NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

U.S. Department of Education
Institute of Education Sciences
NCES 2006–029

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

Technical Report

May 2006

Teresa Smith Neidorf
NAEP Education Statistics Services Institute
American Institutes for Research

Marilyn Binkley
National Center for Education Statistics

Kim Gattis
NAEP Education Statistics Services Institute
American Institutes for Research

David Nohara
Independent Consultant

U.S. Department of Education
Margaret Spellings
Secretary

Institute of Education Sciences
Grover J. Whitehurst
Director

National Center for Education Statistics
Mark Schneider
Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006

May 2006

The NCES World Wide Web Home Page is http://nces.ed.gov.
The NCES World Wide Web Electronic Catalog is http://nces.ed.gov/pubsearch.

Suggested Citation
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, DC: National Center for Education Statistics. Retrieved [date] from http://nces.ed.gov/pubsearch.

For ordering information on this report, write to
U.S. Department of Education
ED Pubs
P.O. Box 1398
Jessup, MD 20794-1398

or call toll free 1-877-4ED-Pubs or order online at http://www.edpubs.org.

Content Contacts:

Marilyn Binkley
(202) 502-7484
Marilyn.Binkley@ed.gov

Elois Scott
(202) 502-7489
Elois.Scott@ed.gov

Executive Summary

The National Center for Education Statistics (NCES) collects information on student performance in key subject areas through the National Assessment of Educational Progress (NAEP), as well as through participation in international studies of student achievement. Information from these studies is used to inform policymakers, educators, researchers, and the public about the knowledge and skills of U.S. students and how these compare with students in other countries.

This technical report describes a study that was undertaken to compare the content of three mathematics assessments conducted in 2003: the NAEP fourth- and eighth-grade assessments; the Trends in International Mathematics and Science Study (TIMSS), which also assessed mathematics at the fourth- and eighth-grade levels; and the Program for International Student Assessment (PISA), which assessed the mathematical literacy of 15-year-old students. Its aim is to provide information useful for interpreting and comparing the results from the three assessments, based on an in-depth look at the content of the respective frameworks[1] and assessment items.

The report draws upon information provided by the developers of the assessments, as well as data obtained from an expert panel convened to compare the frameworks and items from the three assessments on various dimensions.[2] The frameworks were compared with respect to

- how each assessment organizes and defines the mathematics content and process skills to be assessed at each grade (or age) level;
- the main content areas included and the set of topics covered in each; and
- other aspects, such as item format and calculator policy.

Item comparisons were based on

- cross-classification of NAEP and TIMSS items to each other's assessment framework in terms of the mathematics content covered and grade-level expectations;
- classification of PISA items to the NAEP framework on these same dimensions;
- classification of all items with respect to their level of *mathematical complexity*;[3] and
- comparisons based on other framework dimensions related to cognitive processes, item formats, and item contexts.

---

[1] Assessment frameworks define what will be assessed, including the content to be covered, the types of test questions, and recommendations for how the test is administered.
[2] The panel members—experts in mathematics, mathematics education, and mathematics assessment, with familiarity and experience across the three assessments—are listed in appendix C.
[3] *Mathematical complexity* reflects the demands on thinking that an item makes, assuming that a student is familiar with the mathematics content of the task. The classifications in this report are based on the definitions in the NAEP 2005 framework for three levels of *mathematical complexity*—low, moderate, and high—that form an ordered description of the demands an item may make on a student (described in appendix B).

While comparisons between NAEP, TIMSS, and PISA were focused on the common classification systems based on the NAEP framework, the study also included a limited comparison between PISA items and NAEP eighth- and twelfth-grade *problem solving* items in light of the dimensions in the PISA framework. Example items are referenced throughout the report to illustrate some key similarities and differences.[4]

The results of this study indicate that although the NAEP, TIMSS, and PISA 2003 mathematics frameworks address many similar topics and require students to use a range of cognitive skills and processes, it cannot be assumed that they measure the same content in the same way. A hypothetical student who takes all three assessments might indeed perform equally well on them, but depending on the curriculum they have been exposed to and their skill and experience in various types of mathematical thinking, other students might exhibit quite different levels of performance across the three assessments. For NAEP and TIMSS, this is also true within each of the five corresponding content areas related to *number*, *measurement*, *geometry*, *data*, and *algebra*.

At the overall level, there is apparent agreement between the NAEP and TIMSS frameworks on the general boundaries and basic organization of mathematics content across the fourth and eighth grades, with nearly all items from each assessment being placed in one of the major content areas of the other assessment framework at the broadest level. Furthermore, both NAEP and TIMSS place similar emphases on each of the five major content areas, as evidenced by similar distributions of items across the main content areas of both frameworks at both the fourth- and eighth-grade levels. These types of comparisons, however, do not consider the grade level correspondence or the level of content match based on the distribution of items across the specific set of topics and subtopics included at each grade level in each of the assessments.

Despite the similarity between NAEP and TIMSS at the broadest content area level, there are differences between the two assessments when considering more detailed comparisons of the mathematics content covered and the grade level correspondence between the items in each assessment and the intentions of the other assessment framework. Differences between the NAEP and TIMSS assessments emerge with more detailed content analyses that consider the level of content match to specific topics and subtopics in the other assessment framework, with 20 percent of fourth-grade and about 15 percent of eighth-grade items from both assessments not classified to specific subtopics in the other assessment framework at any grade level. This finding indicates that both assessments contain items that might not be included in the other assessment and supports the general claim that NAEP and TIMSS do not necessarily assess the same mathematics content.

Most NAEP and TIMSS items were placed at the same grade on the other assessment framework, but this was not always within the corresponding content area. The overall grade-level correspondence between the NAEP items and the TIMSS framework, 86 percent at fourth grade and 73 percent at eighth grade, was lower than that between the TIMSS items and the NAEP framework (at least 90 percent). This is related at least in part to the inclusion of cross-grade items in NAEP that were administered at multiple grade levels. There are notable differences across content areas in the level of grade match between the two assessments. In the TIMSS assessment, *measurement* and *geometry* account for most of the items classified at different grade levels to the NAEP framework (10 percent or more). In the NAEP assessment, the content area of *data analysis, statistics and probability* has the largest percentage of fourth-grade items classified at a higher grade level (almost

[4] Additional released item sets from each assessment are also available on the NAEP, TIMSS, and PISA websites: http://nces.ed.gov/nationsreportcard, http://isc.bc.edu/timss2003.html, and http://www.pisa.oecd.org.

half), most notably from items covering probability topics which are not assessed in TIMSS until the eighth grade. In the NAEP eighth-grade assessment, between 10 and 43 percent of items in all content areas were classified at the fourth-grade level in TIMSS, with the largest proportion of items in *measurement* and *geometry and spatial sense* (37 and 43 percent, respectively) and the smallest in *data analysis, statistics, and probability* (10 percent).

Within each content area, detailed comparisons of content coverage and grade correspondence indicate that NAEP and TIMSS items are not necessarily measuring the same content. Some of the differences in topic and subtopic emphasis or grade level match in each content area include the following:

- *Number*: At the eighth grade, TIMSS has a relatively larger emphasis on ratio, proportion, and percent compared to whole numbers in NAEP. There is a somewhat greater emphasis on computation in TIMSS at both fourth and eighth grades. Only the NAEP eighth-grade assessment includes scientific notation (data not shown).

- *Measurement*: A larger proportion of NAEP fourth-grade items involve the selection and use of appropriate measurement instruments and units. While TIMSS has a greater emphasis at both grades on problems involving properties (area, perimeter, volume, surface area) of two- and three-dimensional shapes, a number of fourth-grade TIMSS items were classified to the NAEP eighth-grade framework (16 percent). In each assessment, at least 25 percent of eighth-grade items was classified at the lower grade level of the other assessment. In addition, there is an overlap of NAEP *measurement* items with topics in the TIMSS *geometry* framework.

- *Geometry*: A larger proportion of NAEP items involve two- and three-dimensional shapes, while TIMSS has a greater emphasis on congruence and similarity. There are differences in the nature of problem-solving items (TIMSS with more application of geometric properties and NAEP with more use of geometric models). Forty-three percent of NAEP eighth-grade items were classified to the TIMSS fourth-grade framework, while 13 percent of TIMSS eighth-grade items were classified to the NAEP twelfth-grade framework.

- *Data*: NAEP includes probability items in the fourth-grade assessment, while TIMSS does not include this topic until the eighth grade. In TIMSS, there is a greater emphasis on reading and interpreting data in tables and graphs at the fourth grade. In NAEP, there is a higher proportion of eighth-grade items involving the organization and display of data.

- *Algebra*: TIMSS has a greater emphasis on algebraic expressions and operations at eighth grade. Some of the eighth-grade NAEP items involving patterns, equations, and functions were classified to the fourth-grade TIMSS framework (18 percent). There is an overlap of NAEP *algebra and functions* topics involving the use of number lines and coordinate systems with the TIMSS *geometry* framework, and some TIMSS eighth-grade items were classified to the NAEP twelfth-grade framework (5 percent).

NAEP and TIMSS appear to be quite similar overall in terms of the distribution of items across the low, moderate, and high *mathematical complexity* levels. Sixty-four percent of fourth-grade items and more than half of eighth-grade items were classified at the low complexity level and less than 5 percent were classified at the high complexity level at both grade levels in NAEP and

TIMSS. The content areas with the highest proportion of items (more than 60 percent) classified at the moderate or high complexity level are *algebra and functions* in fourth-grade NAEP, *data analysis, statistics, and probability* in eighth-grade NAEP, and *measurement* in eighth-grade TIMSS.

PISA stands apart from NAEP and TIMSS in a number of important areas, including the organization of its mathematics content framework (which is based on *overarching ideas*), its focus on problem solving in real-world applications, and the fact that it samples students based on age (15-year-olds) rather than grade level. Interestingly, PISA items, which are distinct from NAEP and TIMSS items in numerous ways, do have a relatively high degree of content match to NAEP subtopics from a purely mathematics content perspective (more than 90 percent classified to a NAEP subtopic). Grade-level analyses based on classifications to the NAEP content framework also indicate that although the target population of PISA is somewhat older than the students taking the NAEP and TIMSS eighth-grade assessments, the mathematics content of most of the PISA items (85 percent) are at the eighth grade level.

The different nature of PISA makes it complementary to both NAEP and TIMSS. The mathematics topics addressed may not necessarily be substantially different, although PISA places greater emphasis on data analysis and less on algebra than do either NAEP or TIMSS, but it is in how that content is presented that makes PISA different. In terms of item type and level of *mathematical complexity*, PISA is quite different from NAEP and TIMSS. Not only does PISA use multiple-choice items to a far lesser degree, but it also contains a substantially higher proportion of items (71 percent) classified at the two upper levels of *mathematical complexity* (moderate and high).

