their counterparts in every one of the 13 other countries and states represented here;
- fewer hours of state-mandated tests than their counterparts in 12 of the 13 other countries and states;
- fewer numbers of systemwide tests than their counterparts in 9 of the 13 other countries or states;
- fewer numbers of criterion-referenced systemwide tests than their counterparts in all 13 other countries or states;
- a greater number of individual administrations of short, norm-referenced systemwide tests with low or no stakes attached than their counterparts in all 13 other countries or states: and
- a greater-than-the-international-average frequency of classroom tests in mathematics or science and a less-than-the-international-average frequency of classroom tests in reading, but no absolute superiority in the frequency of classroom testing in any of the three subject areas.

Based on just the data included in this study, then, it would appear that U.S. students may face more systemwide testing than most of their foreign counterparts if one counts tests by their number of individual test administrations and ignores their duration, their mandates, their stakes, and their *referencing* (either norm or criterion). In other words, U.S. students seem to frequently face short, low, or no-stakes tests.

One should remember, however, that this study only includes what may be incomplete information on testing from some of these 13 other countries or states, that there are many other countries in the world other than just those included here, and that some of them may also conduct more systemwide testing than the U.S. average.

Discussion

Two more points are relevant and deserve discussion.

First point. It is this author's observation that standardized testing's most vociferous critics in the United States focus their objections on "important," "external," and "high-stakes" standardized tests. These are the bad tests, in their opinions. These are the tests that do harm. These are the tests of which there are too many. These bad tests distort and "narrow the curriculum," cause undo stress, and intrude, interfering with the natural good instincts of well-meaning and well-trained teachers by imposing artificial, external constructs, restrictions, and standards. Teachers may respond by "teaching to the test."

By contrast—it is my impression—these same critics would argue that low-stakes standardized tests which are used merely for system monitoring or student diagnosis are fine and their use should be encouraged relative to that of high stakes and mandated tests.

It is quite ironic to learn, then, that U.S. students may already be seeing the lowest amount of "bad" standardized testing in the world and the greatest amount of "good" standardized testing. Indeed, given the apparent state of affairs in testing around the world, why are the U.S. critics of standardized testing complaining? Our students face the lowest amount of high-stakes, mandated, and criterion-referenced testing in the world. Instead, our students face a plethora of . . . well . . . *unimportant* tests.

Second point. Enormous advantages in efficiency are created by wholly integrating examinations into the structure of a country's or state's education system. In many countries and states, examinations are systemwide, curriculum-based, high-stakes, and set at transition points between levels of education. Done this way, the curriculum determines the tests, and the tests determine the curriculum.

Done this way, every teacher, administrator, and student has clear goals, standards, rewards and punishments. Students who don't pass an exam do not go on to the next level. Administrators of schools whose average student score on an exit exam is especially good or poor may face public questioning when their school's average test score is compared to that of other schools, or the systemwide average. Teachers of a particular subject area in which a school's average test score is espe-

cially good or poor may face public questioning from administrators or parents when that score is compared to that of other schools, or the systemwide average. National or state education ministers may face questions if average student scores trend down or up under their tenure. Common standards and measurements help form coherent systems. Clear goals, standards, and tests clarify the process of achieving them.

Another benefit of important, high-stakes tests is that they buttress the power of teachers by imposing another standard on students' behavior other than just the teacher's.

Other researchers have argued that the high-stakes tests in Europe are used for selection, credentialing, or certification of individuals only and not for system monitoring and accountability or instructional feedback to teachers and students (Feuer & Fulton, 1994, p. 36; Madaus & Kellaghan, 1991; U.S. Congress, 1992, pp. 135, 142–146). I believe these assertions are naive. The OECD surveys reveal at least four countries—Finland, Belgium, France, and Norway—that explicitly claim to use student tests for both student accountability and system accountability and monitoring. The OTA report claimed that only Sweden and China used systemwide tests for both purposes (U.S. Congress, 1992, p. 138). Even in countries and states where student performance on systemwide tests is not *officially* considered to be part of teacher and administrator performance evaluation, it may be anyway. Parents and journalists in other countries are no more prone to ignore such information than they are here.

As for the alleged lack of instructional feedback from high-stakes tests in other countries . . . it is human nature to try to find out what went wrong on a less-than-perfect test performance. I would argue that it is far *more* likely that students, teachers, and administrators will pay attention to a test performance when the test has high stakes.

Notes

The author would like to thank Marilyn Binkley, Tom Jirele, Keith Rust, Maryellen Schaub, T. Neville Postlethwaite, John H. Bishop, Jay Moskowitz, David Baker, JoAnn Blue, the editor, and several anonymous reviewers for their help or comments. Any mistakes and annoying opinions that may remain are the responsibility of the author alone.

