The TIMSS "Final Year" Study and Report: A Critique

GERALD W. BRACEY

Educational Researcher, Vol. 29, No. 4, pp. 4–10

The Third International Mathematics and Science Study (TIMSS) has produced curriculum analyses, case studies, and videotapes of instruction in several countries and more reports are to come. TIMSS has also generated test scores for many countries at Grades 4 (Population 1 in TIMSS terminology) and 8 (Population 2) and "the final year of secondary school" (Population 3) as well as links between the eighth-grade scores and 1996 National Assessment of Educational Progress (NAEP) mathematics and science scores[1] (Martin et al., 1997; Mullis et al., 1997; Beaton, Martin, et al., 1996; Beaton, Mullis, et al., 1996; Johnson & Siegendorf, 1998).

Quotes occur in the last sentence to draw attention to the fact that in a number of countries the final year of secondary school is not equivalent to "Grade 12" in the United States, rendering the scores non-comparable. Indeed, the systems and cultures of the nations involved differ to an extent that renders the scores uninterpretable. The problems in the Final Year Study are diffuse and multifold, rendering a precise interpretation of how they affected the data impossible.

The results have, unfortunately, been widely accepted in the culture at large as reflecting the scores of tests given to "twelfth graders" around the world. The following quotes reveal the way the media and, therefore, the popular culture has viewed the results:

> American high school seniors have scored far below their *peers* from many other countries on a rigorous new international exam in math and science. (Sanchez, *Washington Post*, 1998, p. A1, emphasis added)

> American high school seniors—even the best and brightest among them—score well below the average for their *peers* participating in TIMSS. (Viadero, *Education Week*, 1998, p. 1, emphasis added)

> TIMSS presents findings concerning the standing of U.S. 12th-graders . . . compared to their *peers* participating in TIMSS. (Valverde, 1998, p. 4, emphasis added)

> The conclusion is unmistakable: The longer students stay in American schools, the farther they fall behind their *agemates* in most industrialized nations of the world. (Dobson, *Family News*, 1998, p. 1, emphasis added)

> U.S. Twelfth-Graders Rank Poorly in Math and Science Study. (Bronner, *New York Times*, 1998, p. A1)

> American 12th-graders scored at the very bottom of the rankings. (Raspberry, *Washington Post*, 1998, p. A25)

> Poor academic showing hurts U.S. high schoolers. (Henry, *USA Today*, 1998, p. 1A)

> Hey! We're No. 19! (Leo, *U.S. News & World Report*, 1998, p. 14)

However, the American students in the TIMSS Final Year Study (Mullis et al., 1998) were mostly *not* age-mates of those in other countries. In the math/science literacy assessment, American students averaged 18.1 years. In 4 of the other 20 nations, students were almost 19 years of age, in 6 they were over 19, and in 2 they were over 20, with Iceland garnering first place for the eldest cohort tested, 21.2 years. In the 16 nations that participated in the physics assessment, only the Russian Federation's students were quite a bit younger than American students (16.9 years vs. 18.0), but Russia only tested 2% of them (compared with 14% in the U.S). Students in Denmark, Italy, Austria, Germany, and Switzerland were all over 19 and had one or two more years of schooling than U.S. high school seniors.

Given the wide acceptance but misinterpretation of the TIMSS Final Year Study, a clarification seems in order. This paper describes the tests used in the Final Year Study and discusses the various problems that render any interpretation of the study problematic.

The report on the tests given to Population 3, students in their final year of secondary school, is hereafter referred to as the "Final Year Report." This report contains the results from three separate tests. One, a general test of "mathematics and science literacy," was to be administered to a representative sample of students in their final year. Two other tests, of physics and advanced mathematics, were to be administered to whatever population the participating countries deemed appropriate.

Earlier TIMSS reports contained data about math and science achievement in Grades 4 and 8. Tests administered in 26 nations at Grade 4 found students in the United States above average in both math and science; tests given in 41 countries at Grade 8 found American students slightly above the international average of all 41 countries in science (58% correct vs. 56%), and slightly below the average in mathematics (53% correct vs. 55%) (Martin et al., 1997; Mullis et al., 1997; Beaton, Martin, et al., 1996; Beaton, Mullis, et al., 1996). Removing the countries that failed to meet the TIMSS exclusion and/or participation rate criteria still leaves a large

GERALD W. BRACEY *is an independent researcher and writer living in Alexandria, VA; e-mail: gbracey@erols.com. His areas of specialization are educational assessment and educational policy.*

number of countries and does not materially affect the results in terms of the relative rank of the United States.

