How Do American Students Measure Up? Making Sense of International Comparisons

How Do American Students Measure Up? Making Sense of International Comparisons

Daniel Koretz

Summary

In response to frequent news media reports about how poorly American students fare compared with their peers abroad, Daniel Koretz takes a close look at what these comparisons say, and do not say, about the achievement of U.S. high school students. He stresses that the comparisons do not provide what many observers of education would like: unambiguous information about the effectiveness of American high schools compared with those in other nations.

Koretz begins by describing the two principal international student comparisons—the Trends in International Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA). Both assessments, he stresses, reflect the performance of students several years before they complete high school. PISA, which targets fifteen-year-old students, measures students' abilities to apply what they have learned in school to real-world problems. By contrast, TIMSS tests fourth and eighth graders. Unlike PISA, TIMSS follows the school curriculum closely.

Because the findings of the two tests are sometimes inconsistent, Koretz stresses the importance of considering data from both sources. He cautions against comparing U.S. students with an "international average," which varies widely from survey to survey depending on which countries participate, and recommends instead comparing them with students in other nations that are similar to the United States or that are particularly high-achieving.

Many observers, says Koretz, speculate that the lackluster average performance of American students in international comparisons arises because many, especially minority and low-income U.S. students, attend low-performing schools. But both TIMSS and PISA, he says, show that the performance of American students on the exams is not much more variable than that of students in countries that are socially more homogeneous or that have more equitable educational systems.

Koretz emphasizes that the international comparisons provide valuable information and are a useful source of hypotheses about American secondary schooling to be tested by researchers. Studies designed to explain differences between U.S. students and those in very similar countries, he says, might provide especially useful suggestions for changes in policy and practice.

www.futureofchildren.org

Daniel Koretz is a professor at the Harvard Graduate School of Education.

One reason for the widespread dissatisfaction with American secondary schools is the view that U.S. students perform poorly compared with their peers in other nations. For years, the drumbeat of bad news from international student comparisons has been unrelenting. In 1983, *A Nation at Risk*, the report that did much to spur ongoing efforts to reform American education, stressed the weak performance of U.S. students compared with students abroad, and negative international comparisons have been a staple of public debate about American education ever since.[1] International comparisons of student achievement are now carried out frequently, and newspapers never fail to highlight their disappointing results. A comment a few years ago in *Education Week*, the leading trade paper in K–12 education, is typical: "In their most recent lackluster showing on the world stage, students in the United States scored below average in mathematics literacy and problem-solving in an international comparison of the academic skills of teenagers in developed nations."[2]

The data, however, are more limited and more complex than is often realized, and the story they properly tell is not quite so straightforward. The results that receive the most attention—the simple ranking of countries in terms of their average mathematics achievement—are less clear-cut than most observers think. Moreover, the data include useful information beyond the horse race that gets little or no attention, some of which flies in the face of common expectations. Data about student performance at the end of high school are scarce and especially hard to collect and interpret. International comparisons of student achievement are valuable, but they cannot provide a clear evaluation of the performance of American high schools.

In this article I explore what these international comparisons do accomplish. I begin by describing the available data from the two principal international student surveys. After raising several cautions and offering advice about how to interpret the data, I describe some key findings of the international assessments. Finally, I discuss their implications.

What Are the Data?

Just whom is the press talking about when it reports gloomy news about the comparative achievement of American students? The great bulk of the news reflects two ongoing international surveys: the Trends in International Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA).[3] The two are cited interchangeably in the press, but they are quite different, and their results are sometimes different as well—occasionally strikingly so.

International comparisons of student achievement are valuable, but they cannot provide a clear evaluation of the performance of American high schools.

PISA's target population is a single group: fifteen-year-old students attending educational institutions, including part-time students. Thus, in most instances, PISA tests students near but not yet at the final grade of secondary schooling. Individuals not being schooled or being schooled at home, in the workplace, or out of the country are intentionally excluded.[4] The survey, which is repeated

every several years, assesses mathematics, science, and reading. The PISA assessment is intended to measure students' abilities to apply what they have learned in school to real-world problems. For that reason, the framework from which the PISA tests are constructed does not closely mirror school curricula. The PISA tests are organized by broad themes, such as "change and growth," rather than curricular areas, such as geometry, and some of the test items are intended to look rather different from what one might find in a typical curriculum-based test.

