WHAT DO INTERNATIONAL TESTS REALLY SHOW ABOUT U.S. STUDENT PERFORMANCE?

BY MARTIN CARNOY, STANFORD GRADUATE SCHOOL OF EDUCATION AND EPI
AND RICHARD ROTHSTEIN, EPI
Executive summary

Education policymakers and analysts express great concern about the performance of U.S. students on international tests. Education reformers frequently invoke the relatively poor performance of U.S. students to justify school policy changes.

In December 2012, the International Association for the Evaluation of Educational Achievement (IEA) released national average results from the 2011 administration of the Trends in International Mathematics and Science Study (TIMSS). U.S. Secretary of Education Arne Duncan promptly issued a press release calling the results "unacceptable," saying that they "underscore the urgency of accelerating achievement in secondary school and the need to close large and persistent achievement gaps," and calling particular attention to the fact that the 8th-grade scores in mathematics for U.S. students failed to improve since the previous administration of the TIMSS.

Two years earlier, the Organization for Economic Cooperation and Development (OECD) released results from another international test, the 2009 administration of the Program for International Student Assessment (PISA). Secretary Duncan's statement was similar. The results, he said, "show that American students are poorly prepared to compete in today's knowledge economy. ... Americans need to wake up to this educational reality—instead of napping at the wheel while emerging competitors prepare their students for economic leadership." In particular, Duncan stressed results for disadvantaged U.S. students: "As disturbing as these national trends are for America, enormous achievement gaps among black and Hispanic students portend even more trouble for the U.S. in the years ahead."

However, conclusions like these, which are often drawn from international test comparisons, are oversimplified, frequently exaggerated, and misleading. They ignore the complexity of test results and may lead policymakers to pursue inappropriate and even harmful reforms.

TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Executive summary</td>
<td>2</td>
</tr>
<tr>
<td>Part I. Introduction</td>
<td>6</td>
</tr>
<tr>
<td>Part II. PISA 2009—the comparative performance of U.S. students by social class group</td>
<td>9</td>
</tr>
<tr>
<td>Part III. PISA trends from 2000 to 2009</td>
<td>23</td>
</tr>
<tr>
<td>Reading, 2000–2009</td>
<td>25</td>
</tr>
<tr>
<td>Mathematics, 2000–2009</td>
<td>31</td>
</tr>
<tr>
<td>Part IV: Defining social class for comparative purposes</td>
<td>40</td>
</tr>
<tr>
<td>Part V. Comparing PISA and TIMSS results in mathematics</td>
<td>42</td>
</tr>
<tr>
<td>Part VI. Comparing NAEP, PISA, and TIMSS trends</td>
<td>56</td>
</tr>
<tr>
<td>Part VII. Population and curricular sampling issues</td>
<td>66</td>
</tr>
<tr>
<td>Population sampling flaws</td>
<td>69</td>
</tr>
<tr>
<td>Population sampling inconsistency between tests</td>
<td>72</td>
</tr>
<tr>
<td>Decisions about curricular sampling</td>
<td>73</td>
</tr>
<tr>
<td>Assessment by age or by grade</td>
<td>77</td>
</tr>
<tr>
<td>Part VIII. Discussion</td>
<td>79</td>
</tr>
<tr>
<td>Part IX. Conclusion</td>
<td>82</td>
</tr>
<tr>
<td>Appendix A</td>
<td>83</td>
</tr>
<tr>
<td>Appendix B</td>
<td>86</td>
</tr>
<tr>
<td>Books in the home (BH) and the Economic, Social, and Cultural Status (ESCS) indices</td>
<td>86</td>
</tr>
<tr>
<td>Mother's education, parents' education, ESCS, and books in the home as correlates of students' test scores</td>
<td>86</td>
</tr>
<tr>
<td>Endnotes</td>
<td>92</td>
</tr>
<tr>
<td>References</td>
<td>97</td>
</tr>
</tbody>
</table>

Both TIMSS and PISA eventually released not only the average national scores on their tests but also a rich international database from which analysts can disaggregate test scores by students' social and economic characteristics, their school composition, and other informative criteria. Such analysis can lead to very different and more nuanced conclusions than those suggested from average
national scores alone. For some reason, however, although TIMSS released its average national results in December, it scheduled release of the international database for five weeks later. This puzzling strategy ensured that policymakers and commentators would draw quick and perhaps misleading interpretations from the results. This is especially the case because analysis of the international database takes time, and headlines from the initial release are likely to be shaped in conventional wisdom by the time scholars have had the opportunity to complete a careful study.