The Final Year Report, *Mathematics and Science Achievement in the Final Year of Secondary School,* shows that U.S. 12th graders appear to score far below students in 20 other nations in math/science literacy and 15 other nations in physics and advanced mathematics. Yet its opening pages describe the problems in obtaining comparable results across nations:

> testing this "grade" was a special challenge for TIMSS. First there was the question of how many students of the age-eligible cohort are even in school by the final year, and how this might differ across countries. Second, it was no small task for many countries to describe the final year of school. In most TIMSS countries, students' final year of school depends on their course of study (e.g., academic, technical or apprenticeship). Thus, the final year of schooling varies across and within countries, with some students completing secondary school after a two-, three-, four-, or even five-year program. (Mullis et al., 1998, p. 1)

This quote clearly indicates that the TIMSS staff was aware of the problems in comparing countries at the final-year level.

Misleading Statements

Given the diversity of the curricula across countries and given that "secondary school" can last from 2 to 5 years, one might have expected that TIMSS staff and the U.S. Department of Education, which funded much of the study, would have emphasized the differences among these countries and would have tried to tease out how differences among the systems affect and produce differences in outcomes. One might have expected, as well, that the two organizations would have taken care to ensure that when the data were released, these many differences among systems were made clear to the media and other interpreters. As the quotes above from reporters and others (which could be multiplied many times over) show, however, this did not happen.

At a Brookings Institution panel some 6 weeks after the report was released, William Schmidt, TIMSS research coordinator, did assert that the TIMSS Final Year assessments compared *systems,* not *students* (Schmidt, 1998). At the same panel, U.S. Commissioner of Statistics, Pascal Forgione, declared that "The purpose of this [Final Year] component of TIMSS was not to compare students of the same age or years of schooling, but rather to compare students at a similar point in the education system: the end of secondary school" (Forgione, 1998).

These comments are germane to an understanding of the Final Year Report. Unfortunately, as the media citations indicate, by the time that Schmidt and Forgione made these statements, a different understanding of the study already prevailed across the country. The study has been universally portrayed, mostly in front-page media stories, as comparing *students* not *systems*, often using the word "peers."

If comments such as these were limited to the media, they might be regretted but dismissed. Other people, though, including some educators, have made similar comments. In *Education Week,* Senta Raizen of the National Center for the Improvement of Science Education was quoted as saying, "Some of our science and math folks . . . have been saying for years that our best kids are the best in the world. Well, they're not" (Viadero, 1998, p. 1). The same story cited Schmidt as lamenting that "These essentially are just devastating results. There's no other way to cast them" (p. 1).

Without a single exception, the study has been characterized by the media as a study of "peers" in various nations. The following composite description has emerged: "A school is a school is a school in any country. Twelfth graders in 24 nations were tested.[2] Our best 12th-grade students went up against their best 12th-grade students and got trounced."

How did such a misinterpretation occur? It occurred in part because neither the U.S. Department of Education nor TIMSS spokespersons emphasized the comparison among *systems* in contrast to *student*s, nor did Department of Education publications. Nor did either offer any explanations about which differences in different systems produced differences in outcomes.

Whereas the TIMSS Final Year Report noted the *differences* among the structure of secondary education in different countries (Mullis et al., 1998), *Pursuing Excellence,* the report from the U.S. Department of Education, accentuated *similarities* (National Center for Education Statistics [NCES], 1998). Indeed, it not only contained misleading statements about the comparability of the students, but made factual errors about such comparability. Consider this paragraph:

It remains that the U.S. chose to have 50% of its test takers sit for an exam where a quarter of the items covered material that they had not studied.