TIMSS differs from PISA in all of these respects. TIMSS samples students by grade, not by age. In its first iteration, TIMSS surveyed three groups: fourth-grade students, eighth-grade students, and students at the end of secondary school.[5] Defining comparable groups at the end of high school is a daunting task because of the great variation across countries in the structure of secondary schooling. The sample for this part of TIMSS in 1995 was described as follows:

> The intention of the assessment of final-year students was to measure what might be considered the "yield" of the elementary and secondary education systems of a country with regard to mathematics and science. The international desired population, then, was all students in the final year of secondary school. Students repeating the final year were not part of the desired population. For each secondary education track in a country, the final grade of the track was identified as being part of the target population, allowing substantial coverage of students in their final year of schooling. For example, grade 10 could be the final year of a vocational program, and grade 12 the final year of an academic program. Both of these grade/track combinations are considered to be part of the population [but grade 10 in the academic track is not].[6]

As in the PISA surveys, out-of-school youth were not sampled. This complex sampling makes comparisons among countries extremely hard to interpret. Perhaps for that reason, this end-of-school component has not been repeated in subsequent TIMSS surveys.

TIMSS differs from PISA also in the characteristics of its assessment. TIMSS is intended to follow school curricula reasonably closely. For this reason, the content of the test, the mix of items across content areas, and even the characteristics of the items themselves differ appreciably from those of PISA. For example, in recent tests, TIMSS devoted 25 percent of its items to algebra, while PISA allocated 11 percent. As I shall show, these differences do matter, and they pose a challenge for people using the results.

Interpreting International Comparisons: Some Essential Cautions

Some years ago, the U.S. Department of Education offered the following summary of the performance of American students on the first (1995) TIMSS survey: "On the eighth-grade TIMSS assessment, U.S. students scored somewhat above the international average in science and somewhat below average in mathematics."[7] As the quotation from *Education Week* with which I began this article suggests, similar statements comparing U.S. students with an international average have been common.

Comparisons with an "international average," however, are nearly meaningless. An average is useful if it represents a clear comparison group. For example, telling a parent that her fourth-grade child scores below the average of all fourth graders in the state is useful information. However, in the case of international comparisons, the "international

mean" reflects the collection of countries that happened to participate in a given assessment in a given year. That happenstance group is not always a sensible comparison, and it changes over time, moving the average up or down considerably. For example, in 1999, a year after the statement above was published, the TIMSS mathematics assessment was administered again. In the main presentation of the results, the United States was shown as scoring above the "international average" of countries that participated in that year. A few pages later, however, the report showed the United States scoring well below the average of a different group of countries, those that that had participated in both 1995 and 1999.[8] This problem is easily avoided. American performance should not be compared with a slippery "international average," but with the performance of other countries that provide an informative contrast. For example, it is useful to compare the United States with the nations that consistently perform best, such as Japan and Singapore, as well as with nations that are in many respects more similar, such as Australia and Canada. These comparisons are generally stable over time, although they are not always consistent from one survey to another, for example from TIMSS to PISA.

A second complication in interpreting international comparisons involves differences among assessments. International assessments measure very broad domains of achievement, such as the cumulative mastery of mathematics over the first eight years of schooling. All tests of broad domains use a relatively small number of test items to estimate mastery of the entire domain, most of which remains untested. In this respect, tests function much like political polls, which use the views of a few people to predict the voting behavior of a much larger group of people, most of whom are not surveyed. The fact that tests are only small samples of performance has many important implications, one of which is that different tests sample somewhat differently from the domain and therefore may yield different views of performance. These variations may not indicate that something has gone wrong with one of the tests, although they may.[9]

Some little-noticed results from TIMSS illustrate the importance of decisions about sampling content. The eighth-grade mathematics assessment comprises five content areas, such as algebra and data representation. Some nations perform appreciably better in some of these areas than in others. For example, the United States and Australia performed more poorly in geometry than in the other four areas, while Singapore performed markedly better in fractions and numbers than in the others.[10] As a result, the rankings of countries that are reported by the press can be modified, although not dramatically, simply by changing the relative emphasis given to the five content areas in that particular test.[11] Larger differences among tests, such as some of those between TIMSS and PISA, can be expected to have even larger effects.

