AMERICAN INSTITUTES FOR RESEARCH®

Reassessing U.S. International Mathematics Performance: New Findings from the 2003 TIMSS and PISA

PREPARED FOR:
U.S. Department of Education
Policy and Program Studies Service (PPSS)

PREPARED BY:
American Institutes for Research®
1000 Thomas Jefferson Street, NW
Washington, DC 20007-3835

November, 2005

Reassessing U.S. International Mathematics Performance: New Findings from the 2003 TIMSS and PISA

Alan Ginsburg
U.S. Department of Education

Geneise Cooke
U.S. Department of Education

Steve Leinwand
American Institutes for Research

Jay Noell
U.S. Department of Education

Elizabeth Pollock
American Institutes for Research

This paper was supported by funds from the U.S. Department of Education and The Urban Institute. The paper does not necessarily represent the official positions of the U.S. Department of Education or The Urban Institute. The contents are the sole responsibility of the authors.

ACKNOWLEDGEMENTS

This report has been reviewed in draft form by individuals chosen for their diverse mathematics perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that assist our project team in making its published report as sound as possible and to ensure that the report is objective and responsive to the goals of our study. Their review raised some excellent methodological issues and suggested important analytical directions. Also, the reviewers suggested editorial changes that greatly assisted with the clarity of this report. We wish to thank the following individuals for their participation in the review of this report.

John Dossey, Illinois State University
Michael S. Garet, American Institutes for Research
Patsy Wang-Iverson, Research for Better Schools

TABLE OF CONTENTS

LIST OF EXHIBITS

Page

EXECUTIVE SUMMARY

In 2003, U.S. students' mathematics performance was assessed by the Trends in International Mathematics and Science Study (TIMSS) for students in grades 4 and 8 and by the Program for International Student Assessment (PISA) for students at age 15. Results in the press and later summaries, such as that found in The Condition of Education (NCES, 2005b) reported U.S. rankings in relation to all participating countries on each assessment. Because of the variability in the composition of countries participating in each assessment, these discussions have given an inaccurate impression that U.S. students' performance on PISA experienced a precipitous decline compared with favorable U.S. rankings on TIMSS at grades 4 and 8.

A total of 24 countries participated in TIMSS-4, 45 countries in TIMSS-8, and 40 countries in PISA. Notably, many higher-performing European countries that participated in PISA and contributed to the lower U.S. rankings were absent from the TIMSS results in which U.S. performance ranked above average. Reform proposals, such as those from the National Academy of Science and the Business Roundtable, have focused on improving mathematics at the secondary school level and may have been mislead by the differing country comparisons.

The current study reexamines the TIMSS and PISA results to correct this comparison bias by analyzing U.S. mathematics performance relative to a common set of 12 countries that participated in all three assessments. Along with the United States, the comparison countries are Australia, Belgium, Hong Kong, Hungary, Italy, Japan, Latvia, Netherlands, New Zealand, Norway, and the Russian Federation. These countries span the four continents of Australia/Oceana, Asia, North America and Europe and constitute a broad range of primarily industrialized nations (i.e., above the world average per capita income).

This study first examines the results for the United States compared with the results for the comparison countries. The country results are then used to explore various educational factors associated with international mathematics performance at different stages of students' mathematical development and through different country-specific characteristics.

U.S. results. Once the composition of countries participating in the three assessments is controlled, there is no evidence of a sharp decline on PISA compared with TIMSS, but a instead relative consistency of U.S. international performance. Within the common 12-country group:

- U.S. mathematics scores rank 8th on TIMSS-4; 9th on TIMSS-8; and 9th on PISA.

- On TIMSS-4, the scores of seven countries were statistically above the U.S. score and four were below. On TIMSS-8, the scores of five countries were statistically above the U.S. score and three were below. On PISA, six countries' scores statistically exceeded the U.S. score and three were below.

Like the United States, other countries generally show a consistent ranking across assessments. Overall, the correlation of scores between TIMSS-4 and TIMSS-8 is high (above .9, where a correlation of 1 indicates perfect correspondence). The correlation is lower between TIMSS-8 and PISA, although it is still a significant .67. This correlation would likely have been much higher if PISA had not differed from TIMSS in its sampling approach and its stress on items measuring real-world mathematics applications. We conclude that, in general, a country's initial grade 4 international performance is likely to be where that country ends up performing internationally for

15-year-olds. Thus, countries that want to improve their mathematics performance should start by building a strong mathematics foundation in the early grades.

Next we examine how the strengths and weaknesses of United States and other countries' international performance is related to the mathematical complexity of items, the mathematics content area assessed, and the gender of the test taker.

Mathematics rigor. Low-rigor items require students to use simple mathematical skills such as recalling mathematics definitions or applying mathematics formulas. High-rigor items require students to use cognitively-demanding skills, including reasoning through the information to formulate a problem mathematically. Rigor can also be defined in terms of the difficulty of mathematics items. Mathematics difficulty is related to skills, but it has a different dimension. A problem requiring the performance of lengthy arithmetic computations, although a low-level skill, may be a difficult problem as measured by a low percentage of correct answers. Hence, this study measures mathematics rigor as a combination of the type of mathematics skills required in answering an item and the difficulty of an item as measured by the percentage of students who correctly answer it. On these measures:

- The U.S. performance is below the 12-country average at both low- and high-skill levels and low and high-levels of item difficulty.

- For all countries, there is a high correlation between a country's score on items of low and high mathematical rigor, whether measured by skills or difficulty.

These results suggest that the mathematical abilities to solve problems at different levels of mathematics rigor are complementary. Therefore, the evidence does not support proposals to reduce attention to learning computational and simpler mathematics skills in order to focus on strengthening students' ability to handle more complicated mathematics reasoning.

Content areas (number/quantity, algebra/change and relationships, measurement, geometry/space and shape, data/uncertainty). Relative to the United States' average score for all content areas, U.S. performance is significantly weakest in measurement in grades 4 and 8 and in geometry in grade 8 and at age 15. The United States scores consistently strongest in data and statistics in all three assessments. The results also show a strong correlation between countries' grade 4 and 8 performance on number, algebra, and data and statistics, suggesting the importance of developing a strong early foundation in these content areas.

Although the overall amount of instructional time devoted to mathematics in the United States is similar to the average amount of time in the other countries, the distribution of that time across mathematics content areas differs in ways consistent with our findings about relative performance across content areas. The TIMSS analyses find that elementary classrooms in the United States spend proportionally less time on geometry in grade 8 than those in most other comparison countries, and consistently more time on data/statistics, especially in the early grades.

Gender. Boys in the United States consistently outperform girls in all three assessments. Although the differences are small—less than a tenth of a standard deviation—the United States and Italy are the only countries out of the 12 countries in which boys consistently outperform girls on all three assessments. Across all countries, there is no evidence from the international data of a persistent gender gap in mathematics on TIMSS-4 or TIMSS-8. However, on PISA, boys outperform

girls across all countries. The boys' score advantage is greatest on the more difficult items, which is consistent with some prior gender literature. Although girls are more likely than boys to believe that they are not good at mathematics, on PISA differences in girls' self-perceptions of their mathematics ability are not related to the size of the girls' mathematics disadvantage across countries.

Instructional background factors. The TIMSS data provide a rich set of information about mathematics instruction in the participating countries and show that several features of the mathematics system in the United States differ from those in most of the comparison countries. Some of the more important instructional differences follow:

The United States does not have national mathematics standards. The de facto U.S. standards prepared by the National Council of Teachers of Mathematics (NCTM) are voluntary and describe the mathematical content only in terms of general expectations over broad grade bands, rather than specific mathematics topics grade-by-grade.

- Grade 4 classrooms in the United States use calculators more often than those in most of the other comparison countries.

- A disproportionately high percentage of grade 4 teachers have no mathematics specialization and grade 8 teachers have no mathematics degree.

Conclusion. These findings demonstrate that carefully conducted international comparisons over a common group of countries offer a broader context in which to examine the U.S. mathematics system than is possible from domestic research alone. By looking outward, we can collect important new evidence to improve our understanding of our mathematics system's performance and its internal weaknesses and strengths.

In our K-twelve we were doing okay at the fourth-grade level, we were doing middle-of-the-road in the eighth grade, and by twelfth grade we were hovering near the bottom in international tests related to math.
—Tracy Koon, Intel's Director of Corporate Affairs, quoted in T. Friedman, *The World Is Flat (2005)*

INTRODUCTION

On December 6, 2004, the U.S. Department of Education released the Program for International Student Assessment (PISA) results along with a press statement concluding: "America's 15-year-olds performed below the international average in mathematics literacy and problem-solving" (ED, 2004a).

Only one week later on December 14, 2004, a U.S. Department of Education press release stated: "America's fourth- and eighth-grade students significantly outperformed many of their international peers, scoring well above the international average in both mathematics and science, according to the latest results from the Trends in International Mathematics and Science Study (TIMSS)" (ED, 2004b). Citizens exposed to both releases could understandably be confused. On the surface, the findings clearly painted a portrait of higher-than-average U.S. mathematics performance at the elementary and middle grades on TIMSS-4 and TIMSS-8 and a faltering performance among 15-year-old students in the United States on PISA.