While we await the release of the TIMSS international database, this report describes a detailed analysis we have conducted of the 2009 PISA database. It offers a different picture of the 2009 PISA results than the one suggested by Secretary Duncan's reaction to the average national scores of the United States and other nations.

Because of the complexity and size of the PISA international database, this report's analysis is restricted to the comparative test performance of adolescents in the United States, in three top-scoring countries, and in three other post-industrial countries similar to the United States. These countries are illustrative of those with which the United States is usually compared. We compare the performance of adolescents in these seven countries who have similar social class characteristics. We compare performance in the most recent test for which data are available, as well as trends in performance over the last nearly two decades.

In general, we find that test data are too complex and oversimplified to permit meaningful policy conclusions regarding U.S. educational performance without deeper study of test results and methodology. However, a clear set of findings stands out and is supported by all data we have available:

**Because social class inequality is greater in the United States than in any of the countries with which we can reasonably be compared, the relative performance of U.S. adolescents is better than it appears when countries' national average performance is conventionally compared.**

- Because in every country, students at the bottom of the social class distribution perform worse than students higher in that distribution, U.S. average performance appears to be relatively low partly because we have so many more test takers from the bottom of the social class distribution.

- A sampling error in the U.S. administration of the most recent international (PISA) test resulted in students from the most disadvantaged schools being over-represented in the overall U.S. test-taker sample. This error further depressed the reported average U.S. test score.

- If U.S. adolescents had a social class distribution that was similar to the distribution in countries to which the United States is frequently compared, average reading scores in the United States would be higher than average reading scores in the similar post-industrial countries we examined (France, Germany, and the United Kingdom), and average math scores in the United States would be about the same as average math scores in similar post-industrial countries.

- A re-estimated U.S. average PISA score that adjusted for a student population in the United States that is more disadvantaged than populations in otherwise similar post-industrial countries, and for the oversampling of students from the most-disadvantaged schools in a recent U.S. international assessment sample, finds that the U.S. average score in both reading and mathematics would be higher than official reports indicate (in the case of mathematics, substantially higher).

- This re-estimate would also improve the U.S. place in the international ranking of all OECD countries, bringing the U.S. average score to sixth in reading and 13th in math. Conventional ranking reports based
on PISA, which make no adjustments for social class composition or for sampling errors, and which rank countries irrespective of whether score differences are large enough to be meaningful, report that the U.S. average score is 14th in reading and 25th in math.

- Disadvantaged and lower-middle-class U.S. students perform better (and in most cases, substantially better) than comparable students in similar post-industrial countries in reading. In math, disadvantaged and lower-middle-class U.S. students perform about the same as comparable students in similar post-industrial countries.

- At all points in the social class distribution, U.S. students perform worse, and in many cases substantially worse, than students in a group of top-scoring countries (Canada, Finland, and Korea). Although controlling for social class distribution would narrow the difference in average scores between these countries and the United States, it would not eliminate it.

- U.S. students from disadvantaged social class backgrounds perform better relative to their social class peers in the three similar post-industrial countries than advantaged U.S. students perform relative to their social class peers. But U.S. students from advantaged social class backgrounds perform better relative to their social class peers in the top-scoring countries of Finland and Canada than disadvantaged U.S. students perform relative to their social class peers.

- On average, and for almost every social class group, U.S. students do relatively better in reading than in math, compared to students in both the top-scoring and the similar post-industrial countries.

Because not only educational effectiveness but also countries' social class composition changes over time, comparisons of test score trends over time by social class group provide more useful information to policymakers than comparisons of total average test scores at one point in time or even of changes in total average test scores over time.