Differences in the demands that the problem-solving items place on students' mathematical thinking skills are also found when comparing PISA items and NAEP eighth- and twelfth-grade *problem solving* items with respect to the PISA *competency clusters*.[5] From the perspective of the PISA framework, the mathematical thinking skills required of the NAEP *problem solving* items are focused more on *reproduction* and much less on *reflection* than PISA. This is consistent with their different purposes—NAEP being more closely aligned with curriculum-based mathematics outcomes at fourth, eighth and twelfth grades and PISA assessing the preparedness of 15-year-olds to be able to apply mathematics to solve novel, real-world problems. The *situations or contexts*[6] involved in the NAEP *problem solving* items also differed from PISA, with NAEP having a relatively higher proportion of items focused on *educational/occupational* and *scientific* contexts and lower proportions involving *personal* and *public* contexts than PISA. A number of the NAEP *problem solving* items investigated were judged by the panel as not appropriate for the PISA assessment (due to contexts or mathematical applications that were not authentic) or requiring revisions related to the level of instructions, general formatting, and sequencing in order to be included in the PISA assessment.

This report illustrates the complementary nature of the assessments, as there are certainly cases, especially looking within content areas, where results from NAEP, TIMSS, or PISA might be more informative than the others regarding a specific topic or skill. However, as scores are not reported at the topic or subtopic level, the ability to use assessment results to make statements about these student skills or abilities is limited to performance on individual items.

[5] The PISA framework defines three *competency clusters*—*reproduction*, *connections*, and *reflection*—to describe the mathematical cognitive processes required by its mathematics items. These are described in section 2.3.
[6] The PISA framework includes a situations or contexts dimension with four categories—*personal*, *educational/occupational*, *public*, and *scientific*.

For all three assessments, when reviewing results, it is important to look beyond the overall scores and content area subscales and examine in detail what each assessment measures. This study has yielded data that can be used to make informed readings of results. While there is no single factor that may be related to differences in student performance, the numerous differences noted here, whether dramatic or more minor, may have a substantial effect overall. As each assessment program continues, this type of research can continue, not only to help explain differences in student scores, but also to understand the complementary nature of the three assessments.

This report provides a first-level comparison of items in each assessment in terms of the coverage of broad content areas and distribution across mathematics topics as defined in the frameworks. All items in each assessment were considered in order to make overall comparisons of content coverage and grade-level expectations as well as distributions with respect to three broad levels of mathematical complexity. In addition, the types of item classifications conducted within the time constraints of this study permit comparisons at the mathematics topic level for each content area. While this method provides a broad view of some of the similarities and differences between the assessments, it is limited in terms of the types of comparisons that are provided at the item level. More in depth analyses of the exact nature of the items from each assessment within topics would reveal other important differences related to difficulty, scope, depth, complexity, and other item attributes. These types of more focused comparisons were outside the scope of this study, but may be important to include in future comparative studies of the assessments.

Acknowledgments

Many people's contributions made this report possible, and the authors wish to thank all those who have assisted with various aspects of the report, including data analysis, reviews, and design.

Members of the expert panel listed in appendix C provided all of the item classification data used as the foundation for the results presented in this report. The authors would like to thank the panel members for their expertise and contributions to the study.

Thanks to Patrick Gonzales, Eugene Owen, and Mariann Lemke of the National Center for Education Statistics (NCES) for their input on the design and their role in reviewing the report. The authors are grateful to NCES technical reviewers, Marilyn Seastrom and Shelley Burns, as well as to NCES program directors (Andrew Malizio and Elois Scott) and associate commissioners (Peggy Carr and Val Plisko) who provided direction and support for this publication. The authors wish to thank Lisa Bridges of the Institute of Education Sciences, Christine Brusselmans-Dehairs of Ghent University, Jeff Haberstroh of the Educational Testing Service, Marcie Dingle of Synergy Enterprises, and two anonymous reviewers for their very helpful comments.

Several people, formerly or currently from the Education Statistics Services Institute (ESSI), provided support in the form of research assistance and/or review: Ben Dalton, Kristy David, Aaron Douglas, J. Lane Glenn, Jamie Johnston, Juan Carlos Guzmán, Dana Kelly, Janine Emerson, David Miller, Melanie Ouellette, Lisette Partelow, Anindita Sen, and Margaret Woodworth. The authors wish to thank ESSI technical reviewers, Lauren Gilbertson and Zeyu Xu, for their hard work and expertise. Thanks also to David Miller in helping to shepherd the report through the final stages of the review and publication process, to Heather Block, also of ESSI, for her work on graphic design for this report, and to Brian Henigin of Westat for the final publication formatting of the report.

For permission to use secure items and publish released items, the authors would like to thank the Trends in International Mathematics and Science Study (TIMSS) and the International Association for the Evaluation of Educational Achievement (IEA), as well as the Program for International Student Assessment (PISA) and the Organization for Economic Cooperation and Development (OECD). Thanks also to the staff of the Educational Testing Service for participation in the expert panel meeting and for supplying item information and review materials from the National Assessment of Educational Progress (NAEP).

Contents   Page

List of Tables

Table **Page**

List of Figures

Figure **Page**

List of Exhibits

Exhibit **Page**

1. Introduction

Researchers, policymakers, educators, and members of the general public interested in the achievement of U.S. students currently have available several major sources of national-level data: results from the U.S. Department of Education's National Assessment of Educational Progress (NAEP) and U.S. results from various international assessments, such as the Progress in International Reading Literacy Study (PIRLS), the Program for International Student Assessment (PISA), and the Trends in International Mathematics and Science Study (TIMSS). NAEP administers periodic assessments in reading, mathematics, science, and other subjects at the fourth, eighth, and twelfth grades; TIMSS assesses mathematics and science at fourth and eighth grade; and PIRLS is a reading literacy assessment administered to fourth-grade students. In comparison, PISA primarily assesses the literacy[1] of 15-year-old students in reading, mathematics and science. In cases where the different assessments address the same subject areas (e.g., mathematics, reading, science) at the same or similar grade levels, the opportunity exists to measure U.S. student achievement using multiple instruments. Comparing results across assessments can be useful not only for interpreting the results, but also for developing a more complete picture of student achievement than would be possible with the results of just one assessment.

In order to provide useful guidance for comparing the results of the different assessments, the U.S. Department of Education's National Center for Education Statistics (NCES) has periodically conducted studies comparing various assessments in terms of their underlying frameworks, items, and other related features. In 2003, NCES conducted two comparison studies—one in mathematics and one in science—following the 2003 administrations of TIMSS and PISA. This report focuses on a comparison of the mathematics assessments—NAEP 2003, TIMSS 2003, and PISA 2003—while a companion report (Neidorf, Binkley, and Stephens 2006) compares the NAEP 2000 and TIMSS 2003 science assessments.

The 2003 mathematics and science comparison studies build on several earlier studies, which were also undertaken to explore the similarities and differences between NAEP and various international assessments. Such studies comparing frameworks and items are conducted periodically, as NAEP and international assessments evolve, improving their frameworks and test items to reflect current research, policy, and practice.

Previous published studies of mathematics and science assessments included comparisons of the TIMSS 1995 and NAEP 1996 mathematics assessments (McLaughlin, Dossey, and Stancavage 1997) and the NAEP 2000, TIMSS 1999, and PISA 2000 mathematics and science assessments (Nohara 2001). Both studies compared the underlying frameworks and test items from each assessment in terms of content, item format, and thinking skills required.

There also have been several studies comparing reading assessments. The earliest of these compared the NAEP 1992 reading assessment and the 1991 IEA Reading Literacy Study (Binkley and Rust 1994). More recently, Binkley and Kelly (2003) examined the frameworks, reading passages and items from the NAEP 2002 and PIRLS 2001 reading assessments.

---

[1] PISA uses the terminology of "literacy" in each subject area to denote its broad focus on application of knowledge and skills; that is, PISA seeks to ask if the 15-year-olds are mathematically literate, or to what extent they can apply mathematical knowledge and skills to a range of different situations they may encounter in their lives.

The goal of this mathematics comparison study is to identify similarities and differences between the 2003 NAEP, TIMSS, and PISA assessments based on a detailed comparison of their frameworks and items. This information may be used to help inform interpretations of student performance in mathematics on the three different assessments. While there are other important aspects that might be compared, such as item difficulty, sampling, and scaling procedures, this study focuses on a comparison of the content of the assessments. This content comparison is based on the main dimensions of the assessment frameworks and focuses on a comparison of the set of assessment items as a reflection of how the frameworks are implemented. The main questions driving the study are as follows:

- How do NAEP, TIMSS, and PISA define the domain of mathematics to be assessed and its main content areas, in terms of both the topics that are included and the distribution of items across topics?

- How do NAEP, TIMSS, and PISA define the content and process skills appropriate for the assessments at different grade or age levels? How do the items in each assessment compare to the grade-level expectations specified by the other frameworks?[2]

- How do the items in the NAEP, TIMSS, and PISA assessments compare with respect to the level of *mathematical complexity* demanded of students?

- How do NAEP, TIMSS, and PISA compare with respect to the types and distribution of item formats used? How do the items in the different assessments compare in terms of their problem-solving contexts?

To answer these questions, NCES convened an expert panel (appendix C) to examine the mathematics frameworks and items for each assessment. The panel cross-classified NAEP and TIMSS fourth- and eighth-grade items to each other's assessment frameworks with respect to mathematics content and grade level. PISA items also were classified to the NAEP framework on the same dimensions. The panel classified the items from all three assessments with respect to a common definition of *mathematical complexity* level based on the NAEP 2005 framework.[3] A limited comparison was also made between PISA items and NAEP eighth- and twelfth-grade *problem solving* items. Although TIMSS and PISA were not compared directly, this approach permits the comparison of NAEP, TIMSS, and PISA through the common classification systems based on the NAEP framework. In addition to the classification data from the panel, the study draws upon information provided by the NAEP, TIMSS, and PISA assessment developers that describes how each item is classified according to the main dimensions of its own framework, as well as other relevant characteristics such as item format and scoring rubrics.

[2] The 2003 mathematics and science comparison studies are the first to compare the assessments in terms of grade level—the extent to which items from one assessment map to the same grade level of the framework of the other assessment.

[3] The rationale for using this dimension from the NAEP 2005 framework as the basis for item classifications is described in appendix D.

Section 2 of this report presents an overview of the NAEP, TIMSS, and PISA assessments and a comparison of their respective mathematics assessment frameworks. Section 3 reviews the methods used for this comparison study. The results of the study are then presented in three major sections. The first, section 4, compares the NAEP, TIMSS, and PISA assessments overall with respect to content coverage, grade level, *mathematical complexity* level, and item format. The overall comparisons are followed by comparisons of the NAEP and TIMSS assessments with respect to each of the following main content areas (section 5): *number*, *measurement*, *geometry*, *data*, and *algebra*. This section provides more detailed comparisons of the extent to which items in one assessment map to the mathematics framework of the other assessment. It compares the content distribution of the items for each of the NAEP and TIMSS mathematics subscales. Section 6 contains additional comparisons made between the NAEP and PISA assessments, including detail on the mathematics topics covered by the PISA items and how NAEP eighth- and twelfth-grade *problem solving* items compare to those included in the PISA assessment. The report concludes with a summary of key findings (section 7).