Turning from press releases to the *The Condition of Education* (2005b) published by the National Center for Education Statistics (NCES), we find a similar conclusion about the falloff in U.S. international mathematics performance:

- "U.S. students at grades 4 and 8 scored above the international average in 2003. . . . U.S. 4th-graders scored higher, on average, than students in 13 countries, while students in 11 countries outperformed U.S. students.[i] At grade 8, the average U.S. mathematics score was higher than those of students in 25 countries, but below the average scores of students in 9 countries."[ii]

- "U.S. 15-year-olds, on average, scored below the international average for participating OECD [Organization for Economic Cooperation and Development] countries in combined mathematics literacy. . . . In combined mathematics literacy, students in 20 OECD countries and 3 non-OECD countries outperformed U.S. students, while U.S. students outperformed students in 5 OECD countries and 7 non-OECD countries."[iii]

This perception of a secondary-school mathematics decline is reflected in the National Academy of Sciences (NAS) description of U.S. mathematics performance: "Our 4th-grade students perform as well in math and science as do their peers in other nations, but 12th-graders in 1999 were almost last among students who participated in the Trends in International Mathematics and Science Study" (NAS, 2005). Similarly, the Business Roundtable proposal for mathematics and science education describes the U.S. mathematics problem as a secondary-school problem: "Although U.S. fourth graders score well against international competition, they fall near the bottom . . . . by 12th grade in mathematics . . . ." (Business Roundtable, 2005). The recommendations of these groups also stress reform of secondary school mathematics instruction.

A central issue in drawing these conclusions from the international comparisons is how to adjust for the changing mix of 24 countries participating in the 2003 TIMSS-4, 45 countries participating in TIMSS-8, and 40 countries participating in PISA, age 15. This paper reexamines the most recent TIMSS and PISA findings about the decline in U.S. mathematics performance from elementary to secondary grades by focusing on the countries that participated in all three assessments. Our specific focus is to examine the performance of the 12 industrialized countries (including the United States) that participated in the TIMSS-4, TIMSS-8, and PISA assessments.[iv] The other 11 countries are Australia, Belgium, Hong Kong, Hungary, Italy, Japan, Latvia, Netherlands, New Zealand, Norway, and the Russian Federation.

With respect to this common 12-country comparison group, it is appropriate to ask:

- Was U.S. student performance really above the international average in grades 4 and 8, and does U.S. student performance really decline precipitously after grade 8?

- Once the pattern of performance across assessments is clarified, are there factors that can help explain U.S. performance on international assessments?

When the results of these assessments were released, press reports quoted experts who focused their explanation for the apparently weak U.S. performance on PISA's assessment by its focus on assessing the ability of students to apply mathematics in real-world contexts, as compared with TIMSS's emphasis on assessing mathematics content knowledge found in schools' curriculum (Cavenagh, 2004; Cavenagh and Robelen, 2004). These experts urged the United States to strengthen instruction in solving real-world mathematical problems. Our analysis of the common set of 12 countries takes a broader view, however, and examines several potential causes for lower U.S. performance. This analysis examines how U.S. mathematics performance differs by the mathematical rigor of assessment items, the mathematics content area (e.g., algebra or statistics), gender of students taking the assessments, and certain important instructional features of a country's mathematics system.

It is worth noting that interpretations of the similarities and differences of country-level results on large-scale international assessments provide evidence only of association, not causation. International assessments that include widely different education systems offer a natural laboratory in which to identify characteristics of mathematics systems associated with performance differences, but require further study of the effect of these characteristics within a particular country to demonstrate their applicability and validity.

The paper consists of seven sections:

- Section I describes the core features of the TIMSS and PISA assessments.

- Section II looks at how U.S. scores rank internationally across the common set of 12 countries at grades 4 and 8 and at age 15 (i.e., modal grade of 10). It also explores the stability of scores for the other countries to determine whether countries' initial grade 4 performance is a good predictor of their grade 8 performance and their grade 8 performance is a good predictor on PISA at age 15.

- Sections III, IV, and V examine whether U.S. and other countries' international performances differ because of the mathematical rigor of assessment items; the content area of assessment; and the results of boys versus girls.

- Section VI compares several instructional characteristics of mathematics systems in the United States and other countries that research suggests could explain international differences in mathematics performance.

- Section VII summarizes the implications of the findings for strengthening mathematics instruction in the United States.

I. CHARACTERISTICS OF THE TIMSS AND PISA ASSESSMENTS

Any valid interpretation of U.S. performance drawn from TIMSS or PISA requires an understanding of the differences between these two international assessments.

Exhibit 1 compares the key characteristics of the TIMSS and PISA assessments with respect to the following key features:

Purpose. TIMSS and PISA differ in their aims. TIMSS serves a traditional purpose of testing student knowledge of the mathematics content that students in the participating countries would typically have had an opportunity to learn by grades 4 and 8. "Every effort was made to ensure that the tests represented the curricula of the participating countries and that the items exhibited no bias toward or against particular countries" (Mullis et al., 2004). The TIMSS conceptual model is framed in terms of measuring countries' "achieved curriculum" on the assessment in relation to "intended curriculum" taught by the schools.

Whereas TIMSS assesses knowledge of the curriculum taught, PISA's design has a different purpose—to assess students' ability to apply mathematical ideas to solving real-world problems. PISA states its aim as follows: "The assessment is forward-looking, focusing on young people's ability to use their knowledge and skills to meet real-life challenges, rather than merely on the extent to which they have mastered a specific school curriculum. This orientation reflects a change in the goals and objectives of the curricula themselves, which are increasingly concerned with what students can do with what they learn at school, and not merely whether they can reproduce what they have learned" (OECD, 2004, p. 20).

Mathematics framework: Content areas. TIMSS and PISA frameworks consist of different but related mathematical content areas, although the overlap is imperfect. The TIMSS frameworks are organized around the five mathematics content areas that many countries use to organize their primary and middle school mathematics curriculum: number, patterns or algebra, measurement, geometry, and data/statistics. In the United States, the National Council of Teachers of Mathematics (NCTM, 2000) framework and the majority of state frameworks are organized around these same five mathematical content areas (Ginsburg, Leinwand, Anstrom, and Pollock, 2005; National Research Council, 2001).

PISA defines its mathematics content in terms of four "overarching ideas" (quantity, space and shape, change and relationships, and uncertainty) and only secondarily in relation to "curricular strands" (such as number, algebra, and geometry). Although PISA assesses an older age group and appears to organize mathematics content in a different way, in practice, the four PISA content categories can be related to the five TIMSS categories, although with differing degrees of common topics, as follows:

- PISA's quantity category describes "numeric phenomena" [v] (e.g., representing numbers, meaning of operations, magnitude of numbers, computations, mental arithmetic, and estimation) and aligns quite well with the mathematics content covered in TIMSS's traditional number category (e.g., whole numbers, integers, ratios/fractions, and percents).

Exhibit 1. Key Features of TIMSS Grades 4 and 8 and PISA Age 15 Mathematics Assessments, 2003

	TIMSS	PISA
Purpose	Assess "students' achievement in relation to different types of curricula, instructional practices, and school environments." (Mullis et al., 2004, p. 13)	"Presents students with problems mainly set in real-world situations" and "obtain measures of the extent to which students presented with these problems can activate their mathematical knowledge and competencies to solve such problems successfully." (OECD, 2004, p. 37)
Mathematics framework: Content areas	• Number • Patterns, Equations, and Relationships (grade 4) or Algebra (grade 8) • Measurement • Geometry • Data (e.g., statistics)	• Quantity • Change and relationships • Space and shape • Uncertainty
Mathematics framework: Cognitive skills	• Knowing facts and procedures • Using concepts • Solving routine problems • Reasoning	• Reproductions (simple operations) • Connections (bringing together ideas) • Reflection (deeper mathematical thinking)
Assessment items	• Grade 4: 161 assessment items with 65% multiple choice and 35% constructed response • Grade 8: 194 assessment items with 71% multiple choice and 29% constructed response • Calculator permitted for new items in 2003	• 85 assessment items with 33% multiple choice and 67% constructed response • Calculator use left to country to determine
Background questionnaires	Student, teacher, and principal questionnaires	Student and principal questionnaires
Country participation	• 24 countries in grade 4[vi]; 45 countries in grade 8[vii] • 13 countries participated in TIMSS-4, TIMSS-8, and PISA (excludes Tunisia because our analysis is limited to industrialized countries)	• Age 15 (15.3 yr.–16.2 yr.) • 40 countries and 29 OECD country results[viii] • 13 countries participated in TIMSS-4, TIMSS-8, and PISA (excludes Tunisia because our analysis is limited to industrialized countries)
Assessment cycle	Assessment of mathematics every 4 years	Assessment of some mathematics every 3 years, but next in-depth mathematics component is 2012
Source: Mullis et al., 2003; Mullis et al., 2004; OECD, 2004.		

- PISA's change and relationships category relates to TIMSS's algebra category (patterns, algebraic expressions, formulas, relationships), but TIMSS stresses relationships and does not focus as explicitly as PISA on describing the change process mathematically. Also, PISA's emphasis on the relationships among different mathematical representations (e.g., symbolic, algebraic, graphical) is more advanced than that of TIMSS.

- PISA's uncertainty category (collecting data, data analysis and display/visualization, probability and inference) is quite similar to the mathematics topics covered by TIMSS's traditional data category, which includes data displays and probability.

- PISA's space and shape category (spatial and geometric phenomena and relationships) aligns only imperfectly with TIMSS's measurement and geometry categories. PISA does not cover measurement, a topic stressed primarily in elementary grades mathematics. Like PISA, TIMSS covers the similarities and differences of shapes and objects and 2- and 3-dimensional representations of objects in geometry. However, PISA's treatment of space and shape is more complex and includes the understanding of shadows, perspectives, and the degree of reality in representations of shapes (e.g., maps of cities).

Overall, TIMSS's traditionally defined content categories and PISA's overarching ideas match up well on two content areas—numbers and data/statistics. PISA's treatment of measurement/geometry and algebra emphasizes more advanced concepts than does TIMSS. The differences between TIMSS and PISA content categories have to be kept in mind when examining U.S. strengths and weakness across the content areas on the different assessments.

