Comparing TIMSS with NAEP and PISA in Mathematics and Science

The purpose of this document is to provide background information that will be useful in interpreting the 2007 results from the Trends in International Mathematics and Science Study (TIMSS) by comparing its design, features, framework, and items with those of the U.S. National Assessment of Educational Progress and another international assessment in which the United States participates, the Program for International Student Assessment (PISA).

Background

Reporting results to provide a comprehensive picture of how U.S. students perform in key subject areas is one of the objectives of the National Center for Education Statistics (NCES). In the United States, nationally representative data on student achievement come primarily from two sources: the National Assessment of Educational Progress (NAEP)—also known as the "Nation's Report Card"—and U.S. participation in international assessments, such as the Trends in International Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA).[1]

NAEP measures fourth-, eighth-, and twelfth-grade students' performance in mathematics and science, among other subjects, with assessments designed specifically for national and state information needs. Alternatively, the international assessments enable the United States to benchmark its performance to that of other countries—in fourth-grade and eighth-grade mathematics and science in TIMSS and in 15-year-old students' mathematics, science, and reading literacy in PISA. All three assessments are conducted regularly to allow the monitoring of student outcomes over time.[2]

While the international assessments may appear to have significant similarities with our national assessment program, such as the content areas studied and the ages and grade levels of students, each was designed to serve a different purpose and each is based on a separate and unique framework and set of assessment items (or questions). Thus, there may be differences in results for a given year or in trend estimates among the studies, each giving a slightly different view of U.S. students' performance in these subjects.

NCES is releasing the results from the 2007 administration of TIMSS in December 2008. Also available are results from 2007 for fourth- and eighth-grade mathematics in NAEP and from 2005 for fourth-, eighth- and twelfth-grade science and twelfth-grade mathematics in NAEP.[3] Additionally, PISA 2006 results for 15-year-old students in mathematics and science literacy are available. This document is intended to provide information that will help the public understand the results across studies, grasp the similarities and differences among the studies and results, and identify what each assessment contributes to overall knowledge on student performance.

Comparing Features of the Assessments

TIMSS differs from NAEP and PISA on several key features, including purpose, partners, population, precision of estimates, and content.

Purpose and proximity to curriculum

The goals of the assessments have subtle but important distinctions with regard to U.S. curricula.

NAEP is the primary source for information on mathematics and science achievement in the United States at key stages of education using nationally established benchmarks of performance (e.g., achievement levels of basic, proficient, and advanced). The frameworks and benchmarks are established by the National Assessment Governing Board (NAGB) and are based on the collaborative input of a wide range of experts and participants from government, education, business, and public sectors in the United States. Ultimately, the frameworks and benchmarks are intended to reflect the best thinking about the knowledge, skills, and competencies needed by U.S. students in these subjects at different grades.

TIMSS is a major source for internationally comparative information on the mathematics and science achievement of students in the fourth and eighth grades and on related contextual aspects such as mathematics and science curricula and classroom practices across countries, including the United States. The TIMSS frameworks and specifications are developed in a collaborative process involving international mathematics and science experts, as well as the national research coordinators from each participating country, and thus reflect recent developments and consensus in the international research community and the interests of a wide range of countries.

PISA is the primary source for internationally comparative information on the mathematics and science literacy of students in the upper grades at an age that, for most countries, is near the end of compulsory schooling. The objective of PISA is to measure the "yield" of education systems, or the skills and competencies students have acquired and can apply in these subjects to real-world contexts by age 15. The literacy concept emphasizes the mastery of processes, understanding of concepts, and application of knowledge in various situations within subject matter domains. By focusing on literacy, PISA draws not only from school curricula but also from learning that may occur outside of school.

The tailoring of NAEP to national practices distinguishes it from TIMSS and PISA, the content of which is determined internationally in collaboration with other countries. The focus in PISA on yield and the application of competencies in real-world contexts distinguishes it from TIMSS and NAEP, both of which aim at measuring school-based curricular attainment more closely.

Partners

The international assessments provide benchmarks with different groups of countries.

The TIMSS assessment and the PISA assessment differ in country participation. The sponsorship for TIMSS is the International Association for the Evaluation of Educational Achievement (IEA), which includes in its assessments a diverse group of countries and jurisdictions. The Organization

for Economic Cooperation and Development (OECD) sponsors PISA, with its 30 member countries representing many of the world's most industrialized nations.

Both assessments are well subscribed. Thirty-six countries at grade four and 48 countries at grade eight participated in TIMSS 2007, including 17 countries that are members of the OECD (see table 1).[4] PISA 2006 reported scores for 57 countries and subnational jurisdictions, including all 30 OECD countries. Students from 33 countries (England and Scotland are part of the United Kingdom, and are therefore counted as a single country) participated in both assessments. Still, the international averages in TIMSS and PISA are based on different sets of countries.

Population

The students being studied may represent different groups.

TIMSS, PISA, and NAEP are all sample-based assessments—meaning that each program administers the assessment to a sample of U.S. students in such a way that the results can be generalized to the larger population. However, each assessment defines the population to which it generalizes differently. One distinction between TIMSS and NAEP, on the one hand, and PISA, on the other hand, is that the former use grade-based samples while PISA uses an age-based sample. These choices relate to the purposes of each program described earlier—NAEP and TIMSS to report on curricular achievement and PISA on the yield of systems toward the end of compulsory schooling.