And indeed, in some cases, the results of PISA and TIMSS differ substantially. For example, in recent TIMSS and PISA assessments of mathematics, Scotland, New Zealand, and Norway ranked considerably better on the PISA assessment than on TIMSS; the Netherlands, Hong Kong, South Korea, and the United States had quite similar ranks on both tests; and Russia and Hungary ranked much higher on the TIMSS assessment than on PISA.[12] In the 2003 TIMSS assessment of eighth-grade mathematics, Norway scored far below the United States. In the PISA assessment of the same year, Norway outscored the

United States, not by much, but by enough that the difference was statistically significant (that is, unlikely to have arisen simply by chance because of sampling students).

It is not hard to find "explanations" of these differences, but in fact, the explanations remain speculative. Because of the design of the two assessments, it is not possible to explain differences between their results with confidence. These disparities may reflect intentional differences in content, differences in sampling of students, or unintentional factors that have not yet been identified. Nonetheless, they pose a problem for users of the results. For example, how does the mathematics performance of American students compare with that of students in Norway? That question has no single answer, though some other important patterns in the findings are consistent.

The inconsistencies are no reason to put international comparisons aside. They are simply reason to be careful in interpreting the results. Taking a few precautions can help to interpret the results sensibly. The first is to pay little attention to small differences, because these are particularly likely to depend on relatively unimportant aspects of test design. Careful readers of the reports of TIMSS and PISA will see that the authors specify which differences between nations are statistically significant so that readers can ignore differences that are statistically untrustworthy. However, statistical significance tells one only that a given difference was unlikely to have arisen by chance as a result of the sampling of schools and students. It does not indicate how robust the difference would be to reasonable changes in test design.[13] Therefore, it is wise to ignore small differences even when they are statistically significant.

The second precaution is to be wary of relying on the results of a single assessment. No one test, however well designed, should be treated as a "gold standard." An equally good test, designed differently, will often yield modestly different findings and occasionally markedly different findings. When a finding appears in more than one assessment—particularly, assessments that are quite different, as PISA and TIMSS are—then one can have more confidence that the result is not caused simply by the particular choices made in designing a specific test. For example, in mathematics, the United States has always scored far below the developed countries in East Asia—Japan, Korea, Singapore, and Hong Kong. Although the precise size of these differences will vary from test to test—indeed, they differ between TIMSS and PISA—it is a safe bet that other tests of similar domains would also show the United States well behind these countries.

A final complication, for those interested in the performance of students at the end of high school, is that this group is especially difficult to compare across countries. As noted, the one, highly complex attempt by TIMSS to compare performance at the end of school has not been repeated. But comparisons would be hard to interpret even if more data were available. One reason is youth who have left school, either because of completion (in countries where mandatory schooling ends at younger ages) or because of dropping out. The portion of the cohort that leaves school early varies both in size and in characteristics from one country to another. Leaving them out of an assessment can badly bias international comparisons. Including them, however, would be difficult and expensive and would require different sampling methods than those used for in-school youth.

Differences in school leaving were a major reason why the sole TIMSS study of students at the end of high school was problematic. TIMSS reported a "coverage index," which was the percentage of the school-leaving age cohort tested. For the survey as a whole, the best coverage was in Norway and France, where 84 percent of the age cohort was tested. In the United States, 63 percent was tested; in Italy, 52 percent; and in a few countries, 10 percent or less. A comparison involving 84 percent of Norwegian youth and 52 percent of Italian youth is hard to interpret. For a separate comparison of students taking advanced mathematics and physics, the differences were starker yet; for example, this comparison included 86 percent of youth in Slovenia but 4 percent in the Russian Federation and 22 percent in the United States.[14] Moreover, the majority of participating countries, including the United States, failed to meet TIMSS standards for the minimum quality of the sample, which required following specified guidelines for recruiting the sample and meeting specified criteria for participation rates and coverage of the population.[15]

A second reason why comparisons at the high school level are problematic is curricular differentiation—the routing of students into dissimilar instructional programs. In some countries, such as Germany and the Netherlands, students are sorted into various types of secondary schools that differ in selectivity, curriculum, and, in some cases, length of schooling. These differences were another factor that led to the extremely complex design of the single TIMSS study of the final year of schooling. In other countries, such as the United States, most students attend comprehensive high schools, but curricular differentiation within them is typically substantial, particularly in subjects such as mathematics that are important for admission to selective colleges and universities.

Curricular differentiation is problematic because having useful comparisons across countries in a broad subject area, such as mathematics, requires agreement about the goals of mathematics instruction. To the degree that countries, or educational tracks within a country, differ in their goals, students in countries or tracks whose goals align well with the test will score higher than others. TIMSS approaches this problem by looking for common elements in curricula, but the greater the differences in curricula, the less tenable this approach is, and the more sensitive comparisons will be to the particular makeup of the test. PISA addresses this issue by focusing on application of skills beyond the school context, but even that strategy does not entirely solve the problem. Students are in different instructional tracks in part because the goals for their later use of mathematics differ.