Mathematics framework: Cognitive skills. The TIMSS and PISA frameworks define a set of mathematics skills that students should be able to do within each content area. TIMSS and PISA use different words, but fundamentally both order mathematical skills by the cognitive demands involved. TIMSS processes range from facts to reasoning and PISA from reproduction to reflection (Exhibit 1). The fact that TIMSS and PISA have a similar ordering of mathematical processes should not be interpreted to mean that the categories cover similar cognitive processes, because PISA's processes require a greater ability to represent real-world problems mathematically. [ix]

Assessment items. The different focuses of the TIMSS and PISA assessments show up in the degree to which each relies on multiple-choice or constructed-response questions. The PISA test format uses constructed-response questions for 67 percent of its items, while such questions make up 35 percent of TIMSS items at grade 4 and 28 percent at grade 8. The open-ended items used on PISA allow divergent responses and scoring based on depth of answers, which is consistent with PISA's focus on having students demonstrate their ability to apply mathematics concepts to real-world settings.

The mathematics items also differ in whether students are permitted to use calculators. Until 2003, calculators were not permitted on TIMSS (Mullis et al., 2004. p. 74). TIMSS allows students to use calculators for new items only, in order to maintain comparability with items on previous tests. PISA leaves calculator use optional to the country.

Background questionnaires. TIMSS surveys students, teachers, and principals about the context in which mathematics instruction is provided, but PISA surveys only students and principals, omitting the teacher questionnaire. Without teacher responses, PISA's information about the nature of mathematics instruction is quite weak compared with that of TIMSS. [x]

Country participation. The mix of countries differs considerably on each assessment, which makes it difficult to compare U.S. rankings on TIMSS and PISA. Twenty-four countries participated in TIMSS-4; 45 countries in TIMSS-8; and 40 countries in PISA at age 15 (primarily grades 9 and 10). Several of the high-scoring Asian countries, including Chinese Taipei and Singapore, participated in TIMSS but not in PISA; in addition a substantial majority of the lower-performing,

less-industrialized nations were TIMSS but not PISA participants. Of the 29 OECD members that participated in PISA, 19 did not participate in one or both of the TIMSS assessments.

Assessment cycle. TIMSS and PISA follow regular, but different, mathematics assessment cycles. Having both administer an in-depth mathematics assessment in 2003 was an uncommon occurrence. The TIMSS assessment follows a four-year cycle, with 2007 set as the next administration date. PISA is administered on a three-year cycle, but each administration consists of an in-depth assessment of one subject—mathematics, science, or literacy—and a less complete assessment of the other two subjects. PISA will not conduct another in-depth mathematics assessment until 2012, and the TIMSS and PISA assessments will not comprehensively assess mathematics in the same year until 2039. So this 2003 confluence presents a unique opportunity to examine U.S. student mathematics performance in different international contexts.

II. U.S. MATHEMATICS PERFORMANCE ON INTERNATIONAL ASSESSMENTS

To address the issue of making meaningful interpretations of U.S. performance relative to other countries, we identified the 12 industrialized countries that participated in all three assessments. Along with the United States, the other 11 countries are Australia, Belgium, Hong Kong, Hungary, Italy, Japan, Latvia, Netherlands, New Zealand, Norway, and the Russian Federation. These 12 countries span four continents—North America, Australia/Oceania, Asia, and Europe—and constitute a broad range of industrialized nations for comparison.

Exhibit 2 displays the country scores for each of the 12 countries participating in all three 2003 international mathematics assessments. The TIMSS and PISA scales are both set at a mean of 500 and a standard deviation of 100.[xi]

Exhibit 2. Scores and Rankings[2] of 12 Countries Participating on the 2003 International Mathematics Assessments: TIMSS Grades 4 and 8, and PISA Age 15[1]

Country	TIMSS Grade 4		TIMSS Grade 8		PISA Age 15	
	Score	Rank	Score	Rank	Score	Rank
AUS	499*	10	505	8	524*	5
BEL	551*	3	537*	3	529*	4
HKG	575*	1	586*	1	550*	1
HUN	529*	7	529*	5	490	8
ITL	503*	9	484*	11	466*	12
JPN	565*	2	570*	2	534*	3
LAT	536*	5	508	6	483	9
NLD	540*	4	536*	4	538*	2
NZL	493*	11	494*	10	523*	6
NOR	451*	12	461*	12	495*	7
RUS	532*	6	508	6	468*	11
USA	518	8	504	9	483	9
AVG	524		519		507	
	Countries statistically above U.S. = 7 Countries statistically below U.S. = 4 Difference = 3 countries statistically above U.S.		Countries statistically above U.S.= 5 Countries statistically below U.S.= 3 Difference = 2 countries statistically above U.S.		Countries statistically above U.S.= 6 Countries statistically below U.S.= 3 Difference = 3 countries statistically above U.S.	
	Scorecor (TIMSS4,TIMSS8) = .93* Rnkcor (TIMSS4,TIMSS8) = .96* Scorecor (TIMSS8,PISA15) = .67* Rnkcor (TIMSS8,PISA15) = .66*					

***** Indicates country scored statistically significant above or below the United States at the .05 level.

[2]Country rankings are from highest score (equals 1) to lowest score (equals 12).

[1]Tunisia also participated in all three international results, but it is not an industrialized country and was omitted from our study.

Source: Mullis, Martin, Gonzalez, and Chrostowski, 2004; OECD, 2004.

Quite a different portrait of U.S. international mathematics performance emerges when U.S. scores are compared with the scores of the 11 other countries rather than with a variable number of countries. The United States ranks 8th on TIMSS-4, 9th on TIMSS-8, and 9th on PISA, age 15. At grade 4, seven countries have scores that are significantly above the U.S. score compared with four countries with scores significantly below the U.S. score. For grade 8, five countries have scores significantly above the United States, and three have scores significantly scores below the U.S. score. At age 15, six countries have scores significantly above the United States, while three have scores significantly below. Hence, the comparison changes little as U.S. students progress through school—a consistent picture of overall mediocrity.

The consistent showing of mediocre U.S. performance across international assessments is a robust finding that extends to other ways to compare U.S. performance across assessments. Among countries that participated in two successive assessments, TIMSS-4 and TIMSS-8 or TIMSS-8 and PISA, their ranks change very little between the successive assessments. Also, the correlations among country ranks on successive assessments is .85 or above, which indicates a high degree of stability of country international performance over the groups of countries that participated on successive pairs of assessments.[xii]

In addition to examining the constancy of the relative U.S. score across the three assessments, we can examine the constancy of scores for all 12 countries. This is estimated by computing the correlation coefficients of the country scores between any two assessments. Thus, the correlations are computed between a country's score and rank on TIMSS-4 and TIMSS-8, and between TIMSS-8 and PISA. A correlation of near one indicates that countries' scores or ranks on one assessment have a strong relationship between countries' scores or ranks on the successive assessment; a correlation near zero indicates a weak relationship between assessments. The correlation between country scores on TIMSS-4 and TIMSS-8 is .93 and between country rankings is .96. This indicates a high degree of association between a country's relative performance at grade 4 and its relative performance at grade 8. The correlations between TIMSS-8 and PISA, while still statistically significant, are considerably lower at .67 for country score and .66 for country ranking.

Measurement differences between TIMSS-8 and PISA will reduce the correlations between country scores and rankings on TIMSS and PISA. One measurement difference occurs because TIMSS samples students in a particular grade and PISA samples students of a particular age. A second difference is that TIMSS items stress the content taught in traditional classrooms while PISA items focus on real-world applications of mathematics. We expect that if PISA were more similar to TIMSS in its sample and purpose, the correlations between country scores and ranks on TIMSS-8 and PISA would be higher than the observed two-thirds, which is still statistically significant.

Finally, measurement error may affect the accuracy of all the score correlations because we used the original country scale scores reported by each of the three assessments and did not recalibrate them using the common 12-country sample. To test for possible errors from not rescaling, we computed each country's score by using its average percentage of students who correctly answered each item on an assessment. This score is not sensitive to the fact that we used only a subset of countries instead of all the countries that participated in the assessment. The correlations between countries' reported scale scores and the countries' average percentages correct were .98 or above.[xiii] We conclude that there is strong evidence to support continued use of the published scale scores on TIMSS and PISA because they are highly correlated with scores that are not sensitive to the country mix.[xiv]

In summary, the analysis of international performance among the 12 countries reveals consistently mediocre U.S. results on all three assessments. We also find that overall there is a strong correlation between a country's mathematical performance on TIMSS-4 and TIMSS-8 and generally a positive and significant, although somewhat weaker correlation between TIMSS-8 and PISA (e.g., age 15). These results suggest that efforts to improve U.S. international mathematics performance should include a component that strengthens U.S. mathematics education in the early elementary grades, because, generally speaking, a country's initial performance is correlated with its later performance.

III. MATHEMATICAL RIGOR OF ITEMS—COGNITIVE SKILLS AND DIFFICULTY

One factor to consider in understanding international mathematics performance is how a country's performance compares across items of different mathematical rigor. The relationship between performance on items of lesser and greater mathematical rigor may help to resolve the controversy over how much to emphasize computational and procedural skills in primary-grades mathematics instruction, given the widespread availability of calculators and computers.

One argument favors devoting more classroom time to developing students' 21st century skills, including mathematical reasoning and communication, rather than doing arithmetic. Meta-analyses of different studies suggest that using calculators does not weaken basic skills and may enhance higher-level mathematics understanding (Hembree and Stein, 1997). Similar results were obtained from a long-term study of calculator use in Sweden (Brolin and Björk, 1992) and from studies of a large number of children in Australia (Groves, 1994) and England (Shuard, 1992), although none of these countries represents the high performers on TIMSS.

An opposing line of reasoning, also with empirical support, argues that mastery of basic mathematical operations, including computational skills, remains a prerequisite to solving more-complicated mathematical problems (Loveless, 2004; National Research Council, 2001). Its proponents cite research suggesting that students who are able to quickly cut through the computational aspects of problems are better able to focus their attention on developing solutions to challenging, less-straightforward problems.