2. Overview of the Assessments and Their Frameworks

NAEP

The National Assessment for Educational Progress (NAEP) is the United States' source for nationally representative and continuing information on what American students know and can do and is well known as the Nation's Report Card. NAEP policies and frameworks are established by an independent National Assessment Governing Board (NAGB), and the Department of Education's National Center for Education Statistics (NCES) administers the assessments. For over 30 years, NAEP has periodically collected and reported data on achievement in reading, mathematics, science and other subjects, for students in fourth, eighth, and twelfth grades. The comparisons in this report are based on the main NAEP assessments conducted in 2003 at the fourth- and eighth-grade levels and in 2000 at the twelfth-grade level.[1]

The frameworks established by NAGB for all the NAEP subject areas, including mathematics, are based on the collaborative input of a wide range of experts and involvement by participants from government, education, business, and public sectors. They are informed by common curricular practices in the nation's schools and ultimately are intended to reflect the best thinking about the knowledge, skills, and competencies needed for students to have a deep level of understanding at different grades and in different subject areas.

TIMSS

The Trends in International Mathematics and Science Study (TIMSS) is the United States' source for international comparative information on mathematics and science education in the elementary and middle grades. TIMSS is one of the current studies conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA), which has been conducting international comparative studies since the early 1960s, and is directed by the International Study Center at Boston College. TIMSS collects achievement and background data to provide information on trends in mathematics and science achievement over time as well as on the curricular, instructional, and attitudinal factors that may be related to performance. TIMSS collects data on a 4-year cycle, with the first administration in 1995 (at fourth, eighth, and twelfth grades),[2] the second in 1999 (at eighth grade only), and the most recent in 2003 (at fourth and eighth grades), with about 50 countries participating.

Like NAEP, the TIMSS assessments are based on collaboratively developed frameworks. In contrast to NAEP, however, the framework development and consensus process involved mathematics experts, education professionals, and measurement specialists from many countries.

---

[1] At the time this study was conducted, NAEP 2000 was the most recent mathematics assessment at grade 12, and NAEP 2003 (which did not include grade 12) was the most recent mathematics assessment at grades 4 and 8. NAEP long-term trend assessments in mathematics were also administered in 2003–04 but were not included in this study. Later, in 2005, NAEP conducted a mathematics assessment at fourth, eighth, and twelfth grades.

[2] Defined as the upper of the two grades containing the majority of 9-year-olds or 13-year-olds and the final year in secondary school. These are the fourth, eighth and twelfth grades in the U.S. and most other countries. TIMSS 1995 was also administered in third and seventh grades.

PISA

The Program for International Student Assessment (PISA) is conducted by the Organization for Economic Cooperation and Development (OECD). The main objective of PISA is to provide regular, policy-relevant data on the "yield" of education systems, and so targets students at an age that is near the end of compulsory schooling in most countries (15-year-olds). PISA focuses on literacy—the ability to use and apply knowledge and skills to real-world situations encountered in adult life—in the key subject areas of reading, mathematics, and science. PISA is, thus, the United States' source of comparative information on the reading, mathematical, and scientific literacy skills of students in the upper grades, and it provides benchmarks to international performance levels based on other OECD countries. The frameworks guiding the PISA assessments reflect a consensus across the OECD countries regarding the skills and abilities that demonstrate literacy in these key areas.

A key design feature of PISA is its cycle of rotating emphasis among the three key assessment areas every three years. Each subject area is assessed in each data collection, but the design distinguishes between major and minor domains. When a subject is the major domain it comprises a relatively greater share of the total assessment time, with a larger number of items and an assessment framework that is more fully developed and updated. Reading literacy was the major domain in the first PISA assessment in 2000 (32 countries), mathematical literacy was the major domain in the most recent 2003 assessment (41 countries), and scientific literacy will be the major domain in the next assessment in 2006.[3]

Organization of the NAEP, TIMSS, and PISA 2003 Mathematics Frameworks

Assessment frameworks define what will be assessed, including the content to be covered, the types of test questions, and recommendations for how the test is administered. Exhibits 1-A, 1-B, and 1-C compare schematically the organizing dimensions in the NAEP, TIMSS, and PISA 2003 mathematics frameworks. These organizing dimensions provide the basic framework for the development of the pool of items in each assessment, and the frameworks include target percentages for the distribution of the assessments across the main categories in each dimension to ensure a balanced assessment (discussed in the following sections).[4] As seen in these exhibits, there are some basic organizational differences between the frameworks, especially between PISA and NAEP or TIMSS.

Both the NAEP and TIMSS 2003 mathematics frameworks represented in exhibits 1-A and 1-B are based on two main organizing dimensions—a content dimension and a cognitive dimension—as well as an overarching dimension (along the bottom) that defines processes that go across the content and cognitive categories. Both NAEP and TIMSS include five similarly labeled categories in the content dimension (*content strands* in NAEP and *content domains* in TIMSS) that correspond to major mathematics curricular areas related to *number*, *measurement*, *geometry*, *data*, and *algebra*. In the main cognitive dimensions (*mathematical abilities* in NAEP and *cognitive domains* in TIMSS), NAEP has three broad categories (*conceptual understanding*, *procedural knowledge*, and *problem solving*), while TIMSS has four (*knowing facts and procedures*, *using concepts*, *solving routine problems*, and *reasoning*). There is overlap between the categories defined in the cognitive dimensions in NAEP and TIMSS as well as the processes defined by the overarching dimensions in each assessment (*mathematical power* in NAEP and *communicating mathematically* in

[3] The 2003 PISA assessment also included an additional component assessing cross-disciplinary problem solving. Items from this separate component were not included in this comparison study.
[4] The frameworks only provide target percentages of items or assessment time as guidelines for test development.

TIMSS). All items developed for NAEP and TIMSS are classified with respect to which categories in the two main dimensions they assess. The overarching dimensions are also considered as items are developed.

In contrast to NAEP and TIMSS, the PISA mathematical literacy assessment framework includes three main dimensions as shown in exhibit 1-C. Like NAEP and TIMSS, there is one dimension related to mathematics content (*overarching ideas*); the four *overarching ideas* in PISA, however, do not directly correspond to the main content categories in NAEP and TIMSS. Also like NAEP and TIMSS, PISA includes a cognitive dimension (*competency clusters*). In addition to these two dimensions, the PISA framework includes a third main dimension related to the *situations or contexts* in which the application of mathematics concepts is required. The *situations or contexts* dimension does not have an analogue in the NAEP and TIMSS frameworks. All items developed for PISA are classified with respect to the main categories in each of its three dimensions.

The following sections describe and compare in more detail the mathematics assessment frameworks for NAEP, TIMSS, and PISA. Additional assessment framework summary documents that were used for the comparison study are found in appendixes A and B.

Exhibit 1-A. NAEP mathematics framework dimensions: 2003

Content strands	Mathematical abilities
Number sense, properties, and operations Measurement Geometry and spatial sense Data analysis, statistics, and probability Algebra and functions	Procedural knowledge Conceptual understanding Problem solving
Mathematical power (Reasoning, connections, communication)	

NOTE: The NAEP framework is based on two main organizing dimensions—*content strands* and *mathematical abilities*—as well as an overarching dimension (*mathematical power*) that defines processes that go across the content and abilities categories.
SOURCE: U.S. Department of Education, National Assessment Governing Board, *Mathematics Framework for the 2003 National Assessment of Educational Progress*, 2002.

Exhibit 1-B. TIMSS mathematics framework dimensions: 2003

Content domains	Cognitive domains
Number Measurement Geometry Data Algebra	Knowing facts and procedures Using concepts Solving routine problems Reasoning
Communicating mathematically	

NOTE: The TIMSS framework is based on two main organizing dimensions—*content domains* and *cognitive domains*—as well as an overarching dimension (*communicating mathematically*) that goes across the content and cognitive categories.
SOURCE: International Study Center, Lynch School of Education, Boston College, *TIMSS Assessment Frameworks and Specifications 2003, 2nd ed.*, 2003.

Exhibit 1-C. PISA mathematical literacy framework dimensions: 2003

Overarching ideas	Competency clusters	Situations or contexts
Change and relationship Quantity Space and shape Uncertainty	Reproduction Connections Reflection	Personal Educational/occupational Public Scientific

NOTE: The PISA framework is based on three main organizing dimensions—*overarching ideas* (content), *competency clusters*, and *situations or contexts*.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, 2003.

2.1. NAEP 2003 Mathematics Framework

The framework for the NAEP 2003 mathematics assessment is based on two major organizing dimensions—*content strands* and *mathematical abilities*—as well as an overarching dimension of *mathematical power* (exhibit 1-A).[5] The framework stipulates that every item in the assessment is given a primary classification in the two major dimensions according to certain distribution targets. The NAEP 2003 framework allows secondary classification of items to more than one content category. However, NAEP does not use multidimensional scaling, and these secondary classifications are not used in the analysis of results.

The first major dimension is defined by five broad *content strands*, which are the same for fourth, eighth, and twelfth grades. They are *number sense, properties, and operations*; *measurement*; *geometry and spatial sense*; *data analysis*, *statistics, and probability*; and *algebra and functions*. Each *content strand* is further defined by topics and subtopics.[6] The framework document indicates which topics and subtopics are intended for each grade level (4, 8, and 12). While many are intended for all three grades (4, 8, and 12), there are some topics and subtopics that should be assessed only at a specified grade level(s). In particular, a number of topics and subtopics are not intended to be assessed either at grade 4 or at grade 8. Others may only be introduced at a simple level at a lower grade(s) (such as by using a manipulative or pictorial model).

The NAEP 2003 framework specifies target percentages for the distribution of items across the *content strands.* As shown in table 1, the framework specifies that a large proportion of the fourth-grade assessment be devoted to *number sense, properties, and operations* (at least 40 percent) and that moderate emphasis be placed on *measurement* (20 percent)*, geometry and spatial sense* (15 percent) and *algebra and functions* (15 percent). The least emphasis at fourth grade is placed on *data analysis, statistics, and probability* (10 percent). At the higher grades there is a more even distribution across the *content strands*. The greatest emphasis is placed on *algebra and functions* for both eighth and twelfth grades (25 percent). At eighth grade, 25 percent of the assessment is also focused on *number sense, properties, and operations*.

---

[5] The mathematics framework for the NAEP 2003 assessment was developed in the early 1990s and was used as the basis for the assessments in 1996, 2000, and 2003 (NAGB 2002). The framework was updated for the 2005 assessment (NAGB 2004).

[6] See section 5 and appendix A for more information about the topics and subtopics included in the NAEP content framework.