> As is discussed in more detail in chapter 4, the most recent data indicate that in most countries participating in TIMSS secondary school enrollment rates are similar to that of the United States. **Not only do the TIMSS countries have most of their secondary school-age population enrolled in school, the strict quality controls discussed earlier ensured that the sample of students taking the mathematics and science general knowledge assessments were representative of the entire population at the end of secondary school**." (NCES, 1998, p. 20, bold in the original)

The first sentence in the above paragraph is technically true but misleading; the bolded sentence is incorrect. After noting that "there remains considerable variation [among countries] in completion rates" (Mullis et al., 1998, p. 17), the TIMSS Final Year Report presents a table showing those rates (p. 22). The countries do have "most" of their students in secondary school, but "most" means as few as 77%. Is a country with only 77% of its students enrolled in secondary school comparable to the U.S. with 97% enrollment? By what criterion could we make such a determination?

The likelihood of non-comparability increases when one considers that the figures given are not for students in their final year of secondary school. They are for the aggregate enrollment from ages 12 to 17. The study, however, mostly tested students aged 18 to 21. Using "country" as the unit of analysis, the average student tested in the math/science literacy test was 18.7 years old. It is likely that the enrollments near the actual age of secondary school completion are much lower and that the countries vary more. What infer-

ences of comparability can be made from aggregate enrollments for ages 12 to 17 when in 13 of 21 participating countries the tested students were 18.5 years old or older? Table A5.14 in the 12th-grade edition of *Pursuing Excellence* reveals that in the U.S. only 22% of 18-year-olds are in secondary school. For the 14 nations whose scores exceeded the United States in the test of general knowledge, the average percentage of 18-year-olds in secondary schools is 60%.

That the enrollments don't reflect the population of those secondary school students actually tested is also seen in the TCI—the TIMSS Coverage Index. This is an estimate of the percentage of the school-leaving age cohort covered by the TIMSS final year student sample. The Indices for advanced mathematics and physics would be expected to be small because these tests were administered to a small sample of students. But for only Slovenia at 88% does the Index appear to come close to representing the whole student population, even for the general test of math/science literacy. The other nations' indices range from 43% in Lithuania to 84% in France and Norway. It is 63% for the United States. The mean for all nations is 67%.

The "Devastating" Character of the Results

The earlier quotations from Raizen about the less than scintillating performance of our best students and from Schmidt about the "devastating" nature of the results are somewhat called into question where we actually have some disaggregated data that bear on the matter. For instance, Fairfield High School in Fairfield, Connecticut was one school drawn in the national probability sample by Statistics Canada.

Fairfield High School sits in a community that is affluent but not exceptionally so. The Connecticut Department of Education establishes Educational Reference Groups for its schools based on demographics characteristics such as income and educational level. Fairfield is in the second of these five groups, meaning that there is a group of towns that are wealthier than Fairfield and three groups that are not so wealthy. Fairfield has a 7% minority enrollment. Ninety-two percent of Fairfield's seniors take the SAT and score in the 61st percentile on the verbal test (529) and the 64th on the mathematics test (535).

Tables 1 and 2 show Fairfield's results compared to results for (a) the highest-scoring nation, (b) the United States, and (c) the average of all nations. Although the TIMSS Final Year Report gives one international average math/science literacy score, the results provided to Fairfield separate the scores from these two tests.

Table 1
Math/Science Literacy Scores

	Mathematics	Science
Highest-scoring nation[a]	560	559
United States	461	480
International average	500	500
Fairfield H.S.	536	559

[a] The highest-scoring nation for mathematics was Norway. The highest-scoring nation for science was Sweden.

Table 2
Advanced Mathematics and Physics Scores

	Advanced Mathematics	Physics
Highest-scoring nation[a]	557	581
United States	442	423
International average	501	501
Fairfield H.S.	553	614

[a] The highest-scoring nation for advanced mathematics was France. The highest-scoring nation for physics was Norway.

From Tables 1 and 2 we see that students at Fairfield High School scored as high, or nearly as high, as the top-ranked country in both science and advanced mathematics. They scored somewhat lower than the top-ranked nation in the general mathematics assessment, but substantially higher than the top nation in physics. This last result is particularly interesting since the students in Norway tested in physics had studied the subject for 3 years and constituted only 8% of the relevant class, the fourth lowest proportion among the 16 nations involved. Some[3] of the Fairfield students would have taken 2 years of physics: a general year as juniors and an Advanced Placement (AP) year as seniors. Still, in the national results, American students taking AP physics scored only 474. The students were chosen randomly for participation by TIMSS, not selected by Fairfield.