- The performance of the lowest social class U.S. students has been improving over time, while the performance of such students in both top-scoring and similar post-industrial countries has been falling.

- Over time, in some middle and advantaged social class groups where U.S. performance has not improved, comparable social class groups in some top-scoring and similar post-industrial countries have had declines in performance.

Performance levels and trends in Germany are an exception to the trends just described. Average math scores in Germany would still be higher than average U.S. math scores, even after standardizing for a similar social class distribution. Although the performance of disadvantaged students in the two countries is about the same, lower-middle-class students in Germany perform substantially better than comparable social class U.S. students. Over time, scores of German adolescents from all social class groups have been improving, and at a faster rate than U.S. improvement, even for social class groups and subjects where U.S. performance has also been improving. But the causes of German improvement (concentrated among immigrants and perhaps also attributable to East and West German integration) may be idiosyncratic, and without lessons for other countries or predictive of the future. Whether German rates of improvement can be sustained to the point where that country's scores by social class group uniformly exceed those of the United States remains to be seen. As of 2009, this was not the case.

Great policy attention in recent years has been focused on the high average performance of adolescents in Finland. This attention may be justified, because both math and reading scores in Finland are higher for every social class group than in the United States. However, Finland's scores have been falling for the most disadvantaged students while U.S. scores have been improving for similar
social class students. This should lead to greater caution in applying presumed lessons from Finland. At first glance, it may seem that the decline in scores of disadvantaged students in Finland results in part from a recent influx of lower-class immigrants. However, average scores for all social class groups have been falling in Finland, and the gap in scores between Finland and the United States has narrowed in each social class group. Further, during the same period in which scores for the lowest social class group have declined, the share of all Finnish students in this group has also declined, which should have made the national challenge of educating the lowest social class students more manageable, so immigration is unlikely to provide much of the explanation for declining performance.

Although this report’s primary focus is on reading and mathematics performance on PISA, it also examines mathematics test score performance in earlier administrations of the TIMSS. Where relevant, we also discuss what can already be learned from the limited information now available from the 2011 TIMSS. To help with the interpretation of these PISA and TIMSS data, we also explore reading and mathematics performance on two forms of the U.S. domestic National Assessment of Educational Progress (NAEP).

Relevant complexities are too often ignored when policymakers draw conclusions from international comparisons. Different international tests yield different rankings among countries and over time. PISA, TIMSS, and NAEP all purport to reflect the achievement of adolescents in mathematics (and PISA and NAEP in reading), yet results on different tests can vary greatly—in the most extreme cases, countries’ scores can go up on one test and down on another that purport to assess the same students in the same subject matter—and scholars have not investigated what causes such discrepancies. These differences can be caused by the content of the tests themselves (for example, differences in the specific skills that test makers consider to represent adolescent “mathematics”) or by flaws in sampling and test administration. Because these differences are revealed in the most cursory examination of test results, policymakers should exercise greater caution in drawing policy conclusions from international score comparisons.

To arrive at our conclusions, we made a number of explicit and transparent methodological decisions that reflect our best judgment. Three are of importance: our definition of social class groups, our selection of comparison countries, and our determination of when differences in test scores are meaningful.

There is no clear way to divide test takers from different countries into social class groups that reflect comparable social background characteristics relevant to academic performance. For this report, we chose differences in the number of books in adolescents’ homes to distinguish them by social class group; we consider that children in different countries have similar social class backgrounds if their homes have similar numbers of books. We think that this indicator of household literacy is plausibly relevant to student academic performance, and it has been used frequently for this purpose by social scientists. We show in a technical appendix that supplementing it with other plausible measures (mother’s educational level, and an index of “economic, social, and cultural status” created by PISA’s statisticians) does not provide better estimates. Also influencing our decision is that the number of books in the home is a social class measure common to both PISA and TIMSS, so its use permits us to explore longer trend lines and more international comparisons. As noted, however, data on these background characteristics were not released along with the national average scores on the 2011 TIMSS, and so our information on the performance of students from different social class groups on TIMSS must end with the previous, 2007, test administration.