- The NAEP target population is all students in fourth, eighth, and twelfth grades, and thus reflects the performance of U.S. students in these grades—most recently for fourth- and eighth-grade reading and mathematics in 2007, twelfth-grade mathematics in 2005, and all three grades in science in 2005.

- The TIMSS target population is all students in the grade corresponding to the fourth and eighth years of school, excluding kindergarten. For the United States this is the fourth and eighth grades. Thus, the most recent TIMSS results reflect the performance of U.S. fourth- and eighth-graders in 2007.

- The PISA target population is all 15-year-old students. Operationally in 2006, this included all students who were from 15 years and 3 months to 16 years and 2 months at the beginning of the testing period and who were enrolled in school, between the grades of seven and twelve. Because U.S. PISA data are collected in the fall of the school year, most 15-year-old students are tenth-graders. However, the sample includes students from other grades.

Table 1. Countries participating in TIMSS and PISA

	Both TIMSS 2007 and PISA 2006	TIMSS 2007 only	PISA 2006 only
OECD countries	Australia Austria Czech Republic Denmark Germany Hungary Italy Japan Korea, Republic of Netherlands New Zealand Norway Slovak Republic Sweden Turkey United Kingdom (as a single entity in PISA, as England and Scotland in TIMSS) United States	†	Belgium Canada Finland France Greece Iceland Ireland Luxembourg Mexico Poland Portugal Spain Switzerland
Other countries	Bulgaria Chinese Taipei Colombia Hong Kong-China Indonesia Israel Jordan Latvia Lithuania Qatar Romania Russian Federation Serbia, Republic of Slovenia Thailand Tunisia	Algeria Armenia Bahrain Bosnia and Herzegovina Botswana Cyprus Egypt El Salvador Georgia Ghana Iran Kazakhstan Kuwait Lebanon Malaysia Malta Mongolia[1] Morocco Oman Palestinian Nat'l Authority Saudi Arabia Singapore Syrian Arab Republic Ukraine Yemen	Argentina Azerbaijan Brazil Chile Croatia Estonia Kyrgyz Republic Liechtenstein Macao-China Montenegro, Republic of Uruguay

† Not applicable.

[1] Mongolia participated in TIMSS 2007 but, because the quality of its data was not well documented, it was not included in the main data displays of the international reports.

NOTE: The countries that participated in TIMSS 2007 shown in this table differ from the countries shown in the international TIMSS reports (Mullis et al. 2008; Martin et al. 2008). Eight other educational jurisdictions, or "benchmarking" entities, participated: the states of Massachusetts and Minnesota; the Canadian provinces of Alberta, British Columbia, Ontario, and Quebec; Dubai, United Arab Emirates; and the Basque country of Spain.

Thus for elementary and middle grades mathematics and science, the most recent TIMSS and NAEP results report on similar populations in the same academic year. The PISA population, on the other hand, is older than the TIMSS and NAEP eighth-graders and younger than the NAEP

twelfth-graders. PISA also assesses a different cohort in different years than NAEP or TIMSS. Taking this into account, perhaps the closest comparisons can be made between the NAEP 2005 eighth-grade and PISA 2006 15-year-old student cohorts (for science), some of the former of whom theoretically could have been part of the latter. However, all side-by-side comparisons of PISA with NAEP or TIMSS results should be viewed with these significant population and cohort differences in mind.

Precision of estimates

The assessments are designed to measure at different levels of precision.

TIMSS, NAEP, and PISA are all designed to provide valid and reliable measures of U.S. students' performance in the aggregate and for major subpopulations, as well as over time, and each study draws a sample sufficient for this purpose. In addition, at the fourth and eighth grades, NAEP is designed to provide measures of student performance at the state level and, for those districts participating in the Trial Urban District Assessment, at the district level as well.

Sample sizes are determined to balance needs for precision of estimates against burden to respondents. Because of NAEP's need for reporting at the state (and sometimes district) level, NAEP samples many more students than does TIMSS or PISA (table 2). The NAEP national sample comprises individual state samples of public school students, supplemented by a national sample of nonpublic school students. NAEP thus generally measures performance at a finer level of precision than TIMSS or PISA, and these differences can have an impact on the assessments' sensitivities in detecting changes in student performance.

Table 2. Number of students and schools participating in NAEP, TIMSS, and PISA

	No. of students	No. of schools
NAEP 2007 Mathematics (4th grade)	197,700	7,840
NAEP 2007 Mathematics (8th grade)	153,000	6,910
NAEP 2005 Science (4th grade)	147,700	8,500
NAEP 2005 Science (8th grade)	143,400	6,400
NAEP 2005 Science (12th grade)	13,700	900
TIMSS 2007 (4th grade)	7,900	260
TIMSS 2007 (8th grade)	7,400	240
PISA 2006	5,600	170

NOTE: Numbers have been rounded to the nearest hundred for students and the nearest 10 for schools.
SOURCES: National Center for Education Statistics, *The Nation's Report Card: Mathematics 2007* (NCES 2007-094); *The Nation's Report Card: Science 2005* (NCES 2006-466); *Highlights from TIMSS 2007: Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context* (NCES 2009-001); and *Highlights from PISA 2006: Performance of U.S. 15-Year-Old Students in Science and Mathematics Literacy in an International Context* (NCES 2008-016).

Content

The mathematics and science being assessed differ in terms of the ways in which the frameworks for assessment are organized and in terms of content coverage, item format, and other key features.