One precaution is to be wary of relying on the results of a single assessment. No one test, however well designed, should be treated as a "gold standard."

As a result, almost all discussion in the press about international differences in the achievement of secondary school students focuses on younger students, either those in middle school (the TIMSS survey) or those

aged fifteen (the PISA survey). These choices do not eliminate the problems above, but they do substantially lessen them.

Key Findings of the International Assessments

International comparisons have been conducted in numerous subjects, but those in mathematics and science have received the most attention. Here I focus primarily on mathematics but describe briefly the results in science and reading.

Performance in Mathematics

The focus of this volume is high schools, so ideally the most relevant of the international studies would be the TIMSS comparison of students at the end of high school. The news from that study, if it were taken at face value, would be distressing. The mean score for U.S. students on a composite of mathematics and science was fourth from the bottom of twenty-one participating countries. The U.S. mean was statistically significantly higher than those of only Cyprus and South Africa, although it was statistically not reliably different from those of numerous other countries, including the Russian Federation and the Czech Republic.[16] Given the concerns about the TIMSS noted above, however—the difficulty of defining a reasonable international comparison at the end of high school, the problem of curricular differentiation, the large disparities in the percentages of youth tested, and the failure of most participating countries to meet minimum standards for the quality of their tested samples—the results are not readily interpreted. In the absence of a solid basis for directly comparing performance at the end of high school, it is necessary to rely on data from earlier in students' secondary education.

Mathematics results from earlier stages of secondary schooling (from students aged fifteen in the PISA assessments and students in eighth grade in the TIMSS assessments) are not as bleak, but they appear discouraging enough and have dominated discussion in this country for decades. The consistent finding has been that American secondary school students perform less well in mathematics than their peers in many other countries that might be considered either similar or competitors. And this finding can be trusted: it has appeared time after time, in a number of different assessments.[17]

But just how badly do U.S. students perform, and how comparable in this respect are the findings from the two main ongoing assessments, PISA and TIMSS? To answer these questions adequately takes a bit more work because the scales reported—like those of most large-scale assessments—are arbitrary. Is a 20-point difference on the TIMSS scale large enough to worry about, and is it similar in magnitude to a 20-point difference on the PISA scale? (To illustrate this point with a more familiar example, consider the two competing college admissions tests, the SAT and ACT. In a single subject, the SAT scale runs from 200 to 800, while the ACT scale runs from 1 to 36. A 20-point difference on the SAT is trivial, while on the ACT, it is enormous.) A few calculations are needed to make the results of TIMSS and PISA comparable and more easily interpreted.

A common way to solve this problem is to standardize the scores. When scores are standardized, the average score is given a value of zero, and other scores are given values that express how far above or below the mean they are. These distances are expressed in terms of standard deviations, a common measure of the spread of scores. Although unfamiliar to many people, standardized scales have many advantages. They have the

same meaning from one test to the next, and they can be readily converted to other forms. If the distribution of scores roughly follows the bell curve, as scores on most large-scale assessments like TIMSS do, then a score that is one standard deviation above the average (a standardized score of +1) will be roughly at the 84th percentile: roughly 84 percent of students will have scores below this point. Conversely, a standardized score of –1 will fall roughly at the 16th percentile rank: 16 percent will score below –1, and 84 percent above.[18]

When one first glances at the results of the 2003 TIMSS and PISA assessments, the United States seems to fare somewhat worse in the latter: it is further down in the distribution of country means in PISA.[19] But for the same reason that a comparison with an "international average" is not meaningful, comparisons with the entire groups of countries that happened to participate in the two assessments are not particularly useful either. More informative are comparisons with specific countries, and these show important differences between the two assessments.

In the TIMSS assessments, the nations that score the highest in mathematics are always developed countries in East Asia: Japan, Korea, Hong Kong, Taiwan, and especially Singapore. The average difference between these countries and the United States varies a bit from one TIMSS assessment to the next, but it is always large, typically roughly a full standard deviation. In 2003, the gaps ranged from approximately 0.8 standard deviation (Japan) to almost 1.3 standard deviations (Singapore)—meaning that only one in five American students scored above the Japanese average and only one in ten scored above the Singaporean average.