Table 1. Target percentage of NAEP items distributed across NAEP framework dimensions, by grade: 2003

NAEP framework dimensions	Grade 4	Grade 8	Grade 12
Content strand			
Number sense, properties, and operations	40	25	20
Measurement	20	15	15
Geometry and spatial sense	15	20	20
Data analysis, statistics, and probability	10	15	20
Algebra and functions	15	25	25
Mathematical abilities			
Conceptual understanding	33	33	33
Procedural knowledge	33	33	33
Problem solving	33	33	33

NOTE: Percentages reflect the minimum target percentages specified in the NAEP 2003 mathematics framework. At all grades, distributions across *mathematical abilities* are approximately equal. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Assessment Governing Board, *Mathematics Framework for the 2003 National Assessment of Educational Progress*, 2002.

The framework also specifies cognitive abilities to ensure a balanced assessment that demonstrates mathematical thinking in various situations, which are described by a combination of *mathematical abilities* and *mathematical power*.[7] The cognitive dimension of *mathematical abilities* addresses the aspects of knowing and doing mathematics and is defined by three broad categories. These three categories are *conceptual understanding*, *procedural knowledge*, and *problem solving* and the framework specifies that there should be approximately equal emphasis on each of the *mathematical abilities*.

Mathematical power is defined as consisting of three overall cognitive processes in mathematics. These processes are *reasoning*, *connections*, and *communication*. *Mathematical power* as conceived in the framework reflects cognitive processes within a broader context of reasoning and with connections across the scope of mathematical content and thinking. Communication is a unifying thread, and the framework emphasizes the inclusion of extended constructed-response items as a way for students to provide meaningful responses to mathematics tasks and demonstrate their ability to communicate mathematically. While *mathematical power* is to be used as a foundation when developing items in each of the content and cognitive dimensions, no target percentages are specified for the individual process categories in this dimension of the framework.

Approximately one-third of the NAEP 2003 mathematics assessment comprises items for which students are permitted to use calculators.[8] The NAEP framework specifies that calculators be provided for the assessment. These are four-function calculators at fourth grade and scientific calculators at eighth and twelfth grades. The NAEP framework also specifies that manipulatives (e.g., rulers, protractors, and geometric shapes) be used for some portion of tasks in the mathematics assessment.

[7] The *mathematical abilities* and *mathematical power* dimensions are replaced in the NAEP 2005 framework with a single system of three levels of *mathematical complexity.* As discussed in forthcoming sections, the 2005 *levels of mathematical complexity* were used in this study as a means of comparing the cognitive demands of items across assessments.
[8] Calculators are only available to students while they are working on item blocks designed for use with calculators.

The NAEP framework specifies that multiple-choice, short-answer, and extended-response items be included in the assessment. The framework used for the 1996, 2000, and 2003 assessments does not specify exact proportions to be devoted to each of these item types, but does indicate an increasing emphasis on extended-response items and more balance between short-answer and multiple-choice items than in previous frameworks.

2.2. TIMSS 2003 Mathematics Framework

The TIMSS 2003 framework is based on two main organizing dimensions, *content domains* and *cognitive domains*, as well as an overarching dimension of *communicating mathematically* (exhibit 1-B).[9] All TIMSS items are classified with respect to *content domain* and *cognitive domain*.[10] There are five broad *content domains* assessed at both fourth and eighth grades (Mullis et al. 2003). These include *number*, *measurement*, *geometry*, *data*, and *algebra*. Within the *content domains*, the TIMSS framework further specifies main topic areas and grade-specific objectives within those topics that are appropriate for assessment at each grade.[11]

The TIMSS 2003 framework specifies target percentages for the distribution of assessment time across the content domains (table 2). The framework emphasizes the *number content domain* at fourth grade (40 percent of assessment time), with *measurement* as the domain with the next highest level of emphasis (20 percent). Since algebra is not taught as a formal subject in primary school across TIMSS countries, the *algebra* content area (patterns, equations and relationships) represents a relatively low proportion of the fourth grade assessment (15 percent). *Geometry* reflects the same proportion (15 percent). The *content domain* with the least emphasis at the fourth grade is *data* (10 percent). At eighth grade the largest percentage of assessment time is still *number* (30 percent), but there is an increased coverage of *algebra* (25 percent) and an equal distribution of assessment time across the other *content domains* of *measurement*, *geometry* and *data* (15 percent to each).

[9] The TIMSS mathematics framework was revised for 2003 from the original curriculum framework used as the basis for the 1995 and 1999 assessments. See Mullis et al. (2003), for additional information.
[10] While items developed for TIMSS may address more than one category, only primary classifications were used during test development to ensure framework coverage.
[11] See section 5 and appendix A for more information about the topics and objectives included in the TIMSS framework.

Table 2. Target percentage of TIMSS assessment time distributed across TIMSS framework dimensions, by grade: 2003

TIMSS framework dimensions	Grade 4	Grade 8
Content domain		
Number	40	30
Measurement	20	15
Geometry	15	15
Data	10	15
Algebra	15	25
Cognitive domains		
Knowing facts and procedures	20	15
Using concepts	20	20
Solving routine problems	40	40
Reasoning	20	25

NOTE: Percentages reflect the targets specified in the TIMSS 2003 framework.
SOURCE: International Study Center, Lynch School of Education, Boston College, *TIMSS Assessment Frameworks and Specifications 2003, 2nd ed.,* 2003.

On the cognitive dimension, TIMSS specifies four broad *cognitive domains* to describe the range of skills and abilities that students apply in responding to items in the assessment. These include *knowing facts and procedures, using concepts, solving routine problems,* and *reasoning.* Whereas the definitions of the *cognitive domains* are the same for both fourth and eighth grades, the distribution of assessment time specified in the framework differs somewhat across grades. Both fourth- and eighth-grade assessments include the same emphasis on *solving routine problems* (40 percent) and *using concepts* (20 percent) (table 2). The two grades differ, however, with respect to the distribution across the *knowing facts and procedures* and *reasoning* categories. At the fourth grade, equal emphasis is placed on each of these (20 percent), while at eighth grade a greater emphasis is placed on *reasoning* (25 percent) than *knowing facts and procedures* (15 percent).

The TIMSS framework also specifies *communicating mathematically* as an overarching dimension that is to be demonstrated through description and explanation. Adequate levels of constructed-response items at both grades are included to measure students' ability to communicate mathematically across a wide range of content and processes. Some portion of the items in each of the content and cognitive categories measure abilities related to the overarching dimension of *communicating mathematically*, but a target percentage is not specified in the framework.

TIMSS permitted the use of calculators at the eighth grade for the first time in the 2003 assessment. Students in the eighth grade were permitted to use calculators on about half of the assessment at the discretion of each participating country.[12] Because calculators were not permitted in the previous TIMSS assessments, calculators were permitted only for the new items in 2003 and not the trend items carried over from the 1995 and 1999 assessments. Calculators were not permitted at fourth grade in any TIMSS assessment. The TIMSS assessment also includes some extended problem-solving and inquiry tasks that involve the use of manipulatives such as rulers or geometric shapes.

[12] Calculators can be used only during the second half of the testing session. TIMSS does not provide calculators, but students may use their own calculators or those provided by their schools. In the United States, simple-function calculators were provided to eighth-grade students for the TIMSS 2003 assessment.

The TIMSS 2003 framework also specifies that both multiple choice and constructed-response items (requiring students to provide a written response) be included in the assessment, with up to two-thirds of the assessment time coming from multiple-choice items. About two-thirds of the constructed-response items require a short answer, while the other third require a more extended response.

2.3. PISA 2003 Mathematical Literacy Framework

The PISA 2003 mathematical literacy framework includes three major dimensions related to mathematical content (*overarching ideas*), mathematical cognitive processes (*competency clusters*) and *situations or contexts* (exhibit 1-C).[13] All items in PISA are classified with respect to each of these three main dimensions, and the framework includes specifications for the distribution of the assessment across the categories in each (table 3).

The mathematics content to be assessed in PISA is defined by four *overarching ideas*, which include *quantity*, *space and shape*, *change and relationships*, and *uncertainty*. Collectively, they are intended to cover most of the mathematics domains that students are typically exposed to through their mathematics school curriculum, but these particular conceptual areas also were chosen because they encompass a set of phenomena and concepts within a broad range of situations that students are likely to encounter outside of school. *Quantity* includes the topics of number sense, meaning of operations, mental arithmetic, and estimation. *Space and shape* covers recognizing shapes and patterns, understanding dynamic changes to shapes, similarities and differences, and 2- and 3-dimensional representations and relationships between them. *Change and relationships* refers to functional thinking and covers different types of growth (i.e., linear, exponential, periodic, logistic) and the relationships between them. *Uncertainty* includes such topics as data collection, analysis, and representation; probability; and inference.

In the process dimension, PISA defines mathematical competencies important for mathematical literacy and describes the cognitive activities that these competencies encompass according to three *competency clusters* including *reproduction, connections,* and *reflection*. For each *competency cluster*, the framework addresses the abilities and skills associated with eight competencies: thinking and reasoning; argumentation; communication; modeling; problem posing and solving; representation; using symbolic, formal, and technical language and operations; and use of aids and tools.

The competencies demanded by items in the *reproduction* cluster involve "reproduction of practiced knowledge" and may require students to perform routine operations. The *connections* cluster involves the demonstration of competencies in problem-solving situations that are not routine but still familiar. Items in this cluster usually require evidence of integration of different mathematical concepts or making connections across *overarching ideas*. The competencies in the *reflection* cluster require students to plan solution strategies and implement them. Items in this cluster contain more elements and involve settings that are more "original" (less familiar) than in the other two categories. The framework specifies that approximately half of the assessment should be devoted to problems measuring competencies in the *connections* cluster and the remaining portion should be divided equally between the *reproduction* and *reflection* clusters.

---

[13] The same basic framework was used as the basis for the assessment of mathematical literacy in 2000. However, the framework was updated and expanded in 2003 when mathematical literacy became the major domain (OECD 2003).

Table 3. Target percentage of PISA assessment distributed across PISA framework dimensions: 2003

PISA framework dimensions	
Overarching ideas	
Change and relationships	25
Quantity	25
Space and shape	25
Uncertainty	25
Competency clusters	
Reproduction	25
Connections	50
Reflection	25
Situations or contexts	
Personal	25
Educational/occupational	25
Public	25
Scientific	25

NOTE: Percentages reflect the targets specified in the PISA 2003 framework. The PISA framework specifies that assessment time be distributed as evenly as possible across the four *overarching ideas* and across the four *situations or contexts*. The target percentages across the three *competency clusters* reflect the proportion of items in each cluster specified by the PISA framework.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, 2003.

The last dimension, *situations or contexts*, is an important aspect of the PISA framework. PISA places a heavy emphasis on authentic contexts for the use of mathematics and tasks that might be encountered in real-world situations. Four types of *situations or contexts* are defined and used in developing the problems in PISA. These include *personal*, *educational/occupational*, *public*, and *scientific*.

These *situations or contexts* types are based on a model of "distance" from the students' individual world. The *personal* category reflects situations that are within the immediate realm of students' personal experience and interest. The *educational/occupational* category includes problems encountered in students' school or work life involving the application of mathematics. Items in the *public* category involve problem-solving situations that students might encounter in the local community or as a functioning member of society at large. The *scientific* category involves more hypothetical scenarios or scientific applications of mathematics. The PISA framework specifies that the assessment should be balanced with respect to the proportion of problems from each of these types of *situations or contexts*.