As noted before, the assessments under discussion here are developed from frameworks that define the domain and specify the content and skills to be measured.[5] Thus, a first task in comparing assessment programs is to compare how the frameworks and specifications are elaborated. A second task, which can provide a more in-depth view, is to compare how the frameworks are operationalized through the actual assessment items. This section draws on the results from two recent studies, which compared TIMSS and PISA with NAEP in terms of their frameworks and items.[6]

Framework Comparisons

TIMSS, NAEP, and PISA each are developed from frameworks specifying the content and skills to be measured.

The NAEP 2005/2007 and TIMSS 2007 **mathematics** frameworks are organized similarly. Both assessments include the same five main content areas, described generally as: number, measurement, geometry, data, and algebra. Within these content areas in both frameworks, more detailed topics and objectives are further specified by grade level. Both frameworks also include a cognitive dimension, although they define the cognitive dimension differently. The TIMSS cognitive dimension is categorically defined, with knowing, applying, and reasoning as the three categories. NAEP uses a hierarchical model called levels of mathematical complexity, with low, moderate, and high as the three levels. Close examination at the item level (as discussed in the next section) shows differences in operationalization of the content dimension.

The PISA mathematics framework includes content and cognitive dimensions, as well as a third dimension that describes the contexts in which mathematics is applied. For its content dimension, PISA uses overarching ideas (space and shape, change and relationships, quantity, and uncertainty) rather than curricular-based domains as in TIMSS and NAEP. For its cognitive dimension, PISA describes important mathematical competencies in three clusters: reproduction, connections, and reflection. Finally, the PISA framework explicitly names a context dimension that describes a variety of situations in which students should be able to use and do mathematics (e.g., in school, in society). The inclusion of three dimensions and the way in which they are defined reflects PISA's focus on literacy and real-world relevancy of the items. Closer examination of PISA items in comparison with TIMSS and NAEP makes the design differences apparent, although it also shows where content is similar despite its different organization in the frameworks.[7]

In **science**, the NAEP 2005 and TIMSS 2007 frameworks differ somewhat in organization. The NAEP framework defines science content in terms of three broad fields of science: physical science, life science, and earth science. TIMSS includes life science and earth science, but splits physical science into separate domains for physics and chemistry. Both science frameworks also include a cognitive dimension that, for both assessments (unlike mathematics), is categorically

defined. TIMSS identifies knowing, applying and reasoning as its three cognitive categories. NAEP identifies conceptual understanding, scientific investigation, and practical reasoning as three categories of "knowing and doing science." Also different from mathematics, both assessments' science frameworks have overarching dimensions that cross the content and cognitive dimensions. In NAEP these are nature of science and themes, as well as a number of items designated as hands-on performance tasks, which require students to conduct experiments using materials provided to them and record observations and conclusions by answering related test items. TIMSS includes an overarching dimension called scientific inquiry, which attempts to measure students' abilities to engage in (paper-and-pencil) inquiry tasks.

The PISA science literacy framework also has content and cognitive dimensions, although they are more broadly defined, capturing some of what is included in NAEP's and TIMSS's overarching dimensions. PISA's content dimension includes both knowledge of the natural world (in the fields of life systems, physical systems, Earth and space systems, and technology systems) and knowledge about science itself (scientific inquiry and scientific explanations). PISA's cognitive dimension describes important competencies required for scientific literacy: identifying scientific issues, explaining scientific phenomena, and using scientific evidence. In the PISA framework model, the competencies are prominent—they form the subscales for reporting, whereas in NAEP and TIMSS the subscales are by content area. The PISA framework also is explicit about the situation-based nature of science literacy and thus has a context dimension that describes a range of situations in which individuals deploy their competencies.[8]

Item Comparisons

In the most recent comparison studies, TIMSS 2007 mathematics and science items were classified to the NAEP assessment frameworks (2005/2007 for mathematics and 2005 for science) in terms of the content topics and objectives and grade level in order to allow a direct comparison of the two assessments.[9] In two other studies, PISA mathematics and science items also were placed on the NAEP frameworks, which allow content comparison of the two international assessments via the national frameworks.

Item content

Overall assessment scores can depend on the extent to which the proportion of the items devoted to various topics or skills aligns with the emphases of the education system's curriculum.

Even at the broadest level of the **mathematics** framework, content areas, NAEP 2007 and TIMSS 2007 differed in the distribution of items.[10] At the fourth grade, both TIMSS and NAEP had the highest proportion of items in the number content area (40 and 39 percent of items, respectively; table 3). However, NAEP had a higher percentage of measurement items than TIMSS, and TIMSS had a higher percentage of geometry items and a slightly higher percentage of data analysis and probability items. At the eighth grade, algebra was the content area with the highest proportion of items in both TIMSS and NAEP (30 and 27 percent, respectively). Key differences were the much greater emphasis NAEP placed on geometry and the much greater emphasis TIMSS placed on number at this grade level. PISA's assessment of 15-year-old students had a particularly high percentage of data analysis and probability items and a comparably low percentage classified as algebra compared to the TIMSS and NAEP eighth-grade assessments, which may not be surprising for an assessment targeted to older students.