In TIMSS, the United States fares better when compared with European countries and with Australia and New Zealand. In 2003, the highest-scoring European countries were the Netherlands and Belgium, whose averages were well below those of the Asian nations and roughly 0.4 standard deviation above that of the United States (meaning that about a third of U.S. students scored above the averages in those two countries). Some countries that are in many respects similar to the United States—England, Scotland, Sweden, and Australia—had average scores very similar to that of the United States. Norway's average was more than half a standard deviation lower than the American average.

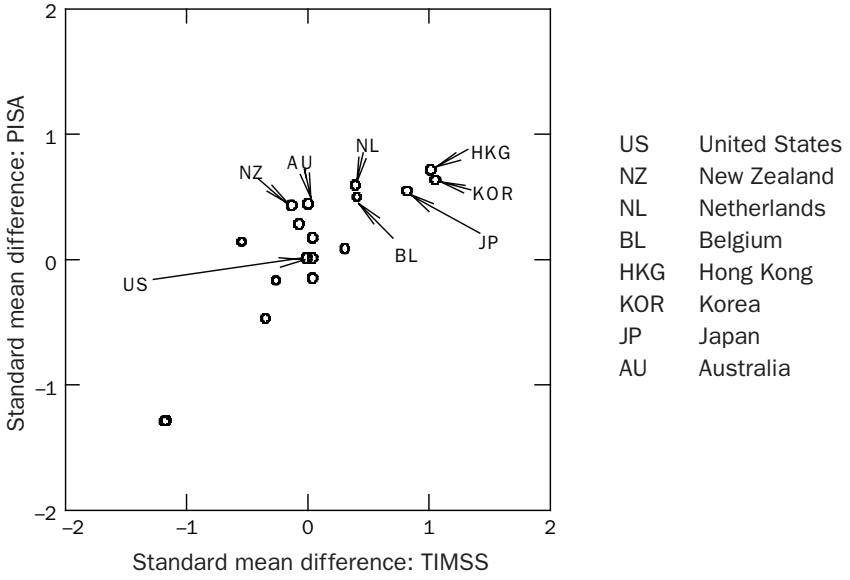
These findings might lead one to the generalization that the United States is far behind East Asian nations but roughly comparable in performance to numerous countries that are more similar, with a few exceptions (high-scoring Holland and low-scoring Norway). If only it were that simple. PISA paints a somewhat different picture, underscoring the risk of placing too much faith in the results of a single assessment.

While confirming that U.S. students do not perform well compared with their peers in many other nations, the results of PISA differ from those of TIMSS in two respects. In one respect, PISA is less discouraging: the United States is not as far behind the highest-ranking countries. In other respects, it is more discouraging: the rankings of countries are somewhat different, and the generalization that the United States performs roughly as well as more comparable nations does not hold up. These differences can be seen by examining the performance of the eighteen countries that participated in both assessments in 2003.

The smaller size of the performance gap between the United States and the top-scoring countries in PISA can be seen by considering countries like Korea and Hong Kong. In figure 1, each circle represents one country's standardized average score. The distribution of scores in the United States was used to standardize scores, so the U.S. average on both tests is zero, and the scores of the other countries represent the fraction of a standard deviation between their averages and the U.S. average. The TIMSS results are arrayed on the horizontal axis, and PISA scores are on the vertical axis. The average score for Korea appears on the far right of the chart because, of the eighteen countries that participated in both assessments, Korea was the highest-scoring on TIMSS. One can see from the figure that the Korean average on TIMSS was a bit more than one standard deviation higher than that of the United States (to be more precise, about 1.1 standard deviations). However, looking at the vertical axis, one can see that the gap between Korea and the United States in PISA was considerably smaller (about 0.6 standard deviation). Hong Kong shows much the same pattern. Among the eighteen countries, the largest gap with the United States was 1.1 standard deviations in TIMSS and 0.7 standard deviation in PISA.

The more discouraging and more obvious disparity in the results of the two assessments is the performance of the highest-scoring western countries. All of them fell substantially short of the top East Asian countries in TIMSS, but a number of them—Australia, New Zealand, the Netherlands, and Belgium —performed roughly similarly to the Asian countries in PISA and therefore did

Figure 1. Standardized Mean Differences between the United States and Eighteen Other Countries in Mathematics, TIMSS and PISA Assessments, 2003



Sources: I. V. Mullis and others, *TIMSS 2003 International Mathematics Report* (Chestnut Hill, Mass.: International Study Center, Boston College, 2004), exhibit 1.1; Organisation for Economic Co-operation and Development, *Learning for Tomorrow's World: First Results from PISA 2003* (Paris: OECD, 2004), figure 2.15a.

considerably better than the United States. (In figure 1, these countries are therefore to the left of the Asian countries but roughly as far up on the vertical axis.) Two of them, Australia and New Zealand, scored almost the same as the United States in TIMSS.