These results do not provide any definitive conclusions, but the enormous differences between the scores of Fairfield and the U.S. average—especially between Fairfield and other American students in AP courses—might raise some questions about the sample of U.S. students. They call into question the characterization of the results as "devastating." And most certainly, they reflect the inappropriateness of responding to the average as if it characterized the performance of the nation as a whole.

The Failure of Quality Controls

The "quality controls" mentioned in *Pursuing Excellence* refer to criteria established by TIMSS for participation rates and sampling procedures. For the math/science literacy assessment, only 8 of the 21 participating countries[4] met these criteria. For advanced mathematics and physics, 6 of 16 nations could meet the criteria.[5] It cannot be determined how the failure to meet quality controls affected the data, but the failure greatly lowers the confidence one can place in *any* interpretation.[6]

The discussion of the TIMSS Final Year Report, though, has proceeded as if the data-quality criteria were not important—indeed, as if the criteria did not exist. Countries that did not meet the criteria are "annotated" in all documents; however, with one exception, these annotations have not been mentioned in any comments from TIMSS or Department of Education staff. Among the media, only *Education Week's* table of scores showed the annotations (using the TIMSS convention of placing parentheses around the countries' names). At the earlier mentioned Brookings Institution seminar, Beaton observed only that these annotations exist (Beaton, 1998).

Even if the criteria were met or did not exist, the data would be difficult to interpret because of the various differences among nations. For example, in connection with overall secondary enrollment, Forgione says, "Thus, while variation in enrollment rates does exist, the countries are *roughly comparable*, and more so than in previous years" (1998, p. 770, emphasis added). A phrase like "roughly comparable" is difficult to interpret precisely in the context of a research study. As noted, the variation in enrollment rates is from 77% up for ages 12 to 17, but there is likely more variation at the end of the Final Year. Even a range from 77% up makes one wonder about the designation of these countries as "roughly comparable." When one considers that the students who were tested were mostly 18 years or older, the meaningfulness of determining comparability across countries based on enrollment figures for ages 12 to 17 is likewise called into question.

Forgione (1998) then argues that "among seventeen-year-olds, the U.S. actually has a smaller proportion in school than the average for the other TIMSS countries for which this information is available (75% vs. 82%)" (p. 770). This figure of 75% is taken from the 1997 edition of the Organization for Economic Cooperation and Development's (OECD) *Education at a Glance*. It appears to be inaccurate, but attempts to track down the source have not been successful.

The OECD figure for U.S. enrollment is contradicted in several places by other publications presenting the same statistic. The figure given by the 1996 edition[7] of *The Condition of Education*, for example, is 92.4% while the 1996 edition of *Education at a Glance* presents the figure as about 86% (NCES, p. 40; OECD, p. 120; the OECD graph cannot be read with exact precision). The 86% figure is much more in accord with other statistics on U.S. attainment.

As noted, the enrollment for ages 12 to 17 is also not an accurate indicator of who was tested in secondary schools in the various countries. The complete table from which Forgione's figure was taken is presented in Table 3. Even though this figure is inaccurate for the real percentage of 17-year-olds enrolled in secondary school, the table still indicates that in the United States, secondary schooling is largely over by the age of 18. American students were tested in May of their senior year and their average age was 18.1 years. While a much larger proportion of 18- and 19-year-olds are still in school in other nations, their proportion is much smaller than the proportion of 12- to 17-year-olds. And it was largely the 18- and 19-year-olds who were tested in other countries.

Table 3
Percentage of Students in Secondary School

	Age			
	17	18	19	20
United States	75	22	4	2
International average[a]	82	60	34	20

[a] These are the mean values. Medians are somewhat different. For instance, the median value for enrollment of 18-year-olds in secondary school outside of the United States is 65%.

The risks of estimating school attendance at age 17 and then testing mostly those 18 years and older can also be understood by looking at the profiles of the various nations' systems (Mullis et al., 1998, pp. A1–A26). The descriptions of the educational systems in the TIMSS Final Year Report show that by age 18 the proportion of students in school has dropped considerably in some countries. For Canada, for instance, the Final Year Report gives the secondary enrollment as 88%, but the description of the Canadian system shows that only 74% were enrolled in the three grades (12, 13, and 14) where the general knowledge test was administered.