Table 3. Percentage distribution of mathematics items across NAEP 2005/2007 mathematics content areas

NAEP mathematics content areas	Grade 4		Grade 8		15-year-olds
	NAEP 2007 (n=166)	TIMSS 2007 (n=162)	NAEP 2007 (n=168)	TIMSS 2007 (n=188)	PISA 2003/2006[1,] (n=85)
Number properties and operations	39	40	22	31	22
Measurement	21	14	17	14	18
Geometry	16	20	19	11	12
Data analysis and probability	12	15	15	15	40
Algebra	12	12	27	30	11
Multiple content areas[2]	0	1	0	1	2

[1] PISA 2006 mathematics items are a subset of PISA 2003 items. Data in this table for PISA mathematics items are taken from the 2003 comparison study. See endnote 7 for more detail.
[2] This row identifies the percentage of items that expert panelists classified to more than one content area, without being able to identify a primary classification in just one of the areas. The items classified to multiple areas also were counted with each relevant area; therefore the columns may sum to greater than 100 percent.
SOURCE: National Center for Education Statistics, previously unpublished tabulations from NAEP-TIMSS 2007 comparison study; and National Center for Education Statistics, *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments* (NCES 2006-029).

Another way to compare assessments is to examine the degree to which items can be fit to more detailed levels of a framework. Over 90 percent of TIMSS (and PISA) mathematics items could be placed in the NAEP mathematics framework at the finest level of detail, the objective level (data not shown). This indicates a relatively strong content match among the assessments. However, it also indicates that both TIMSS and PISA had some items that did not match well to the NAEP framework and thus would not be likely to be found in the NAEP assessment. These accounted for 2, 3, and 8 percent of TIMSS fourth-grade, TIMSS eighth-grade, and PISA mathematics items, respectively. In some cases, the TIMSS items lacked a real-world context, which kept it from fitting a NAEP objective. In other cases, TIMSS items were viewed as having a problem-solving component that went beyond the related NAEP objective. In terms of content, TIMSS had several fourth-grade items on finding the area of triangles, whereas the experts interpreted the NAEP framework as focusing more on other shapes. At the eighth grade, TIMSS included items on solving problems with angles, whereas the framework was interpreted as focusing mostly on the properties of right triangles. In items requiring solving problems using formulas, TIMSS items tended to use less common formulas than those named in the related NAEP objective (e.g., circumference of a circle).TIMSS also had a number of items on number patterns and items requiring drawing symmetrical figures, which were not typical for NAEP.

A third content-related analysis assessed the grade of the NAEP framework with which the TIMSS (and PISA) item topic was most consistent (regardless of whether or not that topic had been able to be classified to a specific topic or objective). The TIMSS mathematics item topics—particularly at the fourth grade—tended to be consistent with topics in the NAEP mathematics

framework at the corresponding grade (table 4). Eighty-six percent of fourth-grade TIMSS items were consistent with topics in the NAEP fourth-grade framework. At the eighth-grade level, 80 percent of TIMSS items were consistent with the NAEP eighth-grade framework, 9 percent with the fourth-grade framework, and 11 percent with the twelfth-grade framework. Additionally, there were small percentages of TIMSS mathematics items (1 percent in both fourth and eighth grades) that did not correspond well with topics or objectives in any grade level of the NAEP mathematics framework; these were all TIMSS geometry items. In PISA-NAEP comparisons, the PISA mathematics items most closely correspond with the NAEP eighth-grade framework.

Table 4. Percentage distribution of TIMSS and PISA mathematics items across NAEP 2005/2007 fourth-, eighth-, and twelfth-grade mathematics framework

	TIMSS 2007		PISA 2003/2006[1] (n=85)
	Grade 4 (n=162)	Grade 8 (n=188)	
NAEP 4th grade	86	9	12
NAEP 8th grade	12	80	85
NAEP 12th grade	1	11	4
No fit[2]	1	1	0

[1] PISA 2006 mathematics items are a subset of PISA 2003 items. Data in this table for PISA mathematics items are taken from the 2003 comparison study. See endnote 7 for more detail.
[2] This row identifies the percentage of items the expert panelists were unable to classify as generally consistent with the topics and objectives within a grade level of the NAEP framework.
NOTES: Data in this table show at which grade in the NAEP framework TIMSS and PISA items were considered generally consistent. The small percentage of fourth-grade items deemed most consistent with the NAEP twelfth-grade framework does not indicate that these items would be likely in twelfth-grade NAEP, but rather that the expert panel had a difficult time finding an explicit match in other parts of the framework, such as for those topics discussed in previous paragraphs. Additionally, for TIMSS, there were five multi-part items whose individual parts classified to different grade levels in the NAEP framework. In this table, however, these items were classified to only one grade: either the higher grade or the grade for which the individual parts were most frequently classified (in the case of items with 3 or more parts). Details may not sum to 100 percent because of rounding.
SOURCE: National Center for Education Statistics, previously unpublished tabulations from NAEP-TIMSS 2007 comparison study; and National Center for Education Statistics, *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments* (NCES 2006-029).

In **science**, NAEP 2005 and TIMSS 2007 differed in the distribution of items across even the broad content areas, as well. TIMSS had a greater emphasis than NAEP on life science in the fourth grade and on physical science in both the fourth and eighth grades (table 5). Conversely, NAEP had a greater emphasis than TIMSS on Earth science at both the fourth and eighth grades.

In comparison with PISA, the eighth-grade TIMSS assessment had a similar percentage of items in Earth science. Eighth-grade TIMSS had a lower percentage than PISA of life science items, however, and a greater percentage of physical science items.