The international assessments provide a source of empirical information about how well countries perform on assessment items that represent different levels of mathematical rigor. Finding that countries that tend to do well on less-rigorous items also do well on more-rigorous items would support those advocating a balanced approach to teaching across the range of mathematics complexity and difficulty. On the other hand, if some countries do very well on more-rigorous items and do not perform as well on less-rigorous items, then this fact may reinforce the position of those who suggest a focus on teaching more-complex mathematical problems and letting calculators do the less-complex work. We examine country performance on items of different mathematical rigor measured two ways: cognitive skills and mathematical difficulty of items.

Cognitive skills. Each item on TIMSS and PISA is classified according to the types of cognitive skills a student is expected to use to solve the problem. Results for each country, as measured by the percentage of students who answered an item correctly, were classified by cognitive skill groups and were available for individual TIMSS but not PISA items. TIMSS uses a four-part skills categorization: knowing facts and procedures, using concepts, solving routine problems, and mathematical reasoning. These categories are ordered so that "In general, the cognitive complexity of tasks increases from one broad cognitive domain to the next" (Mullis et al., 2003). The less demanding end of the cognitive spectrum, knowing facts and procedures, involves recalling definitions, recognizing mathematically equivalent entities, and correctly performing computational procedures. At the other end, reasoning requires students to analyze more-complex problems by breaking them into their parts, making connections between different mathematical ideas, or solving nonroutine problems that they are not likely to have seen before.

Exhibit 3 displays the U.S. and other 11 countries' average percent correct and rank on TIMSS-4 and TIMSS-8 assessment items falling into the least and most demanding cognitive skills categories. U.S. students' scores and ranks are uniformly lodged in the bottom half of the distribution on both TIMSS assessments and for both the high and low ends of the range in cognitive skills. These results suggest that students in the United States need improvements in both their ability to apply facts and procedures and their ability to reason through more-complicated multistep problems.

In general, other countries also display a strong relationship between their scores on items at different levels of cognitive demand. The correlations between scores and ranks at different levels of cognitive demand are somewhat higher on TIMSS-4 (above .9) than on TIMSS-8 (above .8). On TIMSS-8, Latvia, Netherlands, New Zealand, and Russia shifts at least three rank points between knowing facts and procedures and reasoning. This may reflect a smaller difference in the type of reasoning required by the least and most demanding items at grade 4, where the emphasis is on number, than at grade 8, where there is a greater focus on algebra and geometry.

Item difficulty. A second measure of an item's mathematical rigor is the difficulty that students have in answering the item correctly. Item difficulty is determined by the percentage of students who answered the item correctly.xv In general, those items with the highest p-values (percent correct) are assumed to be least difficult and those with the lowest p-values are assumed to

Exhibit 3. Country percent Correct and Rankings[1] on the Least and Most Demanding Cognitive Skills: TIMSS-4, TIMSS-8, 2003.

	TIMSS-4				TIMSS-8			
	Knowing Facts and Procedures		Reasoning		Knowing Facts and Procedures		Reasoning	
	Pct. Cor.	Rank	Pct. Cor.	Rank	Pct. Cor.	Rank	Pct. Cor.	Rank
AUS	55.6	10	45.4	9	50.7	9	44.6	6
BEL	70.3	3	54.4	3	62.0	4	48.5	5
HKG	73.1	1	57.0	2	74.7	1	56.3	2
HUN	62.7	5	51.1	5	63.4	3	50.2	3
ITL	62.2	7	44.1	10	49.6	10	37.3	11
JPN	72.1	2	58.3	1	67.6	2	57.2	1
LAT	62.3	6	52.7	4	59.2	6	40.8	10
NLD	63.8	4	50.3	6	57.7	7	49.6	4
NZL	53.9	11	43.8	11	46.2	11	42.3	8
NOR	46.3	12	35.4	12	35.7	12	35.5	12
RUS	62.0	8	48.4	7	61.1	5	40.9	9
USA	60.2	9	47.6	8	53.5	8	42.9	7
Avg.	62.0	—	49.0	—	56.8	—	45.5	—
	Pctcor (TIMSS4KFP,TIMSS4RE) = .98* Rnkcor (TIMSS4KFP,TIMSS4RE) = .94*				Pctcor (TIMSS8KFP,TIMSS8RE) = .88* Rnkcor (TIMSS8KFP,TIMSS8RE) = .82*			

* Indicates significant at the .05 level.

[1]Country rankings are from highest percent correct (equals 1) to lowest percent correct (equals 12) for each cognitive skills group.

Source: Mullis, Martin, Gonzalez, and Chrostowski, 2004

be most difficult. However, the match between item difficulty and cognitive demand is imperfect. An item that requires computation of multidigit numbers may involve only basic procedures, but still have a high rate of incorrect responses. Hence, examining items by difficulty adds another dimension to understanding mathematical rigor in addition to cognitive demand.

Items on each assessment have been grouped into quartiles by their average percentage of correct responses over all test takers. Exhibit 4 displays each country's percentage of correct responses for the bottom and top quartiles of item difficulty on TIMSS-4, TIMSS-8, and PISA. Countries that rank high in terms of their percentage of correct answers on the least difficult items are also the countries that rank high on the percentage of correct answers on the most difficult items. Conversely, countries with a below-average score on the least difficult items tend to be below average on the most difficult items. The relationship is not perfect. For example, Latvia departs from the constancy of ranks on TIMSS-8 and the United States on PISA as they shift three rank points between the least and most difficult assessment items.

Nevertheless, the overall high correlation coefficients of near .90 or higher on TIMSS-4, TIMSS-8, and PISA are evidence that countries with students who do well on the most difficult items are also the ones whose students do well on the less difficult items. These results are consistent with

Exhibit 4. Country Mathematics Performance by Percent Correct and Rank[1] on Items in the Least Difficult Quartile and the Most Difficult Quartile: TIMMS-4, TIMSS-8, and PISA, 2003

Country	TIMSS-4				TIMSS-8				PISA			
	Least Difficult 25%		Most Difficult 25%		Least Difficult 25%		Most Difficult 25%		Least Difficult 25%		Most Difficult 25%	
	Pct. Cor.	Rnk	Pct. Cor.	Rnk	Pct. Cor.	Rnk	Pct. Cor.	Rnk	Pct. Cor.	Rnk	Pct. Cor.	Rnk
AUS	77	10	30	9	71	8	26	6	80	3	29	5
BEL	87	1	40	3	78	3	30	5	81	1	32	3
HKG	87	1	45	2	84	1	48	1	80	3	36	1
HUN	81	6	34	6	76	5	32	3	76	7	23	7
ITL	78	9	29	11	66	11	21	11	69	11	20	11
JPN	87	1	48	1	82	2	42	2	81	1	33	2
LAT	83	5	36	4	74	6	23	9	74	9	21	9
NLD	85	4	34	6	77	4	31	4	80	3	31	4
NZL	74	11	30	9	68	10	22	10	79	6	29	5
NOR	67	12	23	12	60	12	15	12	75	8	23	7
RUS	81	6	36	4	74	7	26	7	74	9	20	11
USA	81	6	32	8	71	8	24	8	67	12	21	9
AVG	81	—	35	—	73	—	28	—	76	—	27	—
	Pctcor (TIMSS4LD,TIMSS4MD) =.88* Rnkcor (TIMSS4LD, TIMSS4MD) = .92*				Pctcor (TIMSS8LD,TIMSS8MD) = .92* Rnkcor (TIMSS8LD,TIMSS8MS) = .91*				Pctcor (PISALD,PISAMD) = .86* Rnkcor (PISALD,PISAMD) = .91*			

* Indicates statistical significance at the .05 level.

[1]Country rankings are from highest percent correct (equals 1) to lowest percent correct (equals 12) for each item difficulty group.

Source: Mullis, Martin, Gonzalez, and Chrostowski, 2004; OECD, 2004.

the view of those who support a balanced instructional approach that stresses the importance of instruction in both basic and higher-order mathematics skills.

IV. MATHEMATICS CONTENT AREAS

This section examines the relative strengths and weaknesses of U.S. scores and those of other countries across the different mathematics content areas on the three assessments. Content areas identified as relatively weak may warrant special attention to determine the causes of poorer performance.

As described in Exhibit 1, TIMSS breaks down its assessment items into five content areas and PISA divides them into four areas. As discussed in greater detail in section I above, the four PISA content areas relate to the TIMSS content areas as follows: PISA quantity to TIMSS number; PISA change and relationships to TIMSS algebra; PISA space and shape to TIMSS geometry; and PISA uncertainty to TIMSS data. Measurement is normally taught in the elementary grades and therefore appears on TIMSS but not on PISA. Despite their general similarities, the overlap between the TIMSS and PISA content areas is imperfect. A major reason for the difference in content is PISA's concentration on real-world applications of mathematics in contrast with TIMSS's emphasis on traditional classroom content. PISA's change and relationships category, for instance, places more emphasis on the mathematics of the change than TIMSS's algebra content area, which emphasizes such traditional classroom topics as solving two linear equations (Mullis et al., 2003; OECD, 2003).

Exhibit 5 displays a country's relative strength or weakness by content area for the United States and the comparison countries on TIMSS-4, TIMSS-8, and PISA-age 15. A country's relative strength on a content area may be expressed as the difference between a country's content area score and the average of its scores across all the content areas. A positive difference indicates that a country performs better than its average score; a negative difference indicates that a country performs below its average score.

Statistics is the area of clear U.S. strength across all three assessments. For grades 4 and 8, U.S. scores on statistics are over one-quarter of a standard deviation above the U.S. average score, a difference that is considered of moderate to high educational significance (Rosenthal and Rosnow, 1984). U.S. students also perform somewhat better than average on algebra, although the difference is not statistically different from the U.S. average.