Similarly, the enrollment figure for New Zealand is given as 104% (figures above 100% are possible when people outside the age range used are enrolled). But the description of the system shows that only about 80% of the students were still in school at the time of general knowledge testing. The Russian Federation is shown at 88% enrolled but it excluded all vocational students, leaving only some 52% of the students to be tested. Why this did not earn the Russian Federation an "annotation" in the TIMSS Final Year Report is not clear. In any case, an examination of the proportions of students in school in the grades tested is often much smaller than the enrollment figures used and further undermines confidence in the quality controls.

Differences in the Populations Tested

The description of the Russian system raises the issue not only of what age was tested, but *who* was tested. In the United States, the advanced mathematics sample included students enrolled in pre-calculus courses. Yet 23% of the items on this test presumed that the students had taken calculus. It is thus not surprising to find that the pre-calculus American students scored some 100 points lower than American students who had actually taken calculus.[8] American students who had taken calculus scored at the international average (although given the problems in the study, it is not clear what that means).

The inclusion of pre-calculus students has been justified on the grounds of "fairness" and the need to make the proportion of American students tested similar to the proportions tested in other nations: "Would it be fair to compare seven percent in the U.S. with over 16 percent in Canada, 20 percent in France or 33 percent in Austria?" (Forgione, 1998, p. 770). This justification can be questioned from other perspectives on fairness. Is it fair to test students on material they have not studied, namely, calculus? Has not a great deal of testing literature since *Debra P. v. Turlington* turned on the conclusion that the test givers must prove that the tests have instructional validity—for example, that the students have actually had an opportunity to learn the material on the test?

One can legitimately ask why the proportion of American seniors taking calculus is small. The answer would likely be in terms of how this country has historically designed its math curriculum: Calculus has long been viewed in this nation as a college-level course. It might well be that this design needs to be rethought. There is no major cognitive development at this age that would prevent pre-college students from learning calculus. European nations moved to include calculus as part of the secondary school curriculum early in the 20th century (J. Kilpatrick, personal communication, August, 1998). The proportion of high school students going on to college in European nations remains far below

the proportion in the United States. Historically, a nation deeming calculus an important topic of study would have had to provide it at the secondary level.

Rethinking the structure of the U.S. secondary school mathematics curriculum is an important issue, but it is a different issue than the one raised by testing students who have not studied the content of the test, just to increase sample size. The latter procedure is like giving a French test to those taking French and also to those taking Spanish, in order to render the sample size "fair."

Forgione has attempted to minimize the importance of the calculus items in the test: "In *none* of the countries were students chosen on the basis of whether they had taken calculus." (Forgione, 1998, p. 770). This is true but does not bear on the issue. Given that European nations have placed calculus in their secondary curricula for many years and given that many European nations tested only those students in programs focused on math and science, one can assume that the students tested in other nations had taken calculus. It remains that the U.S. chose to have 50% of its test takers sit for an exam where a quarter of the items covered material that they had not studied.

. . . the problems [described here] introduce far too much doubt about the validity of the information to draw any conclusions from the Final Year Study.

Finally, in regard to whether or not students in other countries had taken calculus, the TIMSS test-curriculum matching studies found that only a few countries ruled out even as many as 10% of the items as not being represented by their curriculum. Had the students in other nations not taken calculus, many more items would have been judged as not represented in the curriculum.

The test-curriculum matching analysis, incidentally, was not conducted in the United States. U.S. TIMSS officials at all grades decided to accept all items to see, as it was explained to me at a conference, "how the kids could handle them."

Only the United States and Austria accepted all 82 items of the advanced mathematics test as being addressed by their curricula.[9] Other countries judged between 62 and 81 items as congruent, with an average of 72. The United States and Austria also accepted all of the items on the physics test as addressed by their curricula. The 10 other nations accepted between 38 and 78, with an average of 66. The scores reported in the Final Year Report, though, are based on total scores. For the record, the United States also accepted all items in mathematics and science at Grades 4 and 8 as addressed by its curricula. It was usually the only country to do so.