There also were differences in the degree to which the TIMSS (and PISA) science items could be mapped to more detailed levels of the NAEP science frameworks (e.g., to subtopics or objectives), and these differences were larger than those for mathematics. Twelve percent of TIMSS fourth-grade and 20 percent of TIMSS eighth-grade science items could not be placed on the NAEP science framework at the finest level of detail, the objective level (data not shown). Five and 4

percent of the fourth- and eighth-grade items, respectively, could not be placed on the NAEP frameworks at the next level of detail, the subtopic level.[11] This suggests that there were a number of TIMSS science items that did not exemplify NAEP objectives and thus would not likely be found on a NAEP assessment. The notations by the expert panelists involved in the comparison study indicated that, for TIMSS, these included items on chemical change and reaction, complete circuits, reproduction, and health, which are not explicitly treated in the NAEP 2005 framework.

Table 5. Percentage distribution of science items across NAEP 2005 fields of science

NAEP Field of science	Grade 4		Grade 8		15-year-olds
	NAEP 2005 (n=153)	TIMSS 2007 (n=161)	NAEP 2005 (n=222)	TIMSS 2007 (n=194)	PISA 2006 (n=108)
Earth science	35	19	32	18	19
Life science	35	45	38	35	47
Physical science	29	37	30	47	31
Multiple content areas[1]	0	1	0	0	1
No fit[2]	0	1	0	0	5

[1] This row identifies the percentage of items that expert panelists classified to more than one field of science, without being able to identify a primary classification in just one of the fields. The items classified to multiple fields also were counted with each relevant field; therefore the columns may sum to greater than 100 percent.
[2] This row identifies the percentage of items that expert panelists were not able to classify to any of NAEP's fields of science, as defined by its framework.
SOURCE: National Center for Education Statistics, previously unpublished tabulations from NAEP-TIMSS 2007 comparison study and NAEP 2007-PISA 2006 science comparison study.

PISA stands out even more than eighth-grade TIMSS from the NAEP framework—46 percent of science items could not be placed within the objectives of the NAEP science frameworks—nearly all of which were not placed within a subtopic either (data not shown). This relates to the PISA framework's grounding in "literacy" and the inclusion of many items that focused on knowledge about science itself (scientific inquiry and explanations), without requiring specific science content knowledge.

Compared with mathematics, the TIMSS science item topics did not show as great a consistency with the topics in the NAEP framework at the corresponding grade level. Seventy-nine percent of TIMSS fourth-grade science items were consistent with topics in the NAEP fourth-grade science framework, while 12 percent were more consistent with the eighth-grade framework and 2 percent with the twelfth-grade framework (table 6). Six percent were not consistent with topics at the three grades assessed by NAEP. At the eighth grade, correspondence was lower: 56 percent of TIMSS science items were classified as consistent with the NAEP eighth-grade science framework. The remaining 44 percent of items were distributed across the NAEP fourth and twelfth grade frameworks or classified as not fitting the framework. This strongly suggests that there are some differences in terms of what each assessment defined as fourth-grade versus eighth-grade science content.

In the PISA comparisons, 83 percent of the science items for 15-year-olds were found to be generally consistent with topics in either the NAEP eighth-grade (41 percent) or twelfth-grade (42

percent) science frameworks, with the distribution split fairly evenly between the two. (Again, this statistic relies on the expert judgment of the panel and the *general* consistency of a PISA item topic with the NAEP grade-level framework, since many of the PISA items, as noted previously, did not fit with specific subtopics and objectives in NAEP.)

Table 6. Percentage distribution of TIMSS and PISA science items across NAEP 2005/2007 fourth-, eighth-, and twelfth-grade science framework

	TIMSS 2007		PISA 2006 (n=108)
	Grade 4 (n=161)	Grade 8 (n=194)	
NAEP 4th grade	79	16	9
NAEP 8th grade	12	56	41
NAEP 12th grade	2	15	42
No fit[1]	6	13	8

[1] This row identifies the percentage of items the expert panelists were unable to classify as generally consistent with the topics and objectives within a grade level of the NAEP framework.
NOTES: Data in this table show at which grade in the NAEP framework TIMSS and PISA items were considered generally consistent. The small percentage of fourth-grade items deemed most consistent with the NAEP twelfth-grade framework does not indicate that these items would be likely in twelfth-grade NAEP, but rather that the expert panel had a difficult time finding an explicit match in other parts of the framework. Additionally, for TIMSS, there were five multi-part items whose individual parts classified to different grade levels in the NAEP framework. In this table, however, these items were classified to only one grade: either the higher grade or the grade for which the individual parts were most frequently classified (in the case of items with 3 or more parts). Details may not sum to 100 percent because of rounding.
SOURCE: National Center for Education Statistics, previously unpublished tabulations from NAEP-TIMSS 2007 comparison study and NAEP 2007-PISA 2006 science comparison study.

Item format

A final area of comparison is in terms of item format: whether an item is multiple-choice or is open-ended, requiring the student to construct his or her own response. In mathematics, TIMSS and NAEP both have a majority of items that are multiple-choice, although NAEP has a higher percentage than TIMSS at both grades (table 7). Both NAEP and TIMSS eighth-grade assessments are in contrast to PISA, in which two-thirds of the mathematics items are constructed response. In science, the percentage distribution across item formats is more similar between NAEP and TIMSS, with 54 to 57 percent being multiple choice and 43 to 46 percent being constructed response (table 8). The item formats used in science in PISA are more similar to the NAEP and TIMSS eighth-grade assessments than in mathematics, with 62 percent multiple choice and 38 percent constructed response items.