By contrast, the United States performs relatively poorly on measurement and geometry. Its performance is particularly weak in measurement in TIMSS-4 and in geometry in TIMSS-8. It is probably not a coincidence that these are the grades in which measurement and geometry are stressed, respectively. The results for number, which includes computational operations, are inconsistent across the three assessments.

Correlating the score deviations of any two assessments across all 12 countries determines whether the countries, as a whole, show consistent patterns of content area strengths and weaknesses on the two assessments. The higher the correlation, the more consistent is the pattern of country strengths or weaknesses across two assessments. The correlations generally show the following:

- Countries with a relative strength in number, algebra, and data at grade 4 continue to perform relatively well in these content areas at grade 8. The strong positive correlation, however, fails to hold for measurement and geometry between grades 4 and 8. Perhaps this is because measurement is mainly taught in the earlier grades, so grade 4 scores do not predict grade 8 middle school results; by similar logic, geometry is emphasized in

Exhibit 5. Difference in the Scores of a Mathematics Content Area From Country Average Score on TIMSS (Grades 4 and 8) and PISA (Age 15) for the 13 Countries Participating in All Three International Tests

Country	Number/ Quantity			Algebra/ Change & Relationships			Measurement			Geometry/ Space & Shape			Data/ Uncertainty		
	Gr 4	Gr 8	Age 15	Gr 4	Gr 8	Age 15	Gr 4	Gr 8	Age 15	Gr 4	Gr 8	Age 15	Gr 4	Gr 8	Age 15
AUS	-28*	-8	-7*	-12*	-7	2	7	5	N/A	17*	-15*	-3	18*	25*	8*
BEL	5	5	0	-2	-11*	5*	6*	1	N/A	-11*	-7	0	4	12*	-4
HKG	9*	5	-5	3	-1	-10*	-2	3	N/A	-8	7	8	-3	-15*	8
HUN	-2	3	6*	19*	8*	5	6	-1	N/A	-12*	-11*	-11*	-13*	0	-1
ITL	-2	-3	10*	-8	-6	-13*	0	17*	N/A	18*	-14*	5	-7	7	-2
JPN	-10*	-12*	-9*	-12*	-1	0	2	-10*	N/A	-7*	18*	17*	27*	4	-8
LAT	0	0	0	1	1	5	14*	-7	N/A	-8*	8	4	-5	-1	-8*
NLD	0	4	-11*	-9*	-21*	13*	9*	14*	N/A	-15*	-22*	-13	17*	25*	11*
NZL	-27*	-16*	-13*	-7	-7	3	1	3	N/A	15*	-9	2	20*	29*	9*
NOR	-22*	-9*	-1	-23*	-37*	-7*	13*	16*	N/A	16*	-4	-12*	17*	33*	19*
RUS	5	0	7	4	11*	12*	11*	2	N/A	1	10*	9*	-22*	-21*	-29*
USA	-5	6	-5	3	8	5	-21*	-7*	N/A	-3	-30*	-9*	28*	25*	10*
	Scorecor (4,8) = .82**			Scorecor (4,8) = .78**			Scorecor (4,8) =.30			Scorecor (4,8) = .15			Scorecor (4,8) = .78**		
	Scorecor (8,15) = .35			Scorecor (8,15) =.26						Scorecor (8,15) =.77**			Scorecor (8,15) =.77**		

*Significant at the .05 level (TIMSS and PISA differences are shown significant if a two standard deviation confidence interval around a content area falls outside the country average across all content areas).

Source: Mullis, Martin, Gonzalez, and Chrostowski, 2004; OECD, 2004

middle school, so middle school geometry results are not closely related to a country's grade 4 strengths or weaknesses in geometry.

- Between TIMSS-8 and PISA, countries that perform relatively well in the areas of geometry and data/uncertainty are also likely to perform well in these areas on PISA. Geometry is taught primarily in the middle grades, so results may carry forward into high school. One possible explanation for the strong correlation in data/uncertainty performance on TIMSS and PISA is that countries that emphasize handling data and data applications—skills that are emphasized in PISA—also stress these skills in earlier grades. This is suggested by the strong association between relative performance in data on TIMSS-4 and TIMSS-8. The weak association between TIMSS and PISA on algebra is surprising given that algebra is stressed in middle school. As we noted earlier, PISA stresses particular algebra applications, such as those dealing with change; such applications may not be reflected in the TIMSS scores, which signal more traditional classroom topics, such as solving two linear equations.

In summary, the results call attention to the United States relatively poor performance on measurement and geometry and comparatively strong performance on statistics. The results also

indicate that countries that develop a strong early foundation in data/statistics will find their relative advantage in this area persisting into later grades. In geometry, middle school performance likely determines later results in this content area. The number and algebra results show an early advantage persisting into middle school, but not secondary school. This may be because TIMSS emphasizes traditional classroom mathematics, while PISA emphasizes real-world applications.

V. GENDER

Another factor that potentially explains U.S. and other countries' international mathematics scores is the difference between girls' and boys' mathematics performance. The president of Harvard University recently ignited a firestorm of controversy by theorizing that girls might be genetically less able to perform as well as boys in the areas of science and mathematics, especially at the high end of the ability range (Summers, 2005). The TIMSS-PISA data allow for cross-age international comparisons of girls' mathematics performance relative to that of boys' to determine how common it is for girls to perform less well than boys at different developmental points.

Exhibit 6 displays the difference for each country between its average score for girls minus its average score for boys on TIMSS-4, TIMSS-8, and PISA. Overall, the magnitudes of the gender differences typically are not large. The maximum difference in scores is 9 points at grade 4, 24 points at grade 8, and 18 points at age 15. Ten points equals one-tenth of a standard deviation, or only about three percentage points at the 50th percentile. Nonetheless, these 3 percentage points indicate that about 1.5 million U.S. students could be affected if access to colleges or jobs in certain fields is rank ordered by students' mathematics performance.

Girls in the United States show a small, consistent performance disadvantage of 8 points on TIMSS-4 and 6 points on TIMSS-8 and PISA. We note that across all 12 countries, only the United States and Italy have a consistent and statistically significant difference favoring boys. Across all the

Exhibit 6. Average Score Differences by Gender for 13 Countries Participating in TIMSS Grade 4 and Grade 8 and PISA Age 15, 2003

Country	TIMSS GRADE 4		TIMSS GRADE 8		PISA AGE 15	
	Girls-Boys	Rank	Girls-Boys	Rank	Girls-Boys	Rank
AUS	-3	5	-13*	12	-5	3
BEL	-2	4	-11*	11	-8	7
HKG	0	2	2	5	-4	2
HUN	-3	5	-7	9	-8*	7
ITL	-9*	12	-6*	7	-18*	13
JPN	-4	7	-3	6	-8	7
LAT	1	1	6	1	-3	1
NLD	-6*	10	-7	9	-5	3
NZL	0	2	3	2	-14*	12
NOR	-5	9	3	2	-6	5
RUS	-4	7	3	2	-10*	10
USA	-8*	11	-6*	7	-6*	5
AVG1	-4	—	-3	—	-81	—
Scorecor (4,8)=.41; Rnkcor (4,8) = .27; Scorecor (8,15) = .04; Rnkcor (8,15) = .01						

*Significant at .05 level

[1]From the information available, we are not able to compute the statistical significance of the average girl-boys differences across the 12 countries.

Source: Mullis, Martin, Gonzalez, and Chrostowski, 2004; OECD, 2004

countries, there is no sign of a consistent gender gap on TIMSS-4 and TIMSS-8. In some countries, girls outscore boys; in other countries, boys outscore girls. In fact, the underlying absolute scores show that Hong Kong's girls on TIMSS-4 and TIMSS-8 score as well as or better than boys from Hong Kong or any of the other 11 countries.[xvi]

On PISA, gender differences somewhat favor boys over girls in all countries. The average difference between female and male scores is 8 points on PISA, about twice as high as it is on TIMSS-4 and TIMSS-8. We note that the 2003 PISA results for secondary school students are also similar to those observed a decade earlier, when the mathematics achievement of grade 12 girls on TIMSS was lower than that for boys in 12 of 15 countries (Hanna, Kundiger, and Larouche, 1990). The small but consistently negative differences on the 2003 PISA as contrasted with the mixed results on 2003 TIMSS suggest that gender differences worsen in secondary school.

The international survey evidence provides information to address two common explanations that have emerged from the gender literature about why girls' performance relative to boys' performance may worsen in secondary school (Fenema, 1996; Fenema, 2000). One explanation is that lower expectations among students, teachers, and parents for girls in mathematics lead males to become more confident than females about learning mathematics; furthermore that difference in confidence grows larger over the grades with student maturation (Fenema, 2000; Leder, 1992; Schwartz and Hanson, 1992). The TIMSS and PISA background questions allow us to examine the trends in boys' and girls' self-perceptions of their mathematics abilities over different grades and in different countries, and allow us to associate these differences with country performance.

Student responses to the 2003 PISA background questions regarding girls' and boys' attitudes toward mathematics were available for this paper, but the 2003 questions were not yet available for TIMSS-8. Instead, we used questions from the 1999 TIMSS-8. There is no particular reason to expect large international shifts over the recent period in males' and females' attitudes about mathematics, but if some shifts occurred, they would not be reflected in our analyses.

In as much as the TIMSS and PISA background questions were independently developed, wording of the attitudinal questions was similar, but not identical:

- "I am just not talented in mathematics" on TIMSS, grade 8 from 1999.
- "I am just not good in mathematics" on PISA, 2003.

Girls' answers to these questions showed a lower self-concept about their mathematics ability than did boys' answers on both TIMSS grade 8 and PISA (Exhibit 7). But a separate issue is whether girls' lower self-concept relative to boys' translates into lower mathematics performance by girls relative to boys.