The Structure of Curriculum

Interpretations of the Final Year Study are further complicated by differences in secondary curricula. In most developed nations, virtually all age-eligible students are in school through the eighth grade and are receiving largely the same curricula. Once "high school" begins, however, this is no longer true. As the Final Year Report acknowledges in its opening paragraphs and then shows in its vignettes of secondary school in the participating countries, these countries have marked differences in the structure of secondary education. The publication, *National Contexts for Mathematics and Science Education* reveals additional differences (Robitaille, 1997).

For example, Norway tested students in the final year of a 3-year physics course. Sweden tested students in the final year of the Natural Science or Technology "lines" of the secondary curriculum who had likewise studied physics for several years. Most likely, that is why these two countries topped all others in physics by a considerable margin.

In a number of countries, the students tested were enrolled in programs that focus largely on math and science and that appear to be available only to a small elite. This produced some odd-looking results. For instance, in mathematics, Cyprus fell farther and farther behind in the rankings with each grade tested: Cyprus ranked 18th of 26 nations at Grade 4, 37th of 41 at Grade 8, and 20th of 21 on the Final Year math/science literacy assessment. Yet Cyprus was 6th of 16 nations in advanced math and 8th of 16 in physics. Within the areas of the advanced math test, Cyprus was the number one country in the world in calculus. One wonders what a nation does to produce a group of Final Year students that are nearly last in overall rankings but high within specialty areas. In part, Cyprus tested only those students enrolled in a 3-year program offering a math and science curriculum, 9% of the age cohort. Cyprus was not, however, one of the nations testing older students.

Cultural Variables

It is very difficult to quantify cultural variables that might produce differences in outcomes. The TIMSS Final Year Study did measure one cultural variable, the difference between the United States and other nations in the role jobs play in teenagers' lives. In most nations, a teenager is either a student or a worker, not both.[10] In the United States, many teenagers hold part-time jobs. Some research has indicated that students who work up to 20 hours a week actually have higher school achievement than those who don't work or those who work longer hours (D'Amico, 1984; Gottfredson, 1985; Schulenberg & Bachman, 1993).

The TIMSS Final Year data confirm this curvilinear relation for the United States. American students who worked fewer than 21 hours during the school week scored higher than both those who worked 0–4 hours a school week and those who worked more than 21 hours in a school week. Those working up to 15 hours a week actually scored above the international average in the math/science literacy assessment. However, they constitute only a small minority of American students in the sample, 7%. In total, 55% of American seniors reported that they worked 21 hours a week or more. Those working fewer than 15 hours a week during the school week averaged 508 (international average = 500); for the 28% working 21–35 hours a week, the mean was 474; and for the 27% working more than 35 hours a week, the mean was 448.

From these data alone, one cannot rule out the possibility that longer hours are worked by low-income students out of a need to help support their families. Since students from low-income families score lower on tests than middle-class and affluent students, the lower scores for workers with long hours in TIMSS might be due to factors associated with class. On the other hand, Schulenberg and Bachman (1993) controlled for parental education level as a proxy for afflu-

ence and found that education level had little correlation with hours worked. The usual interpretation of the curvilinear relationship is that working teaches time management, responsibility, and the value of educational credentials in the real world. After 20 hours, though, students start to get too little sleep, eat poorly, and slack off homework.

Pursuing Excellence claimed that work at a job did not affect U.S. performance *relative to other countries*, a claim reiterated by Forgione. Obviously, however, it affected scores *within* this country relative to hours worked. In any case, the pattern in most nations looks more like this one given for Sweden: Less than 1 hour a week, 84%, 563; 5–10 hours, 7%, 506; 15–25 hours, 5%, 474; more than 25 hours, 3%, 424. Given that the United States students are virtually unique in their dual roles, it is not clear what kind of statistic would be appropriate for a comparison with other nations.

Finally, although socioeconomic factors cannot be ruled out in the lower scores of those who work long hours, we can note that socioeconomic factors cannot account for all of the impact. Or, if they do, if all 55% of those who work long hours came from low-income families, then the U.S. TIMSS Final Year test-takers formed a most unrepresentative sample.