Although item difficulty ranges within each of these categories, it is important to know that different assessments employ different balances of item types.

Table 7. Percentage distribution of mathematics items by item format

Item format	Grade 4		Grade 8		15-year-olds
	NAEP 2007 (n=166)	TIMSS 2007 (n=162)	NAEP 2007 (n=168)	TIMSS 2007 (n=188)	PISA 2003/2006[1] (n=85)
Multiple choice	69	57	74	62	33
Constructed response	31	43	26	38	67

[1] PISA 2006 mathematics items are a subset of PISA 2003 items. Data in this table for PISA mathematics items are taken from the 2003 comparison study. See endnote 7 for more detail. Multiple choice for PISA includes the category "complex multiple choice."
NOTE: For TIMSS, there were three multi-part items (all at fourth grade) whose individual parts were different formats. These items were classified as constructed response for the purposes of analysis.
SOURCE: Developers classifications for NAEP 2007 and TIMSS 2007; and National Center for Education Statistics, *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments* (NCES 2006-029).

Table 8. Percentage distribution of science items by item format

Item format	Grade 4		Grade 8		15-year-olds
	NAEP 2005 (n=153)	TIMSS 2007 (n=161)	NAEP 2005 (n=222)	TIMSS 2007 (n=194)	PISA 2006 (n=108)[1]
Multiple choice	56	57	57	54	62
Constructed response	44	43	43	46	38

[1] Multiple choice for PISA includes the category "complex multiple choice."
NOTE: For TIMSS, there were five multi-part items (two at fourth grade and three at eighth grade) whose individual parts were different formats. These items were classified as constructed response for the purposes of analysis.
SOURCE: Developers' classifications for NAEP 2005, TIMSS 2007, and PISA 2006.

Examining Results in the Context of the Distinctions among the Assessments

Both TIMSS and NAEP provide a measure of fourth- and eighth-grade mathematics and science. Both PISA and NAEP provide measures of mathematics and science performance for older students (grade twelve and 15 years old, respectively). It is natural to compare them, but the distinctions described previously need to be kept in mind in understanding the converging or diverging results.[12]

Comparing key results in mathematics

The most recent results from NAEP and TIMSS include information on trends over time in fourth- and eighth-grade mathematics for a largely similar time interval: in NAEP between 1996 and 2007 and in TIMSS between 1995 and 2007. For both grades, the trends shown by NAEP and TIMSS are largely consistent with one another. Both assessments showed statistically significant increases in the mathematics performance of fourth- and eighth-grade students: overall, among boys, and among girls. NAEP also reported increases for each of four racial/ethnic groups (non-Hispanic White, non-Hispanic Black, Hispanic, and non-Hispanic Asian),[13] for students at the top

and bottom extremes of the distribution (at the 10th and 90th percentiles), and for students receiving free and reduced-price lunch, at both grades.[14] TIMSS only detected increases in mathematics performance for some of these groups (e.g., White and Black students in both grades, students in the 10th percentile in both grades) and no change for others (e.g., students receiving free and reduced price lunch, Hispanic fourth-grade students). This is likely because of NAEP's larger sample sizes, which make it more sensitive to picking up small changes among nationally relevant subgroups than TIMSS, which is designed primarily to detect differences among countries.

Comparing key results in science

The most recent results from NAEP and TIMSS also provide trend information for fourth- and eighth-grade science, although covering a slightly shorter time interval in NAEP than in TIMSS. NAEP provides trends for the period 1996 to 2005 and TIMSS for the period 1995 to 2007. Compared with mathematics, the trends shown by NAEP and TIMSS in science are less consistent with one another, which is not surprising given the differing time periods and the relatively greater differences in the assessments discussed in the previous sections.

For example, in fourth grade, NAEP shows that there was an increase in students' science performance both overall and among boys between 1996 and 2005, whereas TIMSS did not detect any change in performance for either of those groups from 1995 to 2007. NAEP also reported increases in science performance for four of five racial/ethnic subgroups,[15] whereas TIMSS only reported increases for Black and Asian students in the fourth grade. At the eighth-grade level, neither NAEP nor TIMSS showed any change in science performance among students overall. But in contrast to the fourth-grade results, TIMSS reported increases for Black, Hispanic, and Asian eighth-grade students, whereas NAEP only reported increases among Black students. This suggests that Hispanic and Asian eighth-grade students performed relatively better on the content unique to TIMSS than unique to NAEP.

Comparing results in upper grades mathematics and science

It is more difficult to compare the results from NAEP and PISA in mathematics and science at the upper grades, not only because of the population and framework differences, but also because PISA is not yet reporting a trend measure for science and is only reporting a 3-year difference (2003 to 2006) for mathematics. The last assessment of mathematics in NAEP showed increases among eighth-graders from 1996 (as before), 2000, 2003, and 2005 to 2007. There also were slight decreases in mathematics among twelfth-graders from 1996 to 2000 on NAEP. PISA did not show any differences in the mathematics performance of U.S. 15-year-olds between 2003 and 2006. However, again, the population and timing of assessment in PISA does not correspond well with NAEP. The last assessment of science in NAEP (2005), showed no statistically significant differences in the performance of eighth- or twelfth-graders since 2000 (although there was a slight decrease since 1996 among the older students).