Why are the results of PISA and TIMSS so dissimilar? It would be tempting to attribute the differences to intentional differences in the design of the assessments, such as PISA's greater focus on applications and TIMSS's greater alignment with school curricula. For example, one might conclude that East Asian countries do better than the highest-performing European countries in giving students mastery of the mathematics curriculum but not in teaching them to apply these skills in real-world contexts. However, as satisfying as such an explanation might be, it remains speculative. Although both PISA and TIMSS are carefully designed to serve their primary purposes, they are not well designed to answer questions such as this. The tests cannot be linked, so one cannot say for certain that differences in content account for the disparities in results. And more generally, these studies are well designed to describe differences among countries, but not to explain them. In the language of social science, they are well suited to generating hypotheses but not to testing them.

One often hears speculation that the lackluster average performance of American students in international comparisons arises because many U.S. students attend low-performing schools. That there are huge disparities in performance—and severe inequities in resources—among American schools is unarguable. Many observers therefore assume that the performance of American students is much more variable than that of students in countries that are socially more homogeneous or that have more equitable educational systems. Therefore, the argument goes, these low-performing students pull down the American average, and international comparisons of higher-achieving students would look different.

One often hears speculation that the lackluster average performance of American students in international comparisons arises because many U.S. students attend low-performing schools.

The argument turns out not to be true: the variability of the performance of U.S. students is unexceptional, and the mediocre achievement of the United States is found across the entire range of performance. To see this, one needs to examine information on the variability of student performance. All of the major reports of both TIMSS and PISA include several indicators of this variation, including standard deviations for each country and performance at a variety of percentiles, but this information has been largely overlooked in the frenzy of attention given to the horse race—that is, the rankings of country averages. Both assessments show that the variability of student performance is reasonably similar among the counties that participated in the studies. And both show that the standard deviation of the scores of American students is well within the typical range. For example, among the eighteen nations participating in both TIMSS and PISA in 2003, almost all had standard deviations between 81 and 101 scale score points. The average

standard deviation was 93. The standard deviation of the scores of American students was 95 points.

Does this mean that inequities have no effect? Hardly. The explanation for this puzzle is a counterintuitive rule in statistics: when scores are highly variable *within* groups, even large differences *between* groups have relatively little impact on the *total* variability of scores. Years ago, to illustrate this principle, I analyzed eighth-grade mathematics and reading scores from two nationally representative American samples, the National Education Longitudinal Study and the National Assessment of Educational Progress. I posed the question this way: if the achievement gap vanished and all of the reported racial and ethnic groups performed exactly like non-Hispanic whites, how much would the total variation (specifically, the national standard deviation) shrink? Very little. Across the four cases, the answer ranged from less than 1 percent to 9 percent. As a determinant of the total variation in scores, the huge variability within each racial and ethnic group swamps the very large mean differences between the groups.

Performance in Science

Both PISA and TIMSS regularly assess performance in science, though these comparisons are somewhat less clear than those in mathematics. Because science curricula vary widely from one country to the next, it is both harder to design a comparative assessment and more difficult to interpret its findings. For this reason, it should not be surprising that the findings in science have been less consistent than those in mathematics.

In PISA, the performance of U.S. fifteen-year-olds in science is similar to their performance in mathematics: mediocre. The U.S. average is far below the averages of the highest-scoring countries, and it is not only East Asian countries that dominate the list. The highest-scoring group includes Finland (by a substantial margin, the best), Hong Kong, Canada, Taiwan, New Zealand, and Australia. The U.S. average is roughly similar to that of Norway, Spain, and Iceland. Many nations that we would consider somewhat comparable, such as the United Kingdom and Germany, are arrayed in between.[20]

TIMSS provides a much more positive portrayal of American eighth graders' performance in science. As in mathematics, most of the highest-scoring countries were East Asian, although Estonia ranked with Japan. However, the United States and a number of other Western countries scored quite well, only a modest distance below some of the East Asian countries. That high-scoring group of Western nations also included the Netherlands, Australia, and Sweden.[21]

In science as in mathematics, one can only speculate about the reasons for the different views provided by TIMSS and PISA. The answer could lie in the nature of the tested material, the nature of the samples (PISA students are older), or incidental characteristics of the studies. However, given the problem of curricular differences in science, it remains a plausible hypothesis that differences in tested content played an important role.