A Few Miscellaneous Concerns

1. *The motivation issue.* Archie Lapointe, former director of the National Assessment of Educational Progress (NAEP) has commented that the big problem with NAEP is keeping kids awake during the test (Lapointe, personal communication, June 1995). NAEP is analogous to TIMSS in that it has no importance to the students and in that no one—students, parents, teachers, administrators, school board members—received any feedback from it.

More importantly, the TIMSS Final Year tests occurred in April and May.[11] American students know their post–high school futures by then. For those going to college, the SAT or ACT is long past and admissions decisions have arrived. This would compound the tendency not to take a test seriously. For all nations, the test was administered equally late in the school year, but there is no data concerning any international differences in test-taking motivation among students this close to school termination. One *Washington Post* story on the "senior slump" seen in the U.S. began this way: "Donny Watkins, a slender 18-year-old, wants to spend this spring doing what most American high school seniors do in their last semester. Next to nothing" (Mathews, 1998, p. B1). A more recent *Post* article suggests that a growing number of seniors shut down academically as soon as the results of their early-admissions applications are known (Mathews, 2000, p. A1).

2. *The variability issue.* The variabilities for many countries are as large as that for the United States. This surprised some observers who contend that, given our diversity and extremes of poverty, one ought to see larger variability here than elsewhere. However, the variabilities reported in the TIMSS Final Year Report are *student* variabilities. A better variance indicator might be the variability of *schools*.

Merging data from the Second International Assessment of Educational Progress with data from NAEP reporting categories, Bracey (1997, pp. 83–84) found that the top third of American schools had eighth-grade mathematics scores as high as the top two countries in the study, Taiwan and Korea. The bottom third of American schools, however, had scores below that of the lowest-scoring nation, Jordan. Most countries do not have the extremes of poverty that affect not only individuals but large portions of whole cities and rural areas. It would be interesting to see how the nations compare in terms of the school variability but this data cannot be derived from the study.

3. *The physics test.* There would appear to be something peculiar about a test where only the two highest scoring nations got substantially more than 50% of the items right—two nations whose students had studied physics for 3 years. The next two highest nations got barely half of the items right, 42 of 81. In the advanced mathematics test, 10 of the 16 participating nations had more than half the items correct. The reliability coefficients for the physics test are not impressive, reaching .80 in only one nation and falling as low as .49 with a median of .70 and a mean of .67.

TIMSS in general is a rich source of information in comparison to earlier international comparisons. The videotapes, curriculum analyses, case studies, and other forthcoming aspects of TIMSS provide much food for thought and many pointers for the reform of mathematics education. The test data, at all grades, much less so. The test data from the Final Year component least of all. While these data might survive any one of the problems noted here, taken together, the problems introduce far too much doubt about the validity of the information to draw any conclusions from the Final Year Study.

Notes

[1] There are also test scores for Grades 3 and 7, but little has been made of them in comparison to the results from Grades 4 and 8.

[2] Twenty-four nations participated in some aspect of the 12th-grade assessment. Twenty-one nations posted results for the math/science literacy test, while 16 tested students in physics and advanced mathematics.

[3] "Some" because the number cannot be determined precisely from Fairfield documents and TIMSS officials do not release data disaggregated below the state level.

[4] Under the "rules" for participating, any nation taking part in TIMSS was required to test eighth graders, resulting in 41 nations at this grade compared to 26 at Grade 4 and between 16 and 21 (depending on the test) in the Final Year.

[5] For reasons not clear, the TIMSS Final Year Report fails to note some violations and shows 10 of the 16 nations meeting the criteria, while *Pursuing Excellence* notes that only 5 do so.

[6] Some countries violated these criteria at the fourth- and eighth-grade assessments. However, if these countries are deleted from the study, one still has international comparisons for 16 nations at Grade 4 and 26 at Grade 8.

[7] The statistic cited does not appear in the 1997 edition.

[8] I say "some 100 points" because the exact figure cannot be determined from the reports. *Pursuing Excellence* (NCES, 1998) shows that students with calculus scored 50 points higher than the total American sample. Since *Pursuing Excellence* also shows that these students constitute 50% of the total, it follows that the pre-calculus students as a group scored about 100 points lower than the calculus students.