Summary

In reporting results on how U.S. students perform, NCES draws on multiple sources of data in order to capitalize on the complementary information presented in national and international assessments and thus offer a more comprehensive picture of U.S. student performance. NAEP

measures in detail the mathematics and science knowledge of U.S. students as a whole, and can also provide trend information for individual states, different geographic regions, and demographic population groups. International assessments like TIMSS (and PISA) add value by providing a method for comparing our performance in the United States to the performance of students in other nations.

Examining the assessments in depth, comparing their features, frameworks, and items, can provide specific information on how they are similar and different and allow us to understand how they complement one another. As described here, some of the differences between TIMSS, NAEP, and PISA include:

- The goals of the assessments have subtle but important distinctions with regard to the U.S. curricula. NAEP is tailored specifically to practices and standards operating in the United States, which distinguishes it from the other two assessments. The TIMSS and PISA content is determined internationally in collaboration with other countries and reflects consensus views of key content. Also, PISA's specific focus on the "yield" of the education system and the application of competencies in real-word contexts, distinguishes it from both NAEP and TIMSS, which aim at measuring school-based curricular attainment more closely.

- TIMSS and PISA provide benchmarks with different groups of countries. Thirty-six countries at grade four and 48 countries at grade eight participated in TIMSS 2007, 17 of which are industrialized OECD countries. Fifty-seven countries and subnational jurisdictions participated in PISA 2006, 30 of which are industrialized OECD countries. Thirty-three countries (England and Scotland are part of the United Kingdom and therefore counted as a single country) participated in both TIMSS and PISA.

- The students being assessed may represent different groups. Both NAEP and TIMSS use grade-based samples and both target fourth- and eighth-grade students. Both studies last assessed students in 2007 so their results can be generalized to the same population. PISA uses an age-based sample, which targets 15-year-old students, who are most often tenth-graders in the United States.

- The assessments are designed to measure student performance at different levels of precision. TIMSS, NAEP, and PISA are all designed to provide valid and reliable measures of U.S. students' performance in the aggregate and for major subpopulations, and each study draws a sample sufficient for this purpose. NAEP, however, is designed to also provide estimates for individual states, which requires an increased sample size. NAEP thus measures performance at a finer level of precision than TIMSS or PISA. These differences can have an impact on the assessments' sensitivities in detecting changes in student performance.

- The components of mathematics and science being assessed can differ in terms of the ways in which the frameworks for assessment are organized and in terms of content coverage, item format, and other key features. Examinations of the frameworks for TIMSS, NAEP, and PISA in mathematics and science showed areas of overlap and difference in terms of the content and skills being measured in the respective subject areas and grades. Item analyses showed that NAEP and TIMSS have more similarities in terms of the content dimension of items than do NAEP and PISA, in both mathematics and science. NAEP and TIMSS are more alike in mathematics than in science. In science, there are larger differences between NAEP and

TIMSS than in mathematics in terms of the distribution of items across content areas and the fit of items to corollary grade levels of the frameworks.

Because there are differences in the features, frameworks and items of the national and international assessments, direct comparisons among the assessments are not useful. Rather the results from different studies should be thought of as different lenses through which to view and better understand U.S. student performance.

Contact Information

Daniel J. McGrath
Director, International Activities Program
National Center for Education Statistics
U.S. Department of Education
1990 K Street, NW
Washington, DC 20006
Tel.: (202) 502-7426
E-mail: Daniel.McGrath@ed.gov

Useful Websites

NAEP	http://nces.ed.gov/nationsreportcard
TIMSS	http://www.timss.org (international)
	http://nces.ed.gov/timss (national)
PISA	http://www.pisa.oecd.org (international)
	http:// nces.ed.gov/surveys/pisa (national)

[1] TIMSS is conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA). PISA is sponsored by the Organization for Economic Cooperation and Development (OECD). The United States also participates in the Progress in International Reading Literacy Study (PIRLS), conducted under the auspices of the IEA. See "Comparing PIRLS and PISA with NAEP in Reading, Mathematics, and Science," available at http://nces.ed.gov/surveys/PISA/pdf/comppaper12082004.pdf for a comparison of the most recent PIRLS and NAEP.

[2] All statements about NAEP in this paper refer to main NAEP (versus long-term trend NAEP). NAEP currently assesses fourth- and eighth-grade mathematics every 2 years, and twelfth-grade mathematics, as well as science at all three grades, every 4 years. TIMSS is on a 4-year cycle. PISA is on a 3-year cycle.

[3] See:

Lee, J.; Grigg, W.S.; and Dion, G.S. (2007). *The Nation's Report Card: Mathematics 2007* (NCES 2007-494). U.S. Department of Education. Washington, D.C.: National Center for Education Statistics.

Grigg, .W; Donahue, P.L.; and Dion, G. (2007). *Nation's Report Card: 12th-Grade Reading and Mathematics 2005* (NCES 2007-468). U.S. Department of Education. Washington, D.C.: National Center for Education Statistics.

Grigg, W.; Lauko, M.A.; and Brockway, D.M. (2006). *Nation's Report Card: Science 2005* (NCES 2006-466). U.S. Department of Education. Washington, D.C.: National Center for Education Statistics.