The answer is "maybe" on TIMSS-8 because there was a correlation of -.56 between the difference in girls' and boys' self-view of their mathematics ability and the difference in achievement scores (Exhibit 7). Thus, as more girls relative to boys agreed that they were not talented in mathematics, the mathematics scores of girls worsened relative to those of boys. The correlation of -.56 was significant at .1, but not at the traditional higher standard of .05.

Exhibit 7. Comparison of Student Attitudes About Own Mathematics Ability on TIMSS Grade 8 (1999) and PISA (2003)

Country	TIMSS Grade 8: % Who strongly agree or agree that "I am just not talented in mathematics"			Girls' Minus Boys' TIMSS-8 Mathematics Score	PISA Age 15: % Who strongly agree or agree that "I am just not good in mathematics"			Girls' Minus Boys' PISA Score
	Boys	Girls	Girls Minus Boys		Boys	Girls	Girls Minus Boys	
AUS	33.2	44.7	11.6	-13	25.4	38.8	13.4	-5
BEL	39.2	50.0	10.8	-11	30.9	43.0	12.1	-8
HKG	38.1	47.4	9.3	2	50.3	62.6	12.4	-4
HUN	36.8	43.1	6.3	-7	40.7	48.4	7.7	-8
ITL	32.9	39.5	6.6	-6	48.2	50.7	2.5	-18
JPN	21.2	33.1	11.9	-3	45.7	58.4	12.6	-8
LAT	47.5	53.4	6.0	6	33.3	43.9	10.6	-3
NLD	27.3	41.3	14.0	-7	27.6	46.8	19.2	-5
NZL	35.3	44.2	8.9	3	26.6	39.2	12.6	-14
NOR	—	—	—	—	36.0	51.7	15.7	-6
RUS	27.9	27.3	-0.6	3	36.3	37.2	0.9	-10
USA	29.1	37.1	8.0	-6	30.4	40.2	9.8	-6
AVG	33.5	41.9	8.4	—	35.9	46.3	10.3	—
	Correlation of TIMSS-8 gender difference in attitudes (col. 3) with TIMSS-8 gender difference in score (col. 4) = -.56				Correlation of PISA gender difference in attitudes (col. 7) with PISA gender difference in score (col. 8) = .37			

Source: TIMSS 1999 results computed using AIR Lighthouse database and software available at: http://lighthouse.air.org. PISA 2003 results computed from: PISA 2003 international database (www.pisa.oced.org)

On PISA the answer is "no" and, in fact, the correlation turned positive, although not statistically significant. The reason for this anomalous association is unclear. Given the statistical insignificance of the correlation, we do not find empirical support from the international studies to suggest that a worsening of girls' self-perception of their mathematics ability is a likely explanation for their relatively poorer performance on PISA.

A second explanation from the gender literature that could explain an increasing girls' mathematics disadvantage in the upper grades is that girls' poorer scores are concentrated on more cognitively complex items (Fenema, 2000). Because cognitive complexity increases as mathematical assessments progress over grades/age, the girls' disadvantage relative to boys' would be expected to show up on the more cognitively complex PISA items.

To test this hypothesis, we approximated item complexity by the difficulty of an item as measured by its percentage of correct responses (as in Exhibit 4). The average girls' and boys' percentage of correct answers are computed for all items in the least difficult and the most difficult categories on each assessment (Exhibit 8). On TIMSS-4 and TIMSS-8, girls do about as well as boys on both the least difficult and most difficult items (i.e., within one percentage point). However, on

Exhibit 8. Comparison of the Males and Females Percent of Correct Answers on the Least Difficult and Most Difficult Items, TIMSS-4, TIMSS-8, and PISA

	Least Difficult Item Quartile			Most Difficult Item Quartile		
	Boys	Girls	Girls-Boys	Boys	Girls	Girls-Boys
TIMSS-4	80.3	81.0	0.7	35.4	34.4	-1.0
TIMSS-8	73.5	73.4	-0.2	28.8	28.2	-0.6
PISA	75.2	74.1	-1.1	27.4	23.4	-4.0
Source: Mullis, Martin, Gonzalez, and Chrostowski, 2004; OECD, 2004						

PISA, girls score an average of 4 points lower on the most difficult items compared with only a 1-point differential on the least difficult PISA items. The international results are consistent with the finding from other research that girls do somewhat more poorly as assessment items become more difficult in the upper grades.

Suggested explanations include:

- More-difficult items tend to be applications and tend to use experiences that are more familiar to males.

- Girls use different strategies for solving mathematics problems than do boys, and those strategies do not work as well on more complex problems.

It is not possible to examine these explanations from the international data.

In conclusion, the U.S. gender gap in mathematics scores is small, and although its magnitude does not worsen, it is more persistent compared with that in other countries except Italy. Overall, no systematic gender gap is found on TIMSS-4 and TIMSS-8, but a small consistent advantage favoring boys surfaces in the PISA results. The common explanation of increasing negative attitudes among girls about their mathematics ability does not seem to explain the PISA results, but there is support for the hypothesis that girls do somewhat poorer relative to boys on the more cognitively complex items.

VI. INSTRUCTIONAL FACTORS

Because teachers were not surveyed in PISA, it is not possible to extract information about mathematics instruction in relation to the PISA assessment. TIMSS, however, surveys a large number of a country's instructional characteristics, and our examination focuses on curricula and teacher characteristics particularly interesting to U.S. mathematics concerns in four areas: existence of a national mathematics curriculum, curriculum exposure, topic coverage, and teacher postsecondary education. The U.S. result for each factor is compared with the average result for the 11 comparison countries (Exhibit 9). Bivariate correlations of instructional factors with international scores across countries are notoriously invalid causal predictors and are not shown (NAGB, 2003).

Only 3 of the 12 countries, Australia, Belgium, and the United States, do not have national mathematics curriculum, although Belgium has a national test that acts as de facto standards (Exhibit 9). A national mathematics curriculum does not guarantee high performance (Italy is a good example), but conversely, in the absence of a national mathematics curriculum, the U.S. has 50 separate state curriculums. The U.S. National Council of Teachers of Mathematics (NCTM) standards may serve as a de facto curriculum, but it is unofficial and states' mathematics curricula differ considerably in their topic coverage grade by grade (Ginsburg et al., 2005).

Absent a national mathematics curriculum, U.S. textbook publishers are able to market to multiple states only by being inclusive in their topic coverage. As such, their textbooks cover more than double the number of topics at each elementary grade, as do textbooks in high TIMSS-scoring Singapore (Ginsburg et al., 2005). With so many topics, U.S. teachers, in trying to follow the textbooks, rarely get much beyond teaching mathematical procedures and do not develop in their students a deep conceptual understanding of mathematics topics and their applications (Schmidt, Houang, and Cogan, 2002).

Further evidence of the broad topic coverage characterizing U.S. mathematics curriculum is indicated by the low percentage of TIMSS mathematics topics not included in the U.S. curriculum through grades 4 and 8 compared with the percentage of TIMSS topics that other countries do not include (Exhibit 9). On average, the U.S. curriculum omits only 17 percent of the TIMSS grade 4 topics compared with an average omission rate of 40 percent for the 11 comparison countries. The United States covers all but 2 percent of the TIMSS topics through grade 8 compared with a 25 percent noncoverage rate in the other countries. High-scoring Hong Kong's curriculum omits 48 percent of the TIMSS items through grade 4, and 18 percent through grade 8. Less topic coverage can be associated with higher scores on those topics covered because students have more time to master the content that is taught.

Researchers have cited a shortfall in U.S. students' instructional time compared with that of other countries as a major cause of lower U.S. performance (National Education Commission on Time and Learning, 1994), but the TIMSS results in mathematics do not support these claims. In the United States, total time spent on mathematics over a year is comparable with the average for the other 11 countries at grade 4 and somewhat above average at grade 8.

However, U.S. instructional time use among the five content areas is different from that of the other countries in grade 8, although not in grade 4. The United States devotes about half the time to its study of geometry—its weakest subject—that other countries spend. By contrast, the United States spends almost 50 percent more time studying algebra, where its performance is somewhat stronger than the average U.S. performance across all content areas. The United States also spends 50

Exhibit 9. Selected Features of U.S. and 11 Comparison Countries Mathematics Instruction: 2003 TIMSS-4 and TIMSS-8

Selected Instructional Factors	Grade 4		Grade 8	
	U.S.	11 Comparison Countries	U.S.	11 Comparison Countries
Has a national curriculum	No	Yes = 8 of 11	No	Yes = 8 of 11
Percentage of TIMSS mathematics topics not included in the curriculum	17	40	2	25
Average yearly mathematics instructional time in hours	147	147	135	123
Percentage of mathematics class time by subject				
• Numbers	38	45	22	20
• Algebra	19	15	41	29
• Measurement	13	15	10	11
• Geometry	13	12	15	27
• Data/statistics	15	10	12	10
Percentage of teachers who relate what students learn in mathematics to their daily lives in half the lessons or more	—	—	66	36
Percentage of students whose teachers reported that calculators are not permitted	31	53	—	—
Percentage of grade 4 teachers with a major or specialization in mathematics, or grade 8 teachers with a major in mathematics	28	51	48	65
Source: Mullis, Martin, Gonzalez, and Chrostowski, 2004; and OECD, 2004				

percent more time on grade 4 data/statistics, another area of U.S. strength, than the comparison countries. The distribution of time devoted to particular content areas rather than the total amount of instructional time invested, seems more related to U.S. mathematics performance.