[9] Technically, there were 82 "score points"—a few items had more than one part.

[10] The only other countries that have substantial proportions of teenagers working are the other Anglophone nations in the study. Berliner (1999) has contended that this is a legacy of British mercantilism, in which the exploitation of children is acceptable.

[11] The test was administered in May in the U.S. In all nations, it was given as close to the end of the final year as possible.

References

Beaton, A., (1998, April 7). Remarks made at "Heads in the Sand," a Brookings Institution Panel Discussion, Washington, DC.

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1996). *Science achievement in the middle school years.* Chestnut Hill, MA: Boston College.

Beaton, A. E., Mullis, I. V. S, Martin, M. O., Gonzalez, E., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years.* Chestnut Hill, MA: Boston College.

Berliner, D. C. (1999, December). How business hurts education: A minority viewpoint. Presentation at Florida International University.

Bracey, G. W. (1997). *Setting the record straight: Responses to misconceptions about public education in the United States.* Alexandria, VA: Association for Supervision and Curriculum Development.

Bronner, E. (1998, February 21). U.S. 12th graders rank poorly in math and science, study says. *New York Times,* p. A1.

D'Amico, R. (1984). Does employment during school impair academic progress? *Sociology of Education, 57*(3), 152–164.

Dobson, J. (1998, April). *Family news from Dr. James Dobson.* Colorado Springs: Focus on the Family.

Forgione, P. (1998). Responses to frequently asked questions about 12th-grade TIMSS. *Phi Delta Kappan, 79*(10), 769–772. These remarks were originally distributed at "Heads in the Sand," a Brookings Institution Panel Discussion, Washington, DC, April 7, 1998.

Gottfriedson, D. (1985). Youth employment, crime, and schooling: A longitudinal study of a national sample. *Developmental Psychology, 21,* 419–432.

Henry, T. (1998, February 25). Poor academic showing hurts U.S. high schoolers. *USA Today,* p. 1A.

Leo, J. (1998, March 9) Hey, We're #19! *U.S. News and World Report,* p. 14.

Martin, M. O., Mullis, I. V. S., Beaton, A. E., Gonzalez, E. J., Smith, T. A., & Kelly, D. A. (1997). *Science achievement in the primary school years.* Chestnut Hill, MA: Boston College.

Mathews, J. (1998, May 27). Schools treat "senior slump." *Washington Post,* p. B1.

Mathews, J. (2000, February 23). Early admission, later cost? Many seniors take it too easy, educators say. *Washington Post,* p. A1.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1997). *Mathematics achievement in the primary school years.* Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1998). *Mathematics and science achievement in the final year of secondary school.* Chestnut Hill, MA: Boston College.

National Center for Education Statistics. (1996). *The condition of education 1996.* Washington, DC: Author.

National Center for Education Statistics. (1998). *Pursuing excellence: A study of twelfth grade mathematics and science achievement in international context.* Washington, DC: Author.

Organization for Economic Cooperation and Development. (1996). *Education at a glance.* Paris: Author.

Organization for Economic Cooperation and Development. (1997). *Education at a glance.* Paris: Author.

Raspberry, W. (1998, March 6). The good news about U.S. schools. *Washington Post*, p. A25.

Robitaille, D. F. (Ed.). (1997). *National contexts for mathematics and science education.* Vancouver, British Columbia: Pacific Educational Press.

Sanchez, R. (1998, February 25). U.S. high school seniors rank near bottom. *Washington Post*, p. A1.

Schmidt, W. (1998, April 7). Remarks made at "Heads in the Sand," a Brookings Institution Discussion Panel, Washington, DC.

Schulenberg, J., & Bachman, J. G. (1993, March). Long hours on the job? Not so bad for some adolescents in some types of jobs: The quality of work and substance use, affect, and stress. Paper presented at the biennial meeting of the Society for Research in Child Development, New Orleans, LA.

Valverde, G. A. (1998, April). TIMSS high school results released. Report No. 8, U.S. National Research Center, Michigan State University, p. 4.

Viadero, D. (1998, March 4) U.S. seniors near bottom in world test. *Education Week*, p. 1.

Manuscript received May 15, 1998
Revision received September 15, 1998
Accepted March 20, 2000