Baldí, S.; Jin, Y.; Skemer, M.; Green, P.; Herget, D.; and Xie, H. (2007). *Highlights from PISA 2006: Performance of U.S. 15-Year-Olds in Science and Mathematics Literacy in an International Context* (NCES 2008-016). U.S. Department of Education. Washington, D.C.: National Center for Education Statistics.

[4] The total number of countries reported here differs from the total number reported in the international TIMSS reports (Mullis et al. 2008; Martin et al. 2008). In addition to the 36 countries at grade four and 48 countries at grade eight, 8 other educational jurisdictions, or "benchmarking" entities, participated: the states of Massachusetts and Minnesota; the Canadian provinces of Alberta, British Columbia, Ontario, and Quebec; Dubai, United Arab Emirates; and the Basque country of Spain. See: Mullis, I.V.S., Martin, M.O., and Foy, P. (2008). *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: Boston College; and Martin, M.O., Mullis, I.V.S., and Foy, P. (2008). *TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA: Boston College.

[5] See:

Mullis, I.V.S.; Martin, M.O.; Ruddock, G.J.; O'Sullivan, C.Y.; Arora, A.; and Erberber, E. (2005). *TIMSS 2007 Assessment Frameworks*. Chestnut Hill, MA: Lynch School of Education, Boston College.

National Assessment Governing Board (NAGB) (2006). *Mathematics Framework for the 2007 National Assessment of Educational Progress*. Washington, D.C.: Government Printing Office.

National Assessment Governing Board (NAGB) (2004). *Science Framework for the 2005 National Assessment of Educational Progress*. Washington, D.C.: Government Printing Office.

Organization for Economic Cooperation and Development (OECD) (2006). *Assessing Scientific, Reading, and Mathematical Literacy: A Framework for PISA 2006*. Paris: OECD.

[6] This paper draws primarily on comparison studies conducted in April and May 2008. These studies involved the convening of three separate expert panels to classify TIMSS 2007 mathematics items, TIMSS 2007 science items, and PISA 2006 science items to the relevant NAEP frameworks: 2005/2007 mathematics framework and 2005 science framework. Data on PISA mathematics items were taken from an earlier comparison study based on the NAEP 2005 mathematics framework (see next endnote).

[7] The data for PISA mathematics items and how they fit to the NAEP framework were taken from a comparison study conducted by NCES in 2003. The results remain relevant for our purposes here because the items used in

PISA 2006, when math was a minor domain, were a subset of the items used in 2003 and were selected to be representative of the prior administration. Additionally, there were no changes to the frameworks (that were not already accounted for in the study) between 2003 and the most recent administration of the two studies. See: Neidorf, T.S.; Binkley, M.; Gattis, K.; and Nohara, D. (2006). *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments* (NCES 2006-029). U.S. Department of Education. Washington, D.C.: National Center for Education Statistics.

[8] The PISA scientific literacy framework also puts an attitudinal dimension alongside the content/knowledge dimension and embeds attitudinal items (normally placed in a background questionnaire) within the actual assessment. Like students' knowledge, their attitudes (interest and motivation, sense of responsibility, and support for inquiry) are seen in the PISA model as influences on competencies. Although the responses to the attitudinal items are not part of the PISA score—they are reported separately—their presence in the assessment is a feature unique to PISA.

[9] For NAEP, our analyses rely on the developers' classifications of items to the framework.

[10] The present comparison studies were designed to be generally comparable with the 2003 comparison studies, and the same conventions were followed in the analysis phase. For example, in both studies, multi-part items were tabulated as a single unit with one classification contributing to the overall count of items. In cases where there were multiple classifications for individual parts of an item, the overall item was deemed to classify to multiple categories ("m"). For the content analyses, the few cases in which there were multiple classifications were identified in the tables as multiples and also counted with the relevant content areas. For the grade level and item format analyses, for the few cases in which there were multiple classifications, a single grade or item format was assigned based on rounding. For items classified at multiple grades, items were classified at either the higher grade or the grade for which the individual parts were most frequently classified (in the case of items with three or more parts). Items with both multiple choice and constructed response parts were counted as constructed response. Any differences in results between the waves of studies may reflect variability in the panels' judgment, with different sets of panel members over time.

[11] The NAEP mathematics framework specifies topics and objectives within the broad content areas. The NAEP science framework has one additional level of detail, specifying topics, subtopics, and objectives within the broad fields of science. Thus, this section presents results on the degree to which items could be classified not just at the finest level of detail, the objective level, as was done in the section on mathematics, but also at the subtopic level, which does not have a corollary in the mathematics framework.

[12] The TIMSS 2007 results are reported in: Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., and Brenwald, S. (2008). *Highlights From TIMSS 2007: Mathematics and Science Achievement of U.S. Fourth- and Eighth-Grade Students in an International Context* (NCES 2009–001). U.S. Department of Education. Washington, D.C.: National Center for Education Statistics. See endnote 3 for references for the NAEP and PISA results described.

[13] There were no changes for the fifth racial group (Native American/Alaska Native) from 2005 to 2007.

[14] There was one exception: there was no change in the performance of Asian eighth-grade students, although this was calculated over a different time period (1992 to 2007) than the other NAEP trends.

[15] NAEP's race/ethnicity categories include: White, Black, Hispanic, Asian/Pacific Islander, and American Indian/Alaska Native—the first four of which saw the increases in science performance between 1996 and 2005 referred to in this section. Race categories exclude Hispanic origin.