Performance in Reading

Although they have received far less attention in the United States, a number of international studies have compared proficiency in reading. Although reading is not a primary focus of instruction in secondary schools, the reading proficiency of secondary school students—and their proficiency when they enter secondary school—is certainly important.

Two studies, one dated, have shown that the reading proficiency of U.S. students in elementary school is very good by international standards.[22] The limited comparative data about the reading proficiency of secondary school students, while less positive, is still reasonably encouraging. The older of the studies noted above tested middle school students and found that their performance, while relatively speaking not as strong as that of elementary school students tested in the same study, was nonetheless reasonably strong by international standards. More recently, PISA has found much the same thing about the performance of U.S. fifteen-year-olds. The 2000 PISA assessment found that U.S. reading proficiency was very similar to that of many countries we might consider reasonable comparisons (such as Denmark, Switzerland, France, Norway, and Belgium); modestly better than that of some others (Germany, Hungary); but not as strong as that of Finland, Canada, or New Zealand.[23]

Discussion

International comparisons clearly do not provide what many observers of education would like: unambiguous information about the effectiveness of American high schools compared with those in other nations. Most of the data reflect the performance of students years before they complete high school. The findings are in some cases inconsistent from one study to another. Moreover, the data from all of these studies are poorly suited to separating the effects of schooling from the myriad other influences on student achievement. There is no reason to believe that if one dropped students from the United States into schools in Singapore, their performance would match that of Singaporean students, or vice versa, even if one adjusted for the limited range of other factors about which data were collected in these studies.

Despite these limitations, the data can be informative. Used sensibly, they provide us with very valuable descriptive information and a unique basis for generating hypotheses about American secondary schooling. For example, the educational systems of some nations that score particularly well in these surveys differ from systems in the United States in a variety of ways, including governance, curricula, instructional methods, approaches to testing and accountability, and recruiting of teachers. TIMSS and PISA cannot tell us which, if any, of these factors contribute to the stronger performance of these nations, but they provide us with many suggestions that can be tested with more appropriate study designs. Studies designed to explain differences between countries that are socially and culturally similar might provide especially useful suggestions for changes in policy and practice.

As noted, obtaining trustworthy and useful comparisons requires some care. First, one should ignore small differences among countries, as they are too likely to be the result of sampling or unimportant characteristics of the tests. Second, one should ignore the "international average" and select other nations that provide informative comparisons, such as those that are similar or that are particularly high-achieving. Third, when possible, one should consider data from more than one source.

Following these guidelines leads to some important conclusions. For example, TIMSS, considered alone, suggests that the performance of U.S. students is fairly similar to that of students in many similar countries. Adding data from PISA, however, shows that in some other respects, U.S. students fall well behind those in some of those same nations, such as Australia. Both TIMSS and PISA

show that the variability of performance is not anomalously large in the United States, which is contrary to common expectations. This suggests that efforts to lessen educational inequities and other sources of undesirable variation in achievement still must accommodate a very wide range of student performance.

Of course, these conclusions do not reflect performance at the end of schooling. If truly comparable data from the end of schooling were available, they would presumably look somewhat different, though it is unlikely that they would be greatly more optimistic. Other data do not suggest that the final few years of high school in the United States are substantially more effective than schooling in the lower grades, and in recent years achievement has improved less in high school than in elementary and middle school. The National Assessment of Educational Progress has for years shown marked improvements in mathematics in grade four, substantial but somewhat slower improvements in grade eight, but only slow and erratic gains in grade twelve.

In sum, the international comparisons now available do not provide us with a straightforward evaluation of either U.S. secondary schools or the policies that govern them. They do, however, provide rich descriptive information about the performance of our students and a unique opportunity to appraise its adequacy in comparison to that of their peers in competing nations. These studies also provide us with numerous hypotheses about factors that may impede performance or that may be useful in improving it. To evaluate these hypotheses will require other types of data and evidence, some of which are discussed elsewhere in this volume.

Endnotes

1. National Commission on Excellence in Education, *A Nation at Risk* (Washington: U.S. Department of Education, April 1983).