The PISA assessment is organized into a set of tasks designed to be authentic problem-solving *situations or contexts* that involve the application of mathematics concepts from the *overarching ideas* and embody the mathematical processes in the *competency clusters*. In general, the tasks include some stimulus information, an introduction, and a series of related items, although there also are some individual questions. PISA places the decision to use calculators at the discretion of participating countries, with the intention of mirroring common instructional practices.[14] The framework specifies a range of format types, including multiple-choice, closed constructed-response, and open constructed-response, and that about equal numbers of each of these item types be included in the mathematical literacy assessment.[15]

2.4. Comparing the NAEP, TIMSS, and PISA Mathematics Frameworks and Assessments

There are a number of similarities between the frameworks and general design aspects of all three assessments. All three frameworks encompass a broad range of mathematics content knowledge and skills and include classification systems for describing the cognitive skills students use to respond to items. There are nevertheless differences, especially between the framework of PISA and the frameworks of NAEP and TIMSS. Whereas the NAEP and TIMSS frameworks are tied closely to the organizational structures used in traditional school curricula, the PISA framework is based on a situation- and phenomena-based approach that results in noticeably different categories for describing content. Both NAEP and TIMSS are grade-based assessments, with NAEP and TIMSS assessing fourth- and eighth- grade students and NAEP also assessing twelfth grade students.[16] In comparison, PISA defines an age-based population of students in secondary school (15-year-olds). Although the target populations for the eighth-grade NAEP and TIMSS assessments and the twelfth-grade NAEP assessment are reasonably close in age to PISA's target population of 15-year-olds, there is still a difference of roughly two grade levels. Furthermore, even though mathematical literacy was the major domain in the 2003 administration of PISA, it nevertheless included a much smaller number of items than either NAEP or TIMSS, meaning that assessment items may not as fully reflect the breadth and depth of the framework.[17] PISA is also unique in its extensive use of sets of two or more items based on a common stimulus. Although both NAEP and TIMSS do include such item sets, they are not as numerous as in PISA.

The NAEP and TIMSS frameworks are quite similar with respect to the broad structure of their content dimensions, but there are differences at the finer levels. Both are organized into five major areas of mathematics related to *number*, *measurement*, *geometry*, *data*, and *algebra*, although the terminology used for these content areas are not exactly the same and the content dimensions are defined somewhat differently based on the topics included in each. In particular, NAEP and TIMSS differ in terms of how they specify what is to be assessed at each grade level. The NAEP framework includes a list of topics and subtopics within each *content strand* and indicates which are intended to be included at each grade level (grades 4, 8, and 12). Although this indicates grade-level appropriateness of each topic and subtopic, the topics and subtopics themselves are the same for all grades in which they are included. As a result, there are few grade-specific subtopics specified in the NAEP framework. A separate NAEP assessment specifications document provides illustrative examples of appropriate items that might be developed for each topic at each grade level for use by

---

[14] In PISA in the United States, the decision to use calculators was left to schools based on school, district, or state policy.

[15] Although only three basic item formats are described in the PISA framework (multiple-choice, closed constructed-response, and open constructed-response), two additional types (complex multiple-choice and short response) were included in the item classification information provided by the assessment developer. These item types are discussed further in section 4.4.

[16] Although TIMSS was administered at the 12th grade in the 1995 survey, it was not administered at this grade level in the 2003 survey, which is the focus of this comparison study.

[17] As shown in section 4, there are 85 items in the PISA 2003 mathematics literacy assessment compared to at least 180 items in the eighth grade assessments in NAEP 2003 and TIMSS 2003 mathematics assessments.

the assessment developers (NAGB 1992). In contrast, the TIMSS framework includes a list of main topics within each *content domain* and grade-specific assessment objectives for each of the main topics. In most cases, the set of specific objectives in TIMSS is unique for each grade level (grades 4 and 8).

When making direct comparisons related to item content, this report uses a general terminology of content area, topic, and subtopic to refer to the comparable levels of specification used in the NAEP and TIMSS content framework (exhibit 2). For the discussion of content, cognitive or other classifications based on a single framework (NAEP, TIMSS or PISA), the terminology from that framework is used.

Exhibit 2. Terminology used in making comparisons across the NAEP 2003 and TIMSS 2003 content frameworks

NAEP framework		General terminology		TIMSS framework
Content strand	⇐	Content area	⇒	Content domain
Topic	⇐	Topic	⇒	Topic area
Subtopic	⇐	Subtopic	⇒	Objective

Examples of related NAEP and TIMSS content areas, topics, and subtopics

NAEP		TIMSS	
Content strand:	Data analysis, statistics, and probability	Content domain:	Data
Topic:	Read, interpret, and make predictions using tables and graphs	Topic area:	Data interpretation
Subtopic:	Read and interpret data	Objective:	Draw conclusions from data displays

SOURCE: U.S. Department of Education, National Assessment Governing Board, *Mathematics Framework for the 2003 National Assessment of Educational Progress*, 2002; and International Study Center, Lynch School of Education, Boston College, *TIMSS Assessment Frameworks and Specifications 2003, 2nd ed.,* 2003.

Since the purpose of PISA is to measure literacy rather than specific educational outcomes that are closely linked to curriculum-based content areas, its frameworks are structured rather differently than those of NAEP or TIMSS. First, PISA defines the mathematics content in terms of broad *overarching ideas* that go across the content areas defined in NAEP and TIMSS. Also, PISA includes a dimension not found in NAEP or TIMSS related to the *situation or context* of the item and emphasizes the use of “authentic” tasks and the application of mathematics to solve real-world problems. While NAEP and TIMSS do not have the same focus as PISA, there is still considerable language in the NAEP and TIMSS assessment frameworks about the application of mathematical knowledge and skills to problem-solving situations.

All three assessments structure their cognitive dimensions differently, but there is considerable overlap in the specific process skills, abilities, and competencies that are deemed important to be included in each assessment to demonstrate performance in mathematics. In particular, reasoning and communication are explicitly emphasized in all three assessment frameworks in the main cognitive or overarching dimensions. Making mathematical connections is also emphasized as one of the categories within the NAEP *mathematical power* dimension and PISA *competency clusters*. While mathematical connections is not an explicit category in TIMSS, it is included in the abilities to be demonstrated by items assessing the *reasoning* cognitive domain.

NAEP, TIMSS, and PISA all include both multiple-choice items, in which students choose the correct answer from a list of four or five choices, and constructed-response items, in which students generate their own answers. All frameworks allow for both short-answer and more extended constructed-response items, but the exact definition of these may vary across assessments. Some differences in item formats are discussed in the overall results section (section 4.4). The TIMSS framework specifies that one-third or more of assessment time be devoted to constructed-response items, and that about one-third of these items require an extended response. The PISA framework specifies an approximately equal distribution of items across the main format types of multiple-choice, closed constructed-response, and open constructed-response. The NAEP framework does not provide specific targets for proportions of items but emphasizes the importance of extended constructed-response items and a balance between multiple-choice and short-answer items.

All three assessment frameworks outline policies for calculator use. The NAEP 2003 framework permits calculators to be used by students in fourth, eighth, and twelfth grades on some portion of mathematics items (about one-third). In TIMSS, calculators were not allowed during the fourth-grade assessment. However, beginning with 2003, calculators were permitted but not required for newly developed eighth-grade assessment materials. In both TIMSS and PISA, participating countries could decide whether or not students were allowed to use calculators. In NAEP, calculators are provided—four-function calculators in fourth grade and scientific calculators in eighth and twelfth grades.[18] In TIMSS, the United States allowed students to use simple-function calculators that were provided with the test. In PISA, in the United States, the decision to allow calculators was left to schools based on school, district, or state policy.

Both the NAEP and TIMSS frameworks include the use of manipulatives (e.g., rulers, cardboard geometric shapes) in some tasks, and the assessments include small numbers of items involving them. In PISA, items using manipulatives are neither specified in the framework nor reflected in the assessment.

The assessment designs for NAEP, TIMSS, and PISA result in each individual student taking only a portion of the total assessment items, but the testing time for individual students differs across the three assessments. NAEP requires 50 minutes at all three grades, TIMSS requires 72 minutes at fourth grade and 90 minutes at eighth grade, and PISA requires two hours of testing time.

Finally, the NAEP framework was developed within the specific context of the U.S. system and defines a set of achievement levels (basic, proficient, and advanced) that are intended to provide descriptions of what students should know and be able to do in mathematics at each grade level from a national perspective. In contrast, the TIMSS and PISA frameworks reflect a consensus across diverse participating countries about the mathematics content and processes that should be assessed. For TIMSS, the framework reflects a consensus about what mathematics topics are most appropriate and important to assess at fourth and eighth grade; in general, the topics included are in the curricula for a majority of TIMSS countries. The PISA framework reflects a consensus across the OECD

[18] This NAEP policy is being revised for the 2005 assessment in which twelfth grade students will be permitted to use their own calculators.

countries about what knowledge, skills, and abilities reflect mathematical literacy and preparedness for adult life. Some of the differences in mathematics curricula and emphases across countries are reflected in differences between the frameworks and the items included in the assessments. The results presented in the following sections that compare the assessments overall and in each of the mathematics content areas provide some information on these possible differences.

This section provided an overview of the NAEP, TIMSS, and PISA assessments and a comparison of their respective mathematics assessment frameworks. The next section reviews the methods used for this comparison study.

3. Process and Methods

To conduct comparisons of the NAEP, TIMSS, and PISA 2003 assessments, NCES convened a panel of 11 experts in mathematics, mathematics education, and mathematics assessment. All panel members had familiarity and experience with at least one of the three assessments and their frameworks.[19] The panel met over a 3-day period to review the frameworks and classify the items from each assessment. The following three sections describe the organization of the expert panel meeting and the methods used for making the NAEP/TIMSS and NAEP/PISA comparisons reported in this report. Additional methodological notes are included in appendix D.

3.1. Organization of the Expert Panel Meeting

The expert panel meeting opened with a plenary session during which the study organizers presented the goals of the study, provided an overview of the NAEP, TIMSS, and PISA frameworks and assessments, and described the procedures for reviewing items. The expert panel members also had an opportunity during the opening plenary session to review, classify, and discuss several practice items in order to establish a common understanding of the classification procedures.

The first two days of the expert panel meeting were devoted to NAEP/TIMSS comparisons. All of the NAEP and TIMSS fourth and eighth grade mathematics items were reviewed, reflecting a total of about 650 items across the two assessments and grades. The items were divided into three groups according to content, with each group containing items from both NAEP and TIMSS in the content areas of[20]

- *number*;
- *measurement* and *geometry*; and
- *data* and *algebra*.

The panel also was divided into three groups, with each group responsible for reviewing and classifying all of the items in one of the content groups. Panelists and staff were assigned to subgroups to make sure that each group contained participants with expertise in each of the assessments. This division of items and panelists ensured a balance across the groups with respect to the coverage of assessments and grades as well as the number of items to be reviewed.