In this report, we focus particularly on comparisons of U.S. performance in math and reading in PISA with performance in three “top-scoring countries” (Canada, Finland, and Korea) whose average scores are generally higher
than U.S. scores, and with performance in three “similar post-industrial countries” (France, Germany, and the United Kingdom) whose scores are generally similar to those of the United States. We employed no sophisticated statistical methodology to identify these six comparison countries. Assembling and disaggregating data for this report was time consuming, and we were not able to consider additional countries. We think our choices include countries to which the United States is commonly compared, and we are reasonably confident that adding other countries would not appreciably change our conclusions. If other scholars wish to develop data for other countries, we would gladly offer them methodological advice.

Technical reports on test scores typically distinguish differences that are “significant” from those that are not. But this distinction is not always useful for policy purposes and is frequently misunderstood by policymakers. To a technical expert, a score difference can be miniscule but still “significant” if it can be reproduced 95 percent of the time when a comparison is repeated. But miniscule score differences should be of little interest to policymakers. In general, social scientists consider an intervention to be worthwhile if it improves a median subject’s performance enough to be superior to the performance of about 57 percent or more of all subjects prior to the intervention. Such an intervention should be considered “significant” for policy purposes, but, to avoid confusion, we avoid the term “significant” altogether. Instead, for PISA, we consider countries’ (or social class groups’) average scores to be “about the same” if they are less than 8 test scale points different (even if this small difference would be repeated in 95 of 100 test administrations), to be “better” or “worse” if they are at least 8 but less than 18 scale points different, and “substantially better” or “substantially worse” if they differ by 18 scale points or more. Eighteen scale points in most cases is approximately equivalent to the difference social scientists generally consider to be the minimum result of a worthwhile intervention (an effect size of about 0.2 standard deviations). The TIMSS scale is slightly different from the PISA scale; for TIMSS, the cut points used in this report are 7 and 17 rather than 8 and 18.

With regard to these and other methodological decisions we have made, scholars and policymakers may choose different approaches. We are only certain of this: To make judgments only on the basis of statistically significant differences in national average scores, on only one test, at only one point in time, without regard to social class context or curricular or population sampling methodologies, is the worst possible choice. But, unfortunately, this is how most policymakers and analysts approach the field.

The most recent test for which an international database is presently available is PISA, administered in 2009. As noted, the database for TIMSS 2011 is scheduled for release later this month (January 2013). In December 2013, PISA will announce results and make data available from its 2012 test administration. Scholars will then be able to dig into TIMSS 2011 and PISA 2012 databases and place the publicly promoted average national results in proper context. The analyses that follow in this report should caution policymakers to await understanding of this context before drawing conclusions about lessons from TIMSS or PISA assessments. We plan to conduct our own analyses of these data when they become available, and publish supplements to this report as soon as it is practical to do so, given the care that should be taken with these complex databases.

Part I. Introduction

A 2009 international test of reading and math showed that American 15-year-olds perform more poorly, on average, than 15-year-olds in many other countries. This finding, from the Program for International Student Assessment (PISA), is consistent with previous PISA results, as well as with results from another international assessment of 8th-graders, the Trends in International Mathematics and Science Survey (TIMSS).
From such tests, many journalists and policymakers have concluded that American student achievement lags woefully behind that in many comparable industrialized nations, that this shortcoming threatens the nation’s economic future, and that these test results therefore suggest an urgent need for radical school reform.

Upon release of the 2011 TIMSS results, for example, U.S. Secretary of Education Arne Duncan called them “unacceptable,” saying that they “underscore the urgency of accelerating achievement in secondary school and the need to close large and persistent achievement gaps” (Duncan 2012). Two years before, upon release of 2009 PISA scores, Duncan said that “…the 2009 PISA results show that American students are poorly prepared to compete in today’s knowledge economy. … Americans need to wake up to this educational reality—instead of napping at the wheel while emerging competitors prepare their students for economic leadership.” In particular, Duncan stressed the PISA results for disadvantaged U.S. students: “As disturbing as these national trends are for America, enormous achievement gaps among black and Hispanic students portend even more trouble for the U.S. in the years ahead. Last year, McKinsey & Company released an analysis which concluded that America’s failure to close achievement gaps had imposed—and here I quote—‘the economic equivalent of a permanent national recession.’” The PISA results, Duncan concluded, justify the reform policies he has been pursuing: “I was struck by the convergence between the practices of high-performing countries and many of the reforms that state and local leaders have pursued in the last two years” (Duncan 2010).