2. S. Cavanagh and E. W. Robelen, "U.S. Students Fare Poorly in International Math Comparison," *Education Week,* December 7, 2004, www.edweek.org [August 20, 2008].

3. TIMSS initially stood for the "Third International Mathematics and Science Study," a reference to two earlier, related studies of mathematics conducted in the 1960s and 1970s. When a decision was made to repeat TIMSS at regular intervals, the name was changed to "Trends in International Mathematics and Science Study."

4. R. Adams and M. Wu, eds., *PISA 2000 Technical Report* (Paris: Organisation for Economic Co-operation and Development, 2002).

5. The TIMSS sample is actually a bit more complex than this. TIMSS surveys students in the two grades that include the largest proportions of nine- and thirteen-year-olds. In most countries, that means grades three and four and seven and eight, but the primary reporting has usually been of the older of the two grades in each pair.

6. I. V. Mullis and others, *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study* (TIMSS) (Chestnut Hill, Mass.: TIMSS International Study Center, Boston College, 1998).

7. United States Department of Education, *Policy Brief: What the TIMSS Means for Systemic School Improvement* (Washington: November 1998). Archived information available at www.ed.gov/pubs/TIMSS Brief/student.html [June 2, 2002].

8. I. V. Mullis and others, *TIMSS 1999 International Mathematics Report* (Chestnut Hill, Mass.: International Study Center, Boston College, 2002).

9. For a non-technical explanation of this principle and more discussion of its implications, see D. Koretz, *Measuring Up: What Educational Testing Really Tells Us* (Harvard University Press, 2008).

10. Mullis and others, *TIMSS 1999 International Mathematics Report* (see note 8).

11. R. G. Wolfe, "Country-by-Item Interactions: Problems with Content Validity in Scaling," presented at a symposium on "Validity in Cross-National Assessments: Problems and Pitfalls," annual meeting of the American Educational Research Association, May 27, 1997, Chicago.

12. L. S. Gronmo and R. V. Olsen, "TIMSS versus PISA: The Case of Pure and Applied Mathematics," paper presented at the Second IEA International Research Conference, November 8–11, 2006, Washington, D.C.

13. Estimates of statistical significance in the reports of both TIMSS and PISA take into account measurement error as well as sampling error. However, these estimates of measurement error take the design of the test as a given and do not reflect changes in performance that would arise from altering it.

14. Mullis and others, *Mathematics and Science Achievement in the Final Year of Secondary School* (see note 6).

15. Ibid.

16. Ibid.

17. A. E. Lapointe, N. A. Mead, and J. M. Askew, *Learning Mathematics* (Princeton, N.J.: Educational Testing Service, 1992). A. E. Lapointe, N. A. Mead, and G. W. Phillips, *A World of Differences: An International Assessment of Mathematics and Science* (Report No. 19-CAEP-01)(Princeton, N.J.: Educational Testing Service, 1989). A third study with the same finding was the International Assessment of Educational Progress, an extension of the U.S. National Assessment of Educational Progress, conducted in 1988 and 1991, but this study is now rarely noted.

18. In the case of international comparisons, standardization is not entirely straightforward. The standard deviation of the tested population and the average country-level standard deviation are functions of the group of countries tested and are therefore not useful for this purpose. The estimates of within-country standard deviations are not highly stable. The following text uses the 2003 estimates of the U.S. standard deviation. As explained below, choosing another nation's standard deviation would not have greatly changed the results. For a non-technical explanation of standardized score scales and their application to international comparisons and trends in the United States, see Koretz, *Measuring Up* (see note 9).

19. Compare I. V. Mullis and others, *TIMSS 2003 International Mathematics Report* (Chestnut Hill, Mass.: International Study Center, Boston College, 2004), exhibit 1.1, with Organisation for Economic Co-operation and Development, *Learning for Tomorrow's World: First Results from PISA 2003* (Paris: OECD, 2004), figure 2.15a.

20. Ibid.

21. Mullis and others, *TIMSS 2003 International Mathematics Report* (see note 19).

22. W. B. Elley, *How in the World Do Children Read?* (The Hague: International Association for the Evaluation of Educational Achievement, 1992); I. V. Mullis and others, *PIRLS 2001 International Report* (Chestnut Hill, Mass.: International Study Center, Boston College, 2003).

23. National Center for Education Statistics, *Highlights from the 2000 Program for International Student Assessment* (PISA) (Washington: Office of Educational Research and Improvement, U.S. Department of Education, 2002).