[1]Specifically, he said, "Before American students, *already the most heavily tested in the world,* are subjected to yet another testing treatment, . . ." (see Madaus, 1991, p. 2).

[2]Specifically, he wrote "The truth is that our students are already the most overtested in the world, with more than 100 million standardized, multiple-choice exams given each year" (see Neill, 1992, p. 46).

[3]According to the U.S. General Accounting Office, a *systemwide* test is one that is taken by all students, almost all students, or a representative sample of all students in at least one grade level in a school district or state. The category of systemwide tests is approximately equal to the less well-defined category of large-scale tests and comprises a large proportion (about 86%) of all standardized student academic tests in the United States.

[4]For New Zealand, the current year was 1990.

[5]The exact number is 3.4 hours. This statistic represents the mean for all U.S. students; the median was 3 hours per student. If one were to calculate the mean based, not on all U.S. students, but, instead, based on the total number of students tested in 1990–1991, as represented by the total number of separate individual test administrations, one gets a somewhat higher mean of 3.9 hours. The number of students tested equals about 89% of all U.S. students.

[6]There is a difference between a *statewide* test and a state-mandated test. *Statewide* tests are single tests administered verbatim in all school districts throughout the state. *State-mandated* tests are tests that may differ in form and content one from another but are still administered in all districts throughout the state. A state that develops a common test that all students in the state must take is administering a statewide test. A state that simply requires that districts in the state administer any test that meets certain minimal requirements is mandating a test. Statewide tests are a subset of state-mandated tests.

[7]Several countries provided information regarding their participation in the IAEP, which was administered in 1991. But, the IAEP was not included in calculating any of the measures of the extent of systemwide testing. (Survey responses from the IAEP regarding the extent of classroom testing were considered separately.)

[8]The *expected duration* of a test is the duration of the test multiplied by the proportion of the student population taking the test. For a national test that samples 5% of students in a grade level and takes 5 hours to complete, the expected duration would be one quarter of an hour. It is that number, not the actual duration, that is added to the country and state totals for test durations in Table 2.

[9]In making comparisons of the proportion of adolescents in a country's population tested, however, it would be incomplete to just consider the numbers in academic and vocational tracks. One should also consider the numbers who attend school at all. The United States, for example, produces more school dropouts among older teenagers than do most other advanced industrialized countries. School dropouts, of course, take no tests at all.

References

Ambassade de France, Centre d'examens des Etats Unis. (1995a). *Calendrier des épreuves (et pratiques) du baccalauréat session 1995 [Baccalaureate examinations schedule 1995.]* Washington, DC: Author.

Ambassade de France, Centre d'examens des Etats Unis. (1995b). *Calendrier prévisionnel du diplôme national du Brevet session 1995 [Junior high exit examination schedule 1995].* Washington, DC: Author.

Feuer, M. J., & Fulton, K. (1994). Educational testing abroad and lessons for the United States. *Educational Measurement: Issues and Practice, 13*(2), 31–39.

Guthrie, J. W., Binkley, M., & Phillips, G. W. (1990). *First International Survey of National and Intra-National Educational Outcome Assessment Practices.* Center for Educational Research and Innovation of the Organization for Economic Cooperation and Development Project on International Education Indicators. Paris: OECD.

Guthrie, J. W., Binkley, M., & Phillips, G. W. (1991). *Assessing assessments: Considerations in selecting cross-national educational performance indicators* (INES Project, General Assembly, Network A). Lugano, Switzerland: OECD.

Lundberg, I., & Linnakyla, P. (1993). *Teaching reading around the world.* The Hague, Netherlands: IEA.

Madaus, G. (1991, June). *The effects of important tests on students: Implications for a national examination or system of examinations.* Paper presented at the AERA Conference on Accountability as a State Reform Instrument, Washington, DC.

Madaus, G., & Kellaghan, T. (1991). *Student examination systems in the European community: Lessons for the United States* (Contractor report submitted to the Office of Technology Assessment). Washington, DC: U.S. Congress.

Neill, M. (1992). Correcting business leaders' assumptions about testing (Letter). *Education Week, 11*(27), 46.

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.
U.S. Education Department, National Center for Education Statistics. (1991). *First International Survey of National and Intra-National Educational Outcome Assessment Practices.* Unpublished tabulations.
U.S. General Accounting Office. (1993). *Student testing: Current extent and expenditures, with cost estimates for a national examination* (GAO/PEMD-93-8). Washington, DC: U.S. General Accounting Office.
Wolf, G. R., & R. A. MacRury. (1991). *IEA Second International Mathematics and Science Study Data Analyses and Archiving Project.* Unpublished tabulations, Ontario Institute of Studies in Education, Toronto.