This conclusion, however, is oversimplified, exaggerated, and misleading. It ignores the complexity of the content of test results and may well be leading policymakers to pursue inappropriate and even harmful reforms that change aspects of the U.S. education system that may be working well and neglect aspects that may be working poorly.

For example, as Secretary Duncan said, U.S. educational reform policy is motivated by a belief that the U.S. educational system is particularly failing disadvantaged children. Yet an analysis of international test score levels and trends shows that in important ways disadvantaged U.S. children perform better, relative to children in comparable nations, than do middle-class and advantaged children. More careful analysis of these levels and trends may lead policymakers to reconsider their assumption that almost all improvement efforts should be directed to the education of disadvantaged children and few such efforts to the education of middle-class and advantaged children.

Education analysts in the United States pay close attention to the level and trends of test scores disaggregated by socioeconomic groupings. Indeed, a central element of U.S. domestic education policy is the requirement that average scores be reported separately for racial and ethnic groups and for children who are from families whose incomes are low enough to qualify for the subsidized lunch program. We understand that a school with high proportions of disadvantaged children may be able to produce great “value-added” for its pupils, although its average test score levels may be low. It would be foolish to fail to apply this same understanding to comparisons of international test scores.

Extensive educational research in the United States has demonstrated that students’ family and community characteristics powerfully influence their school performance. Children whose parents read to them at home, whose health is good and can attend school regularly, who do not live in fear of crime and violence, who enjoy stable housing and continuous school attendance, whose parents’ regular employment creates security, who are exposed to museums, libraries, music and art lessons, who travel outside their immediate neighborhoods, and who are surrounded by adults who model high educational achievement and attainment will, on average, achieve at higher levels than children without these educationally relevant advantages. We know much less about the extent to which
similar factors affect achievement in other countries, but we should assume, in the absence of evidence to the contrary, that they do.

It is also the case that countries’ educational effectiveness and their social class composition change over time. Consequently, comparisons of test score trends over time by social class group provide more useful information to policymakers than comparisons of total average test scores at one point in time or even of changes in total average test scores over time.

Unfortunately, our conversation about international test score comparisons has ignored such questions. It would be foolish, for example, to let international comparisons motivate radical changes in educational policies in a country whose social class subgroup average scores were below those of other nations, if that country’s subgroups had been improving their performance at a more rapid rate than similar subgroups in other nations, even if the country’s overall average still had not caught up. Just as a domestic U.S. school’s average performance is influenced by its social class composition, so too might a country’s average performance be influenced by its social class composition.

The policy responses of educational reformers should be sufficiently nuanced to respond to such considerations, because policy initiatives might improve in response to more sophisticated inquiries.

For example, consider Country C. Its affluent students achieve better than affluent students in comparable countries, but not as much better as in the past; the performance of affluent students in Country C, while still relatively high, has been declining relative to the performance of affluent students in comparable countries. Country C’s socioeconomically disadvantaged students achieve less than disadvantaged children in comparable countries, but not as much less as in the past. The performance of disadvantaged students in Country C, while still relatively low, has been improving relative to the performance of disadvantaged students in comparable nations. In such circumstances, unsophisticated reformers in Country C might well decide to revamp how disadvantaged students are being taught, even though teaching methods have been successfully raising such students’ achievement relative to the achievement of similarly disadvantaged students in other countries and relative to the achievement of wealthier students in Country C itself. Such unsophisticated reformers might also ignore the condition of education of affluent students, believing that their relatively high performance suggests that no reform is needed, while overlooking the decline of such performance over time. Sophisticated education policymakers, in contrast, who have studied the data trends, might direct their reform efforts to the high-scoring rather than the low-scoring students.