NAEP/PISA comparisons were conducted on the third day of the meeting. These comparisons involved a subset of the full panel, including participants with expertise in both NAEP and PISA and representatives from each of the original content area groups. After an initial orientation and plenary discussion of the PISA framework, the panel was divided into two groups to review and classify items. One group focused on reviewing the 85 PISA mathematical literacy items and the other group reviewed a set of 79 NAEP problem-solving items from the eighth- and twelfth-grade mathematics assessments.[21]

---

[19] A list of panel members and associated staff is presented in appendix C.
[20] The division of items was based on the assessment developers' classifications by content area subscale.
[21] The NAEP items selected for comparison with PISA were the items from the 2003 eighth-grade assessment and the 2000 twelfth-grade assessment that were extended constructed-response and/or were classified in the *problem solving* category of *mathematical abilities*.

Both components of the study concluded with a plenary session during which panelists shared their thoughts on the frameworks, items, and the study overall. While this report draws from these comments, where applicable, it describes primarily the results from the item review and classification sessions, which were the focus of the meeting.

3.2. Methods Used for NAEP/TIMSS Comparisons

In each content area group, the panel conducted a framework-level review to familiarize the panelists with these portions of the content frameworks and to uncover some of the main similarities and differences in how the major content areas covered by each group are interpreted in the two frameworks documents. The panels then classified the items, first classifying the TIMSS items to the NAEP framework and then classifying the NAEP items to the TIMSS framework. All items were classified on the following dimensions:[22]

- **Content**: Each item was classified with respect to the content framework of the other assessment (i.e., TIMSS items to the NAEP framework and NAEP items to the TIMSS framework) by identifying the content area, topic, and subtopic with the best match to the item content. Some items were classified as matching the other assessment framework at only the topic or content area level. Items that could not be classified at any level were also identified.

- **Grade level**: Each item was classified with respect to the grade level corresponding to the best content match in the other framework. For TIMSS items classified to the NAEP framework, grade classification was made to grade 4, 8, or 12. However, for NAEP items classified to the TIMSS framework, grade classifications were limited to grades 4 and 8 since TIMSS does not include grade 12.[23]

- ***Mathematical complexity* level** (low, moderate, or high): All items were classified with respect to *mathematical complexity* level as defined in the NAEP 2005 framework. Items in the *low complexity* category rely heavily on recall and recognition of previously learned concepts and principles. They may require students to carry out a mechanical or stated procedure or recognize an example of a concept. Items in the *moderate complexity* category involve more flexibility of thinking and choice among alternatives and often require more than a single step. These items include those that require students to use informal methods of reasoning and problem-solving strategies such as comparing figures or statements. In the *high complexity* category, items make heavy demands on students to engage in more abstract reasoning, planning, analysis, judgment, and creative thought. Items may require students to generalize a pattern or to describe, compare, and contrast solution methods.[24]

In conducting their evaluations, panelists were given several guidelines, including the following:

---

[22] Additional information about the content categories and definitions of levels of mathematical complexity is provided in appendixes A and B.

[23] The grade-level classifications are based on the content descriptions in the NAEP and TIMSS frameworks and in the NAEP specifications document (provided to panelists for reference). The classifications reflect the judgment of this particular expert panel and their knowledge of the NAEP assessment.

[24] A more complete description of the levels of *mathematical complexity* is included in appendix B. The considerations in selecting this dimension from the NAEP 2005 framework are discussed in appendix D.

- Items should be classified to the most detailed content level possible—ideally, to the subtopic level. (Although panelists were allowed to make some logical inferences about what a content area, topic, or subtopic might include, they were instructed not to classify items further than they believed was appropriate.)

- Each group should consider all content areas of the framework. The content area in one assessment may overlap with another content area in the other assessment (e.g., the best topic match for a *geometry* item may be in the *measurement* content area of the other framework).

- In cases where items appear to address multiple content areas, topics, or subtopics, a primary classification for the item should be identified whenever possible. (In cases where this was not appropriate, panelists indicated multiple or secondary classifications which were recorded. The results in this report are based on primary classifications in nearly all cases.)

- Instances where a number of items that cannot be placed in a framework are of a similar type should be indicated. These instances may indicate a potential gap in the framework to which the item is being classified.

- Grade-level classifications should be based on descriptions found in the frameworks, rather than on common understandings of grade-level content (i.e., items should be placed at the grade-level where they best match the descriptions in the content framework). (As with other content classifications, panelists were allowed to make some logical inferences about what a topic or subtopic might include at a given grade level).[25]

- Classifications to *mathematical complexity* level should be based strictly on the descriptions found in the NAEP 2005 framework and not on more general conceptions of complexity.

Within each group, panelists classified all items individually and then discussed the classifications as a group to arrive at a group classification. In general, consensus was reached, but for some items the final classifications reflect the classifications of the majority of panelists. To monitor consistency in the classifications of *mathematical complexity* level across the three groups, a set of common items was classified by the members of all three groups. The degree to which the three groups classified these items in the same categories on this dimension serves as a measure of the reliability of these classifications. The items in the reliability set were not chosen at random, but rather, were a representative set of 60 items (30 from NAEP and 30 from TIMSS) selected to cover the main categories addressed in the study (content area and grade level). Reliability items were classified at regular intervals throughout the classification process. The reliability procedure and results are described in more detail in the methodological notes (appendix D).

Expert panelists typically spent more time reviewing and classifying the items in the reliability set that were in their primary content area. Thus the classifications by the primary content

[25] Since the TIMSS framework contains grade-specific objectives, the grade-level classification is concurrent with a classification to an objective. For items classified to a topic or content domain, but not a grade-specific objective, the grade classification reflects the judgment of the panel of the grade at which the item is most consistent with the overall framework. Since the NAEP framework provides a set of topics and subtopics that usually apply to more than one grade, the grade classification reflects a judgment of grade-level correspondence in terms of the panelists' knowledge of the NAEP assessment. To assist in this process, the panel also consulted the grade-specific item examples in the NAEP assessment specifications document (NAGB 1992).

area expert panel groups are the most valid and used for all of the results in the report. Results from the secondary classifications of the reliability set were used to monitor the consistency of classification and were not a complete replication of the process used by the primary group, which was most familiar with items in the respective content area.

Panelists' comments on the items were also recorded during the item review process, including observations about specific item characteristics, rationales for the classifications, and judgments about whether items exceeded the grade-level descriptions in the framework. In addition, general comments made by the panel about the assessments and frameworks in plenary or during the separate group discussions were recorded and used to inform the discussions in this report.

3.3. Methods Used for NAEP/PISA Comparisons

As noted in the previous section, the NAEP/PISA comparisons were accomplished by two subgroups of the panelists retained for this component of the comparison study conducted on the third day of the expert panel meeting.

The first group reviewed and classified the 85 PISA items to the NAEP framework for content, grade level, and level of *mathematical complexity*. This group included representatives from each of the previous content groups to ensure a consistent method of classification as that used for the TIMSS items for any part of the NAEP content framework.

The second group classified 79 NAEP *problem solving* and extended-response items from the eighth- and twelfth-grade assessments. These items were selected because of their potential alignment with PISA's emphasis on problem solving. They were drawn from both the eighth-grade and twelfth-grade assessments since PISA's 15-year old target population falls between these two grade levels. The group classified the set of NAEP items with respect to the main dimensions of the PISA framework, including the following:

- *overarching idea*;
- *competency cluster*; and
- *situation or context*.

A set of example items from the PISA framework were used to illustrate the classification procedures. The group also commented on whether or not the NAEP items might appear on the PISA assessment and, if not, how they were different from PISA items that might assess comparable mathematics content and processes.

This section reviewed the methods used for this comparison study. The next section compares the assessments overall with respect to content coverage, grade level, levels of mathematic complexity, and item format.

4. Overall Comparisons

The classifications made by the expert panel as well as the information provided by each assessment provide rich data that can be organized and analyzed in numerous ways. This section compares the assessments overall with respect to content coverage, grade level, *mathematical complexity* level, and item format.

4.1. Content Coverage

Tables 4 and 5 compare the distribution of items from each assessment across the main content areas as defined in the NAEP and TIMSS frameworks. The tables compare NAEP and TIMSS item classifications according to their own respective frameworks, with item classifications according to the framework of the other assessment.[26] Generally speaking, the three assessments appear to share similar boundaries for the definition of mathematics content, with nearly all items from each assessment classified as being consistent with the definitions of the five content areas in both the NAEP and TIMSS frameworks. There were only a few items (2 percent of NAEP items at the fourth grade, 1 percent of NAEP items at the eighth grade, and 1 percent of TIMSS items at the eighth grade) that were not classified by the panel at the broad content area level to the other assessment's framework. These classifications, however, do not consider grade level correspondence, which is discussed in the next section.

Using the five *strands* of the NAEP framework to compare NAEP and TIMSS at both the fourth- and eighth-grade levels, the two assessments have very similar distributions of items across the five strands (table 4). At the fourth-grade level on both assessments, the highest percentage of items was classified to *number sense, properties, and operations* (40 and 42 percent, respectively) and the area with the lowest percent was *data analysis, statistics, and probability* (10 percent for both assessments). A similar pattern is true at the eighth-grade level as well, although the distribution of items was more balanced. The largest difference between the two assessments at eighth grade is in *number sense, properties, and operations*, where one-third of TIMSS items were classified compared to about one-quarter of NAEP items.

The distribution of NAEP and TIMSS items across the five *content domains* of the TIMSS framework is very similar to that based on the corresponding content areas of the NAEP framework, giving at least partial support to the idea that the broad content areas defined in the NAEP and TIMSS framework are similar (table 5). While the NAEP and TIMSS assessments appear to place similar emphases on these broad content areas, there are substantial differences between the assessments noted when the item content is examined more closely, as discussed in the following sections.

[26] The classifications of items to their own framework were provided by the assessment developers. Cross classifications of NAEP, TIMSS, and PISA items to the other's assessment framework were done by the expert panel.

Table 4. Percentage of NAEP, TIMSS, and PISA mathematics items classified to the content strands in the NAEP 2003 mathematics framework, by grade/age and survey: 2003

NAEP content strand	Grade 4		Grade 8		15-year-olds
	NAEP[1]	TIMSS[2]	NAEP[1]	TIMSS[2]	PISA[2]
Total number of items	181	145	197	180	85
	Percentage distribution				
Number sense, properties, and operations	42	40	26	33	22
Measurement	18	19	15	14	18
Geometry and spatial sense	15	15	19	18	12
Data analysis, statistics, and probability	10	10	15	11	40
Algebra and functions	14	15	25	23	11
Classified to multiple strands	0	0	0	0	2
Not classified to a content strand	0	0	0	1	0

[1] NAEP items classified by NAEP developers.
[2] TIMSS items classified by expert panel.
NOTE: Data reflect the percentage of items classified to the NAEP content framework at any level of specificity (*content strand*, *topic*, or *subtopic*). Multi-part items were treated as one item for classification purposes and only contribute one to the total. Items classified to multiple *content strands* were counted in each relevant category. Two PISA items were classified to multiple *content strands*: one to *measurement* and *data analysis, statistics, and probability* and one to *algebra and functions* and *geometry and spatial sense*. One TIMSS *data* item at eighth grade was not classified to a NAEP *content strand*. Detail may not sum to totals because of rounding, omitted items, or items classified to multiple *content strands*.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment; International Association for the Evaluation of Educational Achievement, Trends in International Mathematics and Science Study (TIMSS) 2003 Assessment; Organization for Economic Cooperation and Development, Program for International Student Assessment (PISA) 2003 Mathematical Literacy Assessment; and U.S. Department of Education, National Assessment Governing Board, *Mathematics Framework for the 2003 National Assessment of Educational Progress,* 2002.