Teachers in the United States are likely to devote more instructional time to making mathematics relevant to students. Two-thirds of U.S. teachers include the relevance of mathematics as a topic in at least half their classes, which is nearly double the average percentage of classes that the 11 comparison countries devote to tying mathematics instruction to students' daily lives. This evidence, along with the evidence that data and statistics are emphasized in the United States, suggests that the United States must do more than simply increase its emphasis on real-world mathematics problems as a way to improve its international standings, at least not in the way real-world applications are currently introduced in the classroom.[xvii] It is relevant that PISA's real-world applications stress students' understanding and application of mathematics concepts (e.g., high percentage of open-ended problems) and not mechanical solutions to simple mathematics problems presented in a real-world context.

Research on whether calculator use is detrimental to developing arithmetic skills or promotes advanced skills by simplifying computations is inconclusive (Loveless, 2004; National Research Council, 2001). Still, it is important to note that teachers in most countries, including the highest

scorers, choose to limit calculator use in the primary grades more than U.S. teachers do. Only 31 percent of U.S. grade 4 teachers prohibit the use of calculators, but 53 percent of the teachers in the comparison countries disallow calculators (Exhibit 9).

Turning to teacher preparation, we note that U.S. mathematics teachers are less likely to have specialized college preparation in mathematics than are their peers in comparison countries. Although having content knowledge is not sufficient for becoming a good teacher if pedagogical skills are absent, neither are pedagogical skills sufficient without content knowledge (National Research Council, 2001). Yet only 28 percent of grade 4 teachers in the United States have a college specialization in mathematics, compared with 51 percent for the 11 comparison countries. At grade 8, fewer than half of U.S. teachers of mathematics majored in mathematics compared with two-thirds for the comparison countries.

Collectively, the comparative international findings related to instructional factors indicate differences in the U.S. mathematics system of instruction with respect to its lack of a uniform curriculum, insufficient topic focus, limited instructional time in measurement and geometry, and weaker teacher mathematics content knowledge—characteristics that warrant further follow-up in efforts to improve U.S. mathematics performance.

VII. POLICY IMPLICATIONS FOR CONSIDERATION

International comparisons at different stages of students' mathematical development across a common set of countries offer a unique laboratory for examining the performance of different mathematics education systems. However, because of the many differences that exist between the mathematics education systems found in the United States and in other countries, the policy directions that are suggested by the international findings must be tentative and their applicability to improving U.S. mathematics systems requires more detailed investigation. The findings discussed above suggest that further investigations, based on the international comparisons, should examine six policy directions to improve U.S. mathematics instruction.

First, the U.S. mathematics system must do a better job of establishing a strong foundation of students' initial mathematics knowledge in the early grades. Strengthening mathematics in the early grades is important because, contrary to impressions stemming from published reports of U.S. international mathematics results, U.S. students do not perform better at grade 4 than at grade 8, and do not perform better at grade 8 than at age 15 once the comparison is made based on a fixed set of largely industrialized nations. The strong association observed in most countries between their students' performance in grade 4 compared with later grades and ages further reinforces our conclusion about the importance of building a strong initial foundation in mathematics learning.

Second, the international comparisons support a balanced development of U.S. students' mathematics skills that includes mastery of less cognitively demanding mathematics facts and procedures along with more cognitively demanding mathematical reasoning. Students should also be exposed to a broad range of problems of varying mathematical difficulty.

We base these conclusions on overall results that show that countries in which students perform well on the more mathematically rigorous problems are also the countries whose students perform well on mathematics problems that involve more routine mathematics skills. The ability to solve problems at different levels of cognitive rigor appears to involve complementary mathematics skills that collectively contribute to attaining high levels of mathematics performance.

Third, weak U.S. performance in measurement and geometry in the elementary and middle grades needs to be addressed. We suggest that improving performance in measurement should be a particular focus in the early elementary grades through grade 4, and that geometry should be the particular focus of the middle grades because U.S. scores were particularly weak in these areas and grades. The fact that U.S. classrooms devote much less time at grade 8 to geometry than do most other countries further supports emphasizing geometry more in middle school mathematics.

Fourth, the mathematics preparation and knowledge characterizing U.S. grade 4 and grade 8 mathematics teachers should be strengthened. Fewer teachers of mathematics in the United States have a mathematics specialization at grade 4 or a major in mathematics at grade 8 than in other countries. Furthermore, elementary teachers who are better at mathematics should specialize in teaching mathematics through all the elementary grades beginning in first grade. This form of early-grade teacher specialization is practiced in the major urban areas of China with success (Ma, 1999), and it already occurs informally in the United States in the form of team teaching.

Fifth, the U.S. mathematics system is unlikely to improve its PISA results by simply devoting more time to real-world applications of problems if they continue to taught in the same way. Although PISA is a test of mathematical applications and the United States did not perform well on

PISA, the evidence from the common country comparisons suggests that U.S. students performed no worse on PISA than on TIMSS-8, an assessment of more traditional mathematics skills. Moreover, U.S. teachers are already more likely to devote more lessons that include real-world applications than are teachers in other countries. Our analyses suggest that the best way to guarantee success for 15-year-olds is to provide students with a solid foundation of mathematics in the primary and middle grades. However, the analyses of background factors indicate that the United States covers a much higher percentage of the TIMSS mathematics topics than do other countries, thereby diluting the intensity with which mathematics topics are taught in the primary and middle grades. Increasing topic intensity would also help scores on PISA, which includes a higher percentage of items that are open-ended and generally more cognitively demanding (Mullis et al., 2004; OECD, 2004).

Sixth, the U.S. mathematics system should explore strategies for strengthening girls' mathematics results, which—although only about one-tenth of a standard deviation lower than boys' results—are consistently negative. The United States and Italy were the only countries to show a female gap across all three assessments, although it is noteworthy that the size of the U.S. girls' performance disadvantage did not worsen across successive assessments, as it did in some other countries. Overall, the international evidence does not support even a small, consistent gender gap in grades 4 and 8. High-performing Hong Kong girls in grades 4 and 8 are able to perform as well as boys in all countries. However, our international analyses do suggest that females' mathematics results worsen on PISA at age 15, and that their lower scores are concentrated on the most mathematically difficult items. These results are consistent with the gender literature, but the reasons for the girls' lower scores on the more difficult mathematics items are unclear and more research is recommended into the causes.

In summary, the integrated analysis of the TIMSS and PISA mathematics results have shown that much can be learned about U.S. and other countries' mathematics performance at different stages of students' mathematics development. But the TIMSS–PISA comparisons can produce misleading findings if the analyses are not conducted over a common set of countries taking all three international assessments.

REFERENCES

Brolin, H., and Björk, L-E. (1992). Introducing calculators in Swedish schools. In J.T. Fey and C.R. Hirsch (Eds.), *Calculators in mathematics education* (1992 Yearbook of the National Council of Teachers of Mathematics, pp. 226–232). Reston, VA: NCTM.

Business Roundtable (2005). *Tapping Americas potential: The education for innovation initiative.* Retrieved October, 2005 from: http://www.businessroundtable.org/publications/index.aspx.

Cavenagh, S. (2004). "U.S. gets better showing on latest international math and science exam." *Education Week.* December 14.

Cavenagh, S. and Robelen, E. (2004). "U.S. students fare poorly in international math comparison." *Education Week.* December 7.

Crocker, L. and Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, Calif: Holt, Rinehart, and Winston.

Fenema, E. (1996). Mathematics, gender, and research. In G. Hanna (Ed.), *Towards gender equity in mathematics education* (pp.9–26). Amsterdam: Kluwer.

Fenema, E. (2000). *Gender and mathematics: what is known and what do I wish was known?* Prepared for the fifth annual forum of the National Institute of Science Education, May 22–23, 2000, Detroit, Michigan.

Ginsburg, A., Leinwand, S., Anstrom, T., and Pollock, E. (2005). What the United States can learn From Singapore's world-class mathematics system (and what Singapore can learn from the United States). Washington, DC: American Institutes for Research.

Groves, S. (1994). The effect of calculator use on third and fourth graders' computation and choice of calculating device. In J. da Ponte and J.F. Matos (Eds.), *Proceedings of the Eighteenth International Conference for the Psychology of Mathematics Education* (vol. 3, pp. 33–40). Lisbon, Portugal: PME Program Committee. (ERIC Document Reproduction Service No. ED 383–537).

Hanna, G., Kundiger, E., and Larouche, C. (1990). Mathematical achievement of grade 12 girls in fifteen countries. In L. Burton (Ed.), *Gender and mathematics: An international perspective.* London: Cassell Educational Ltd.

Hembree, R. and Stein, M. (1997). Research on calculators in mathematics education. In J. Fey and C. Hirsch (Eds.), *Calculators in mathematics education* (pp. 23–32). Reston, VA: National Council of Teachers of Mathematics.

Leder, G.C. (1992). Mathematics and gender: Changing perspectives. In D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning.* New York: McMillian.

Loveless, T. (2004). *Computation Skills, Calculators, and Achievement Gaps: An Analysis of NAEP Items.* Paper presented at AERA annual conference.

Ma, Liping. (1999). *Knowing and Teaching Elementary School Mathematics*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mullis, I., Martin, M., Smith, T., Garden, R., Gregory, K., Gonzalez, E., Chrostowski, S., and O'Connor, K. (2003). *TIMSS assessment frameworks and specifications 2003.* International Association for the Evaluation of Educational Achievement (IEA): International Study Center, Lynch School of Education, Boston College. Retrieved May, 2005 from: http://timss.bc.edu/timss2003i/PDF/t03_AF_preface.pdf.

Mullis, I., Martin, M., Gonzalez, E., and Chrostowski, S. (2004). *TIMSS 2003 international mathematics report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved May, 2005 from: http://timss.bc.edu/PDF/t03_download/T03_M_Front.pdf

National Academy of Sciences (2005). *Rising above the gathering storm: energizing and employing america for a brighter economic future*. Committee on Prospering in the Global Economy of the 21st Century: An Agenda for American Science and Technology, National Academy of Sciences, National Academy of Engineering, Institute of Medicine. Retrieved October, 2005 from: http://www.nap.edu/catalog/11463.html.