Thus, in evaluating a country’s educational performance, we should want to know how children from different social class groups perform, in comparison to other social class groups within their own country and in comparison to children from similar social class groups in other countries. Describing only an “average” national score obscures what is likely to be more useful information. Yet it is only in terms of national averages that policy discussion of international test scores typically proceeds. U.S. policymakers would learn more if they also studied the performances of demographic (socioeconomic) subgroups and compared these to the performances of similar subgroups in other nations. To the extent international comparisons are important, it is critical to know whether each subgroup in the United States performs above or below the level of socioeconomically similar subgroups in comparable industrialized nations.

If we identify subgroups that perform relatively well or relatively poorly in one country or another, we should also ask how the performances of these subgroups, compared to the performances of similar subgroups in other nations, are changing over time. Are some subgroups improving their performance unusually rapidly, in comparison to socioeconomically similar subgroups in other nations,
while other subgroups are exhibiting unusual deterioration in performance? Are various subgroups improving or declining in performance at different rates, and are these differences masked when we look only at national averages?

In this report, we also identify inconsistencies between various international tests that may well be related to inaccurate population sampling that has caused some tests to oversample some social class groups and undersample others. Such sampling errors inevitably lead to inaccuracies in reports of how students in a particular country perform, relative to those in other countries where the sampling may have been more accurate.

Other considerations, rarely considered in public debate, also influence the care we should take in the interpretation of international comparisons. One is how the curriculum is sampled in the framework for any particular test. Because the full range of knowledge and skills that we describe as “mathematics” cannot possibly be covered in a single brief test, policymakers should also carefully examine whether an assessment called a “mathematics” test necessarily covers knowledge and skills similar to those covered by other assessments also called “mathematics” tests, and whether performance on these different assessments can reasonably be compared. For example, American adolescents perform relatively well on algebra questions, and relatively poorly on geometry questions, compared to adolescents in other countries. Reports on how the United States compares to other countries show the United States in a more favorable light to the extent a test has more algebra items and fewer geometry items. Whether there is an appropriate balance between these topics on any particular international assessment is rarely considered by policymakers who draw conclusions about the relative performance of U.S. students from that assessment. Similar questions arise with regard to a “reading” test.

Whether U.S. policymakers want to reorient the curriculum to place more emphasis on geometry is a decision they should make without regard to whether such reorientation might influence comparative scores on an international test. It certainly might not be good public policy to reduce curricular emphasis on statistics and probability, skills essential to an educated citizenry in a democracy, in order to make more time available for geometry. There are undoubtedly other sub-skills covered by international reading and math tests on which some countries are relatively stronger and others are relatively weaker. Investigation of these differences should be undertaken before drawing policy conclusions from international test scores.

To stimulate an examination and discussion of these and several other complexities, we analyze data on the performance of adolescents from PISA and TIMSS, as well as from two forms of the National Assessment of Educational Progress (NAEP), a test given exclusively to a sample of U.S. students. The first form, Main NAEP, is modified in small ways over time, so that its coverage tracks modifications in the math curriculum. The second form, Long-Term Trend NAEP (LTT), which changes much less over time, assesses how students’ competence changes over time on a more nearly identical set of skills. The Main NAEP has been administered since 1990, and the LTT since the early 1970s.  

Part II. PISA 2009—the comparative performance of U.S. students by social class group

Disaggregation of PISA test scores by social class group reveals some patterns that many education policymakers will find surprising. Average U.S. test scores are lower than average scores in countries to which the United States is frequently compared, in part because the share of disadvantaged students in the overall national population is greater in the United States than in comparison countries. If the social class distribution of the United States were similar to that of top-scoring countries, the average test score gap between the United States and these top-scoring countries would be cut in half in reading and by one-third in mathematics. Dis-
advantaged U.S. students perform comparatively better than do disadvantaged students in important comparison countries. The test score gap between advantaged and disadvantaged students in the United States is smaller than the gap in similar post-industrial countries; it is generally, although not always, greater than the gap in top-scoring countries. This section explores these findings in greater detail.