1996 NCME Award for Application of Educational Measurement Technology to a Specific Problem

Professor Wim van der Linden was selected as the recipient of the 1996 NCME Award for Application of Educational Measurement Technology to a Specific Problem. Professor van der Linden, who is from the University of Twente in the Netherlands and is a long time member of NCME, accepted the award at the NCME Breakfast at the 1996 Annual Meeting in New York. The Committee's decision was based on Professor van der Linden's on-going program of research in the area of optimal test design, which has culminated in the production of the software package *CONTEST*. Important considerations in the deliberations of the Committee were the package's strong theoretical underpinnings, user-friendly interface, and availability to researchers and practitioners. The Committee agreed that Dr. van der Linden's work in optimal test design coupled with the development of a software product that makes the work accessible to a broad range measurement professionals represents a substantial contribution to the field. A description of Dr. van der Linden's work, as well as some history of psychometrics that compelled that work, was provided by Dr. Ronald K. Hambleton in his letter of nomination (personal communication, January 16, 1996). Selected excerpts from that letter follow.

Frederic Lord and Allen Birnbaum were the first psychometricians to sketch out a general strategy for test construction using item response models. This work in the 1960s involved the use of item and test information functions and was labor intensive for producing tests to meet both content and statistical specifications. As originally formulated by Lord and Birnbaum, the strategy involved only statistical considerations of test items. Even so, the strategy was awkward to implement. With the addition of hundreds of content constraints (as is the case with many important aptitude, achievement, and credentialing exams), nonautomated test construction strategies would be difficult, if not impossible, to implement in practice even with all of the advantages of IRT models.

The major breakthrough came in the middle 1980s with the recognition by Professor van der Linden and several of his exceptionally talented students—most notably, T. J. Theunissen, J. J. Adema, and Ellen Boekkooi-Timminga—that a solution to the problem of automated test construction to meet large numbers of test specifications could be found in the operations research literature. Since about 1986, Professor van der Linden and his students, along with several colleagues in the United States and the Netherlands, began their research program, which involves a complicated interaction among item response theory models and procedures, operations research, and test design. There are more that 50 research papers by Professor van der Linden and his students and colleagues on this topic. By any standards, this is immensely productive output! This research has been published in most of the prominent refereed journals—such as, *Applied Psychological Measurement, Psychometrika,* and the *Journal of Educational Statistics* [now *Journal of Educational and Behavioral Statistics*. Ed.]

Professor van der Linden and his students and colleagues have initiated a program of research and development that is comprehensive in scope and deep in psychometric theory and operations research. This research includes everything from conceptualizing the test development problem in operations research terms, to complicated designs for parameter estimation, to incorporating both classical and modern approaches to test development, to the development of various criteria for test design (to reflect popular test development practices such as designing a test to match or exceed a test information function), to the construction of multiple forms of a test simultaneously. More recently, their research has emphasized the special applications of optimal test design to computer adaptive testing.

A unique and critically important outcome of Professor van der Linden's work is the result of his decision to present it in the form of a user-friendly software package *(CONTEST)* that is available to interested persons. Much of the IRT application work to date has been plagued by a failure of researchers to produce useful software. The other problem is that major testing agencies do much of the research and publish their papers but then are unwilling to make the software available for others to use. Professor van der Linden's work is a major exception to the unfortunate rule. As importantly, Professor van der Linden has remembered that not everyone in psychometric methods is committed to modern test theory, sometimes known as item response theory. By drawing on well-established relationships between classical and modern test theory, Professor van der Linden has made it possible for those with a classical persuasion to test development to benefit from the models, principles, and procedures associated with optimal test design.

With optimal test design, test developers can communicate their test content and statistical specifications in a simple form to the computer. These specifications are then converted into a series of linear equations which can be solved. Then the computer selects a set of test items from the available item bank that best approximates the desired test. This automated approach to test development operates very much like an expert system. Users do not need to understand the technical details of item response theory, classical test theory, item statistics, reliability theory, and so forth. What they need to be able to do is to describe the type of test that they would like in terms clear enough to allow the software to function.

Professor van der Linden's work in this area is on-going. At the time that he was told that he was selected to receive the NCME award, he noted that this sign of recognition and appreciation by his colleagues would serve as an impetus to continue to improve, enhance, and expand his research in optimal test design.