National Center for Education Statistics (NCES). (2005). *The condition of education 2005*, NCES 2005-094, Washington, DC: U.S. Government Printing Office. Retrieved October, 2005 from http://nces.ed.gov/programs/coe/2005/pdf/11_2005.pdf.

National Council of Teachers of Mathematics (2000). *Principals and standards for schools mathematics.* Reston, VA. Available: http://www.nctm.org/standards.

National Education Commission on Time and Learning (1994).*Prisoners of Time.* U.S. Government Printing Office. Retrieved November, 2005 from: http://www.ed.gov/pubs/PrisonersOfTime/TitlePage.html.

National Research Council. (2001). *Adding it up: Helping children learn mathematics.* J. Kilpatrick, J. Swafford, and B. Findell (Eds.). Washington, DC: National Academy Press.

OECD (2003). *PISA 2003 technical report.* Paris. Retrieved May, 2005 from http://www.pisa.oecd.org/dataoecd/49/60/35188570.pdf.

OECD (2004). *Learning for tomorrow's world. First results from PISA 2003.* Paris. Retrieved May 2005 from: http://www.pisa.oecd.org/dataoecd/1/60/34002216.pdf

Rosenthal, R. and Rosnow, R. (1984). *Essentials of behavioral analysis: Methods and data analysis. New York:* McGraw-Hill.

Schmidt, W., Houang, R., and Cogan, L. (2002). A coherent curriculum. *American Educator*: Summer 2002. pp. 1–17.

Schwartz, W and Hanson, K. (1992). *Equal mathematics education for female students*, Educational Developmental Center, Inc., Newton, MA. Center for Equity and Cultural Diversity.

Shuard, H. (1992). CAN: Calculator use in the primary grades in England and Wales. In J. Fey and C. Hirsch (Eds.), *Calculators in mathematics education* (1992 Yearbook of the National Council of Teachers of Mathematics, pp. 33–45). Reston, VA: NCTM.

Summers, L. (2005). *Remarks at NBER conference on diversifying the science & engineering workforce.* Cambridge, Mass. January 14. Retrieved June, 2005 from: http://www.president.harvard.edu/speeches/2005/nber.html.

U.S. Central Intelligence Agency CIA (2005). *World Fact Book.* Washington, DC: Author. Retrieved July, 2005 from: http://www.cia.gov/cia/publications/factbook/index.html.

U.S. Department of Education. ED (2004a). *U.S. Students Show Improvement in International Mathematics and Science Assessment. Press release of December 6.*

U.S. Department of Education. ED(2004b). *PISA Results Show Need for High School Reform. Press release of December 14.*

NOTES

[i] Countries that were statistically higher than the United States on 2003 TIMSS-4 were Belgium, Chinese Taipei, England, Hong Kong SAR, Japan, Latvia, Lithuania, Netherlands, the Russian Federation, and Singapore. Countries that were statistically lower than the United States on 2003 TIMSS-4 were Armenia, Australia, Cyprus, Iran, Italy, Moldova, Morocco, New Zealand, Norway, Philippines, Scotland, Slovenia, and Tunisia.

[ii] Countries that were statistically higher than the United States on 2003, TIMSS-8 were Belgium-Flemish, Chinese Taipei, Estonia, Hong Kong SAR, Hungary, 1Japan, Korea, Netherlands, and Singapore. Countries that were statistically lower than the United States on 2003, TIMSS-8 were Armenia, Bahrain, Botswana, Bulgaria, Chile, Cyprus Egypt, Ghana, Indonesia, Iran, Italy, Jordan, Lebanon, Macedonia, Moldova, Morocco Norway, Palestinian, Philippines, Romania, Saudi Arabia, Serbia, Slovenia, South Africa, and Tunisia. Countries which not statistically different than the United States on 2003, TIMSS-8 were Australia, Israel, Latvia, Lithuania, Malaysia, New Zealand, the Russian Federation, Scotland, Slovak Republic, and Sweden.

[iii] Countries that were statistically higher than the United States on 2003, PISA were Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, France, Hong Kong-China, Germany, Iceland, Ireland, Japan, Korea, Liechtenstein, Luxembourg, Macao-China, Netherlands, New Zealand, Norway, Slovak Republic, Sweden, and Switzerland. Countries that were statistically lower than the United States on 2003, PISA were Brazil, Greece, Indonesia, Italy, Portugal, Thailand, Tunisia, Turkey, and Uruguay. Countries that were statistically the same as the United States on 2003 PISA were Hungary, Latvia, Poland, and Spain.

[iv] Countries below the world average per capita income for 2003 (CIA World Facts, 2005) were excluded from our analysis, as nonindustrialized. The only such country was Tunisia, which also participated in all three assessments. Including Tunisia would have the effect of increasing the correlations between scores and ranks on successive assessments.

[v] OECD, 2004 (p. 39)

[vi] The 24 countries that participated in 2003 TIMMS-4 were Armenia, Australia, Belgium (Flemish), Chinese Taipei, Cyprus, England, Hong Kong, SAR, Hungary, Iran, Italy, Japan, Latvia, Lithuania, Moldova, Morocco, Netherlands, New Zealand, Norway, Philippines, Russian Federation, Scotland, Singapore, Slovenia, Tunisia, United States.

[vii] The 45 countries that participated in 2003 TIMSS-8 were Argentina, Armenia, Australia, Bahrain, Belgium (Flemish), Botswana, Bulgaria, Chile, Chinese Taipei, Cyprus, Egypt, England, Estonia, Ghana, Hong Kong, SAR, Hungary, Indonesia, Iran, Israel, Italy, Japan, Jordan, Latvia, Lebanon, Lithuania, Macedonia, Lithuania, Moldova, Morocco, Netherlands, New Zealand, Norway, Palestinian Nat'l Auth., Philippines, Romania, Russian Federation, Saudia Arabia, Scotland, Serbia, Singapore, Slovenia, South Africa, Sweden, Tunisia, United States.

[viii] The 40 countries that participated in PISA were the OECD countries of Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Korea, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Republic, Spain, Sweden, Switzerland, Turkey, and the United States, and the

partner non-OECD countries of Brazil, Hong-Kong-China, Indonesia, Latvia, Lichtenstein, Maco-China, Russian Federation, Serbia, Thailand, Tunisia, and Uruguay.

[ix] TIMSS uses the term cognitive domain to define the groups of skills that students use to solve mathematics problems. PISA describes skills in terms of competency clusters. In this paper, we use the term cognitive skills to include both TIMSS's cognitive skills and PISA'a competency categories.

[x] PISA does not perceive itself as an assessment of the school mathematics curriculum, but an assessment of the application of mathematics to real-world problem solving. Therefore, PISA perceives itself as having less need to collect teacher background information.

[xi] The TIMSS scale is set to have a mean of 500 in 1995. However, our 2003 analyses do not compare changes in scores with TIMSS scores in prior years and the 2003 analyses only depend on score rankings and correlations, so they should not be affected by the use of the 1995 standardized scale.

[xii] The 21 countries that are common to TIMSS-4 and TIMSS-8 are: Armenia, Belgium, Chinese Taipei, Cyprus, Hong Kong SAR, Hungary, Iran, Italy, Japan, Latvia, Lithuania, Moldovia, Morocco, Netherlands, Norway, Philippines, Russian Fed, Singapore, Slovenia, Tunisia, and U.S. The U.S. rank is 11 on TIMSS-4 and 10 on TIMSS-8 and the rank order correlation is .95. Among the 14 industrialized countries (i.e. above the world 2005 average per capita income) the U.S. rank is 11 on TIMSS-4 and 10 on TIMSS-8.

The 16 countries that are common to TIMSS-8 and PISA are: Belgium, Hong Kong SAR, Hungary, Indonesia, Italy, Japan, Korea, Rep, Latvia, Netherlands, Norway, Russian Federation, Serbia, Slovak Republic, Sweden, Tunisia, and US. Among the 16 countries, the U.S. rank is 10th on TIMSS-8 and 10th on PISA. Among the 13 industrialized nations participating in TIMSS-8 and PISA, the U.S. rank remains 10th on TIMSS-8 and 10th on PISA. The rank order correlation is a .85, which shows a strong correlation across all 16 countries between their ranks on TIMSS-8 and PISA.

[xiii] The correlations between the country scale scores and the percentage correct are .99 for TIMSS-4, .99 for TIMSS-8, and .98 for PISA.

[xiv] Note that the computation of the percent correct is based on the percent of students that received full credit, and does not take into account the fact that students may receive partial credit on some of the items.

[xv] The international assessments offer two methods for computing item difficulty. One uses the value of the difficulty parameter that is calculated as part of the Item Response Theory (IRT) process (Martin, Mullis, and Chrostowski, 2004b;). This IRT difficulty parameter would be the preferred choice if we were examining the difficulty of items in only a single assessment. The drawback to the IRT difficulty parameters is that they are calculated on a scale that is appropriate to each assessment, but do not permit comparisons of the difficulty of items across assessments. For this reason, this study applies the classical test methodology (Crocker and Algina, 1986) that measures the difficulty of an assessment item by the proportion of correct responses on that item.

[xvi] Hong Kong girls' average score is 575 on TIMSS-4, which is the same as Hong Kong boys' score, and higher than the boys scores of any other country taking TIMSS-4. On TIMSS-8, Hong Kong

girls score 587, which is higher than that of Hong Kong boys (585) or the boys' score from any of the other 11 countries.

[xvii] A comparison of Singapore and U.S. mathematics textbooks found that U.S. textbooks often introduced real-world examples through pictures of real-world situations or very simple problems that related to a mathematics concept, but they did not directly introduce the messiness or complexity that students would face in confronting actual real-world mathematics situations (Ginsburg, Leinwand, Anstrom, and Pollock, 2005).