To simplify our comparisons of national average PISA scores and of these scores disaggregated by social class, we focus on the United States and six other countries—Canada, Finland, South Korea (hereinafter simply Korea), France, Germany, and the United Kingdom.

We refer to three of these countries (Canada, Finland, and Korea) as “top-scoring countries” because they score much better overall than the United States in reading and math—about a third of a standard deviation better. Canada, Finland, and Korea are also the three “consistent high-performers” that U.S. Secretary of Education Arne Duncan highlighted when he released the U.S. PISA results (Duncan 2010).

We call the other three (France, Germany, and the United Kingdom) “similar post-industrial countries” because they score similarly overall to the United States. They also are countries whose firms are major competitors of U.S. firms in the production of higher-end manufactured goods and services for world markets. Their firms are not the only competitors of U.S. firms, but if the educational preparation of young workers is a factor in national firms’ competitiveness, it is worth comparing student performance in these countries with student performance in the United States to see if these countries’ educational systems, so different from that in the United States, play a role in their firms’ success.

PISA is scored on a scale that covers very wide ranges of ability in math and reading. When scales were created for reading in 2000 and for math in 2003, the mean for all test takers from countries in the Organization for Economic Cooperation and Development (OECD), the sponsor of PISA, was set at 500 with a standard deviation of 100. When statisticians describe score comparisons, they generally talk about differences that are “significant.” Yet while “significance” is a useful term for technical discussion, it can be misleading for policy purposes, because a difference can be statistically significant but too small to influence policy. Therefore, in this report, we avoid describing differences in terms of statistical significance. Instead, we use terms like “better (or worse)” and “substantially better (or worse)” (both of which are significantly better for statistical purposes), and “about the same.”

In general, in this report, we use the term “about the same” to describe average score differences in PISA that are less than 8 scale points; we use the term “better (or worse)” to describe differences that are at least 8 points but less than 18 scale points, and we use the term “substantially (or much) better (or worse)” to describe differences that are 18 scale points or more. Of course, any fixed cut point is arbitrary, and readers may find it strange when we say, for example, that when two countries have an average difference of 7 scale points they perform about the same, whereas when their average difference is 8 scale points one performs better than the other. This is a necessary consequence of any descriptive system using cut points. However, this caution is in order: Readers without statistical sophistication will be tempted to think that a difference of 7 scale points is almost “better.” This is true. But a difference of 8 scale points is also almost “about the same.” Many readers, accustomed to finding differences where there are none, will be more reluctant to consider the latter than the former, but both are equally true.

Table 1 displays overall average scores in reading and math reported by PISA for 2009. These are the basis (without any socioeconomic disaggregation) of most commonplace comparisons.

The table shows that, on average, U.S. performance was substantially worse than performance in the top-scoring
TABLE 1

Overall average national scale scores, reading and math, for U.S. and six comparison countries, PISA 2009

<table>
<thead>
<tr>
<th>TOP SCORING</th>
<th>SIMILAR POST-INDUSTRIAL</th>
<th>U.S.</th>
<th>U.S. VERSUS:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Canada</td>
<td>Finland</td>
<td>Korea</td>
</tr>
<tr>
<td>Reading</td>
<td>524</td>
<td>536</td>
<td>539</td>
</tr>
<tr>
<td>Math</td>
<td>527</td>
<td>541</td>
<td>546</td>
</tr>
</tbody>
</table>

* Simple (unweighted) average of three countries
Source: Authors' analysis of OECD Program for International Student Assessment (PISA) (2010a)

countries in both math and reading, was about the same
as performance in the similar post-industrial countries in
reading, and was worse than performance in the similar
post-industrial countries in math.