Table 5. Percentage of NAEP and TIMSS mathematics items classified to the content domains in the TIMSS 2003 mathematics framework, by grade and survey: 2003

TIMSS content domain	Grade 4		Grade 8	
	NAEP[1]	TIMSS[2]	NAEP[1]	TIMSS[2]
Total number of items	181	145	197	180
	Percentage distribution			
Number	44	38	29	31
Measurement	16	21	12	16
Geometry	17	14	24	17
Data	10	10	15	13
Algebra	12	16	19	23
Not classified to a content domain	2	0	1	0

[1] NAEP items classified by expert panel.
[2] TIMSS items classified by TIMSS developers.
NOTE: Data reflect the percentage of items classified to the TIMSS content framework at any level of specificity (*content domain*, *topic area*, or *objective*). Multi-part items were treated as one item for classification purposes and only contribute one to the total. Five NAEP items were not classified to *content domains* on the TIMSS 2003 framework. These items included three fourth-grade items in *algebra*, *measurement*, and *geometry* and two eighth-grade items in *algebra*. Detail may not sum to totals because of rounding or omitted items.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment; International Association for the Evaluation of Educational Achievement, Trends in International Mathematics and Science Study (TIMSS) 2003 Assessment; and International Study Center, Lynch School of Education, Boston College, *TIMSS Assessment Frameworks and Specifications 2003, 2nd ed*., 2003.

PISA has a different balance across the main content areas than NAEP or TIMSS, with a much higher percentage of items classified to *data analysis, statistics, and probability* on the NAEP framework (40 percent, compared to 15 and 11 percent on the NAEP and TIMSS eighth-grade assessments, respectively). PISA also has relatively fewer items classified to *algebra and functions* (11 percent, compared to about one-quarter for NAEP and TIMSS at the eighth grade) and *number sense, properties, and operations* (22 percent, compared to 26 and 33 percent for the eighth-grade NAEP and TIMSS assessments, respectively).

At the broad content area level, virtually all NAEP, TIMSS, and PISA items were classified as being consistent with the basic definitions of the content areas in the other assessments when the grade level correspondence to the other framework is not considered. At the topic level, there was also a high level of content match for all three assessments (table 6). For all assessments (and grades), at least 95 percent of items were classified at the topic level on the other assessment frameworks. However, at the finest level of classification (either the subtopic level or the topic level, when no subtopics existed), the match between items and frameworks was not as universal. Still, classification at this level was relatively high, with about 80 percent of fourth-grade items and 85 percent of eighth-grade items for both NAEP and TIMSS being classified at the subtopic level on the other assessment frameworks. The level of content match for the PISA items was even higher, with 92 percent of items classified at the subtopic level to the NAEP framework. These findings indicate that there is a generally high level of agreement between the three assessments regarding the general definitions of the mathematics content areas; the level of agreement decreases at more specific levels of classification.

Table 6. Percentage of NAEP, TIMSS, and PISA mathematics items classified to other assessment framework at the topic or subtopic level, by grade/age and survey: 2003

Level of content classification	Grade 4		Grade 8		15-year-olds
	NAEP items to TIMSS framework	TIMSS items to NAEP framework	NAEP items to TIMSS framework	TIMSS items to NAEP framework	PISA items to NAEP framework
Topic level	96	98	98	99	95
Subtopic level[1]	80	80	84	85	92

[1]Includes items classified to a subtopic or to a topic, when no subtopics exist in the NAEP framework.
NOTE: Data reflect the percentage of items that were classified by the expert panel to the topic and subtopic levels of the other assessment framework in any content area at any grade. Items classified to multiple topics or subtopics are considered to match those levels of classification and are counted only once.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment; International Association for the Evaluation of Educational Achievement, Trends in International Mathematics and Science Study (TIMSS) 2003 Assessment; Organization for Economic Cooperation and Development, Program for International Student Assessment (PISA) 2003 Mathematical Literacy Assessment; U.S. Department of Education, National Assessment Governing Board, *Mathematics Framework for the 2003 National Assessment of Educational Progress,* 2002; and International Study Center, Lynch School of Education, Boston College, *TIMSS Assessment Frameworks and Specifications 2003, 2nd ed*., 2003.

4.2. Grade Level

The cross-classification data were used to examine the extent to which items from each assessment map to the other assessment frameworks at corresponding grade levels. Figures 1-A and 1-B show the percentage of items in the NAEP, TIMSS, and PISA assessments overall that were classified at each grade level in the other assessment frameworks. For these overall comparisons, the percentages classified at each grade level of the other assessment frameworks reflect items that were classified at the subtopic, topic, or broad content area level.[27]

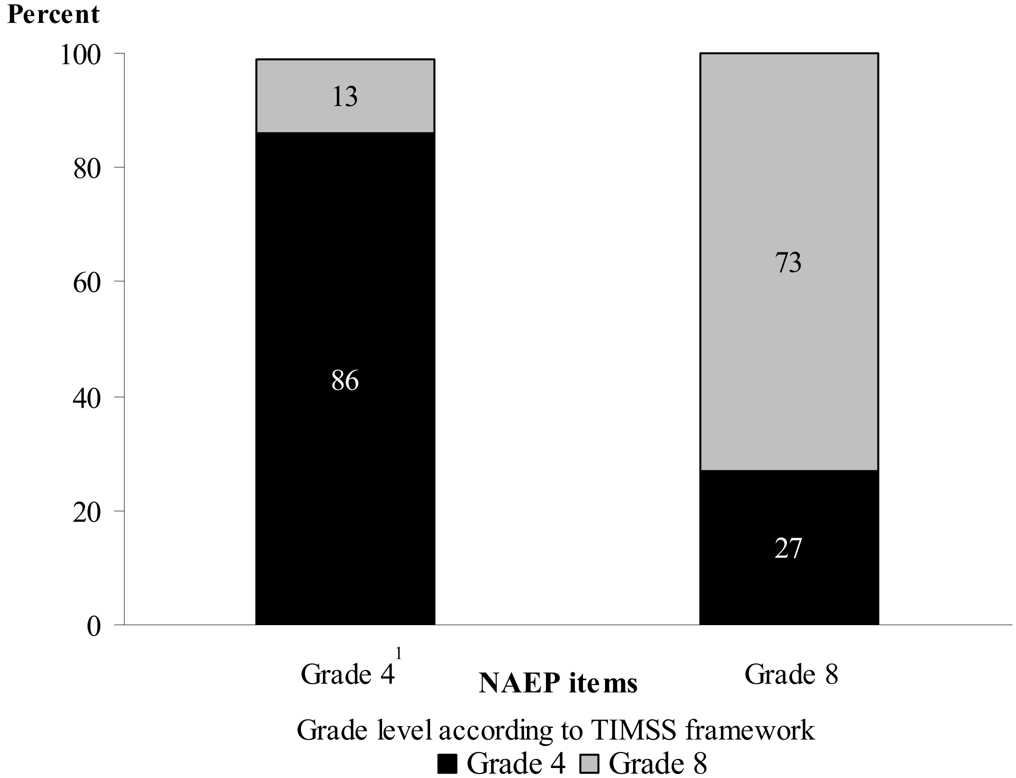
Comparing NAEP and TIMSS, there appears to be a moderately high level of consistency regarding what is considered to be fourth-grade content and what is considered to be eighth-grade content. On both the fourth- and eighth-grade assessments, most NAEP items were placed at the same grade level using the definitions and criteria of the TIMSS framework, and vice versa (figure 1-A). Eighty-six percent of all fourth-grade NAEP items were classified at the fourth grade and 73 percent of all eighth-grade NAEP items were classified at the eighth grade according to the TIMSS framework. Ninety-four percent of fourth-grade TIMSS items were classified at the fourth grade and 90 percent of eighth-grade TIMSS items were classified at the eighth grade on the NAEP framework (figure 1-B). Of the remaining TIMSS eighth-grade items, 6 percent were classified at the fourth grade and 3 percent at the twelfth grade according to the NAEP framework. It should be noted that since the TIMSS framework includes only fourth and eighth grades, it was not possible for the panel to classify NAEP items at a grade level higher than the eighth grade on the TIMSS framework. At the same time, there were no comments recorded that suggested that any of the NAEP items exceeded the TIMSS eighth-grade descriptions.

Although PISA items are not designed to meet criteria for any specific grade, panelists did examine how they correspond to grade level(s) according to the NAEP framework. Most PISA items, 85 percent, were classified as being consistent with the eighth-grade NAEP framework, similar to the results for the TIMSS eighth-grade assessment (figure 1-B). Twelve percent of PISA items were classified at the fourth grade and four percent at the twelfth grade. Thus, according to the definitions

[27] The analyses for each of the content area comparisons in section 5 further examine the degree to which items match topics and subtopics at particular grades.