We next disaggregate scores in the United States and in
the six comparison countries by an approximation of the
social class status of test takers, dividing them into six
groups, from the least to the most advantaged. We refer
to these as Group 1 (lowest social class), 2 (lower social
class), 3 (lower-middle social class), 4 (upper-middle
social class), 5 (higher social class), and 6 (highest social
class). We also refer to Groups 1 and 2 together as disad-
tantaged students, to Groups 3 and 4 together as middle-
class students, and to Groups 5 and 6 together as advant-
gaged students.

There is no precise way to make social class comparisons
between countries. PISA collects data on many charac-
teristics that are arguably related to social class status,
and also assembles them into an overall index. Although
none of the possible indicators of social class differences
is entirely satisfactory, we think one, the number of books
in the home (BH), is probably superior for purposes of
international test score comparisons, and we use it for
our analysis. A very high fraction of students in both
the PISA and TIMSS surveys answer the BH question,
something less true for other important social class indicator questions asked on the student questionnaires. As we
explain in greater detail below, we also examine whether
other social class indicators, such as mother's education or
PISA's overall index, in addition to BH, would produce
meaningfully different results, and determine that they
would not. We conclude that BH serves as a reasonable
representation of social class (home) influences on stu-
dents' academic performance.

Our examination of 2009 PISA scores, disaggregated by
social class group, reveals that:

- In every country, students from more-advantaged
  social class groups outperform students from less-
  advantaged social class groups. The social class
  performance gap is large. In each country we study, the
  reading gap between the highest (Group 6) and the
  lowest (Group 1) social class groups is more than a
  full standard deviation. The math gap is also more
  than a full standard deviation in the United States
  and in four of the six comparison countries. In the other
two, Canada and Finland, the gap is also large, almost
  a full standard deviation. The reading and math gaps
  are larger in France than in any country we studied.

- The reading and math gaps are smaller in the United
  States than in each of the three similar post-industrial
countries we studied.

- The average U.S. scores in reading and math were
  about the same or lower than those in the six compar-
  ison countries in considerable part because a dispro-
  portionately greater share of U.S. students come from
disadvantaged social class groups than do students in
the six comparison countries.
Table 2A displays the share by social class group of the national samples for the United States and the six comparison countries.

From these tables we can see that more U.S. 15-year-olds (37 percent) are in the disadvantaged (Groups 1 and 2) social class groups than in any of the six comparison countries, and we can therefore see why comparisons that do not control for differences in social class distributions between countries may differ greatly from those that do. There are fewer U.S. students in the middle (Groups 3 and 4) social class groups than in the middle social class groups of the three similar post-industrial countries (Germany, France, and the United Kingdom), although the differences are small. Differences in the size of middle-class groups are larger when the United States is compared to the three top-scoring countries (Korea, Finland, and Canada). And in the advantaged (Groups 5 and 6) social class groups there are substantially fewer U.S. students.
About TIMSS and PIRLS

For the past 20 years, TIMSS (Trends in International Mathematics and Science Study) has measured trends in mathematics and science achievement at the fourth and eighth grades. It has been conducted on a regular 4-year cycle since 1995, making TIMSS 2011 the fifth assessment of mathematics and science achievement trends. TIMSS Advanced, which measures trends in advanced mathematics and physics for students in their final year of secondary school, was conducted in 1995 and 2008, and is scheduled for 2015 (with the sixth assessment of TIMSS). For the past 15 years PIRLS (Progress in International Reading Literacy Study) has measured trends in reading comprehension at the fourth grade. First assessed in 2001, PIRLS has been on a regular 5-year cycle since then. Most recently, PIRLS was expanded in 2011 to include prePIRLS, which is a less difficult version of PIRLS. Both TIMSS and PIRLS were assessed in 2011, when the cycles of both studies came into alignment.

In general, participating countries use TIMSS and PIRLS in various ways to explore educational issues, including: monitoring system-level achievement trends in a global context, establishing achievement goals and standards for educational improvement, stimulating curriculum reform, improving teaching and learning through research and analysis of the data, conducting related studies (e.g. monitoring equity or assessing students in additional grades), and training researchers and teachers in assessment and evaluation.

TIMSS and PIRLS results are disseminated through reports and via the web through a well-documented international database for within and across country research.