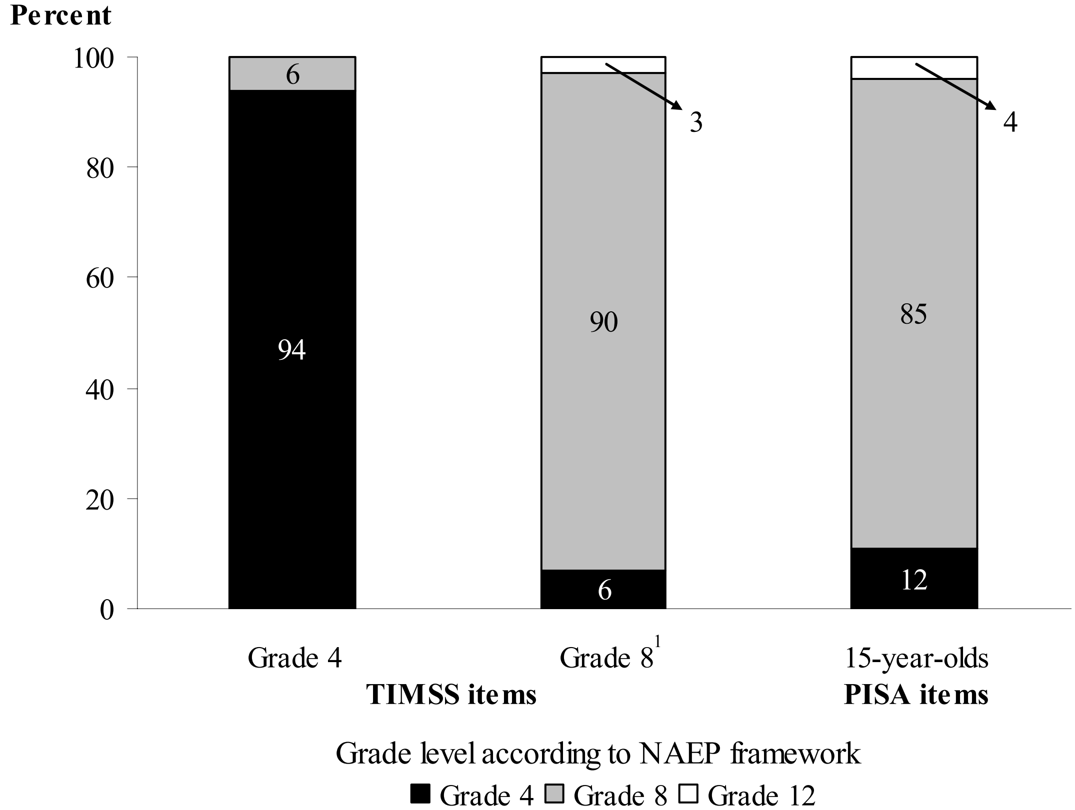
of the NAEP framework, the mathematics content of the PISA assessment is predominantly at the eighth-grade level.

Figure 1-A. Percentage distribution of NAEP mathematics items classified at each grade level according to the TIMSS mathematics framework, by grade: 2003



[1]Two NAEP fourth-grade items that the panel did not classify with respect to grade level on the TIMSS assessment framework are not included.
NOTE: Data reflect expert panel classifications of grade level to the TIMSS content framework. Detail may not sum to totals because of rounding or omitted items.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment; and International Study Center, Lynch School of Education, Boston College, *TIMSS Assessment Frameworks and Specifications 2003, 2nd ed*., 2003.

Figure 1-B. Percentage distribution of TIMSS and PISA mathematics items classified at each grade level according to the NAEP mathematics framework, by grade/age: 2003



[1] One TIMSS eighth-grade item that the panel did not classify with respect to grade level on the NAEP assessment framework is not included.
NOTE: Data reflect expert panel classifications of grade level to the NAEP content framework . Detail may not sum to totals because of rounding or omitted items.
SOURCE: International Association for the Evaluation of Educational Achievement, Trends in International Mathematics and Science Study (TIMSS) 2003 Assessment; Organization for Economic Cooperation and Development, Program for International Student Assessment (PISA) 2003 Mathematical Literacy Assessment; and U.S. Department of Education, National Assessment Governing Board, *Mathematics Framework for the 2003 National Assessment of Educational Progress,* 2002.

The fact that grade-level agreement is less for the comparison between NAEP items and the TIMSS framework than for the TIMSS items to NAEP framework comparison may be related at least in part to the presence of NAEP cross-grade items that were developed to be used at multiple grades. Eighteen percent of fourth-grade NAEP items also appeared on the eighth-grade NAEP assessment and an additional 10 percent appeared on the eighth- and twelfth-grade NAEP assessments (data not shown). This represented 16 and 10 percent, respectively, of the eighth-grade assessment items. An additional 18 percent of eighth-grade items were administered at both the eighth and twelfth grades (but not the fourth grade). Assuming that cross-grade items are designed to be appropriate for all the grade levels at which they are administered, the presence of cross-grade items may affect the grade level correspondence to the TIMSS 2003 framework (TIMSS does not include cross-grade items). As reflected in table 7, most of the cross-grade items were classified by the panel at the lower or lowest grade level according to the TIMSS framework. For example, about 80 percent of NAEP items administered at grades 4 and 8 or grades 4, 8, and 12 were classified at the fourth grade level according to the TIMSS framework. As a result, the grade level match for all fourth-grade items, including both single-grade and cross-grade items, is similar to that for the single-grade items administered only at the fourth grade (86 percent compared to 88 percent). The difference is much greater at the eighth grade, where 93 percent of the NAEP single-grade items were classified at the corresponding grade level according to the TIMSS framework, compared to 73 percent for the NAEP eighth-grade assessment overall when the cross-grade items are included.

Table 7. Percentage of NAEP single-grade and cross-grade mathematics items classified at each grade level according to the TIMSS mathematics framework: 2003

Grade level according to the TIMSS 2003 framework	Total		NAEP item type				
			Single-grade items		Cross-grade items		
	Grade 4	Grade 8	Grade 4 only	Grade 8 only	Grades 4 and 8	Grades 4, 8, and 12	Grades 8 and 12
Grade 4	86	27	88	7	84	79	11
Grade 8	13	73	11	93	16	21	89

NOTE: Data reflect expert panel classifications of grade level to the TIMSS 2003 content framework. Single-grade items are administered at one grade level; cross-grade items are administered at more than one grade (4 and 8; 4, 8, and 12; or 8 and 12); totals reflect single-grade and cross-grade items included in the assessment at each grade level. Two NAEP fourth-grade items that the expert panel did not classify with respect to grade level are not included. Detail may not sum to totals because of rounding or omitted items.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment; and International Study Center, Lynch School of Education, Boston College, *TIMSS Assessment Frameworks and Specifications 2003, 2nd ed.*, 2003.

4.3. Levels of Mathematical Complexity

Comparison of Mathematical Complexity Levels for NAEP, TIMSS, and PISA Assessments

All items from all three assessments were classified based on the three levels of *mathematical complexity* defined in the NAEP 2005 framework (NAGB 2004) and briefly described earlier in this report (see section 3.2). The *mathematical complexity* dimension of the NAEP 2005 framework replaces the previous NAEP cognitive dimensions of *mathematical abilities* and *mathematical power*. *Mathematical complexity* reflects the demand placed on students by the items and focuses on the properties of items rather than inferred abilities of students. *Mathematical complexity* is not necessarily related to item difficulty, which is based on actual student performance. *Mathematical complexity* should also be independent of curriculum, meaning it is determined assuming that students are familiar with the mathematical content of the item. The three levels—low, moderate, and high—are used to describe an increasing level of complexity of steps and processes required of students in order to succeed on an item and are based in part on the degree to which items require flexible or abstract thinking. A more complete description of the levels of *mathematical complexity* is included in appendix B.

Table 8 shows the percentage distribution of items in all three assessments classified by the expert panel at the three levels of *mathematical complexity*. It should be noted that although the classifications of *mathematical complexity* level are based on definitions in the NAEP 2005 framework, the NAEP 2003 items were not originally developed using this new dimension specified in the revised NAEP framework. Overall, NAEP and TIMSS exhibited very similar profiles of *mathematical complexity* level. At the fourth-grade level, the percentages of items classified in the three levels of *mathematical complexity* were nearly identical. At the eighth-grade level, a slightly higher percentage of TIMSS items were placed in the moderate level than NAEP items (46 percent compared to 39 percent), and a correspondingly lower percentage of items were placed in the low level (51 percent compared to 57 percent). For both NAEP and TIMSS, only a few items at each grade level (less than 5 percent) were classified at the high complexity level.

Table 8. Percentage distribution of NAEP, TIMSS, and PISA mathematics items across levels of mathematical complexity, by grade/age and survey: 2003

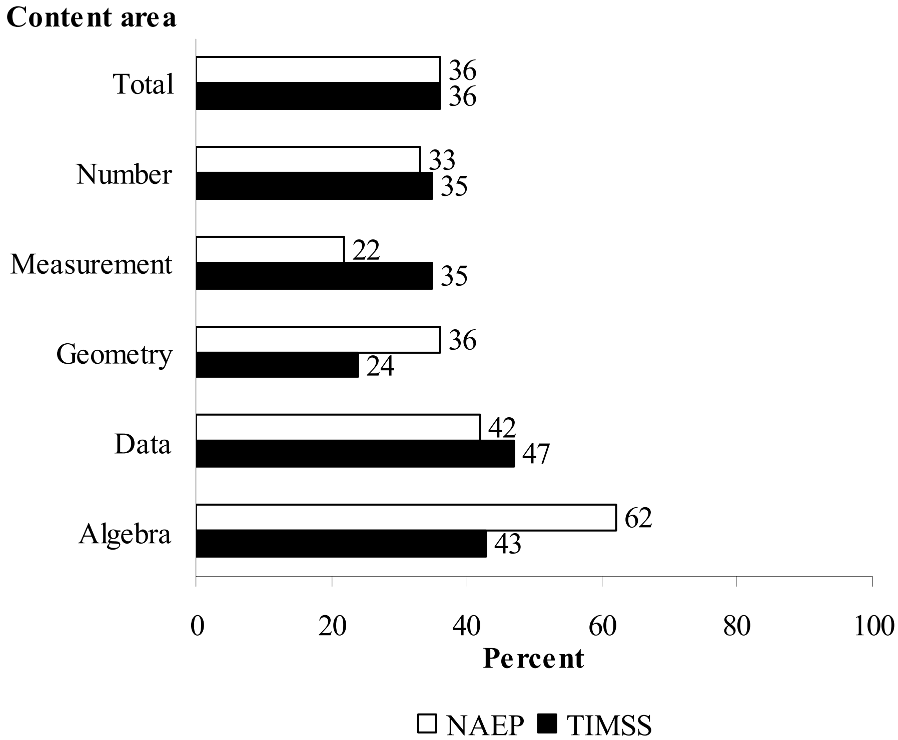
Mathematical complexity level	Grade 4		Grade 8		15-year-olds
	NAEP	TIMSS	NAEP	TIMSS	PISA
Low	64	64	57	51	29
Moderate	33	34	39	46	64
High	3	2	4	3	7

NOTE: Data reflect expert panel classifications of *mathematical complexity* level as defined in the NAEP 2005 mathematics framework. Levels of *mathematical complexity* (low, moderate, and high) form an ordered description of the cognitive demands of the item. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment; International Association for the Evaluation of Educational Achievement, Trends in International Mathematics and Science Study (TIMSS) 2003 Assessment; Organization for Economic Cooperation and Development, Program for International Student Assessment (PISA) 2003 Mathematical Literacy Assessment; and U.S. Department of Education, National Assessment Governing Board, *Mathematics Framework for the 2005 National Assessment of Educational Progress*, 2004.

Relative to both NAEP and TIMSS, PISA has a much higher proportion of items at the moderate level than the low level (64 percent compared to 29 percent). Although a relatively higher percentage of PISA items were placed in the high complexity level than in NAEP or TIMSS, these items nevertheless made up only a small percentage of all PISA items (7 percent).

The percentage of items at the moderate or high complexity level in the NAEP and TIMSS assessments varied across the main content areas (figure 2). At the fourth-grade level, the *data* and *algebra* content areas had the highest proportion of items at the moderate or high complexity level for both NAEP and TIMSS (more than 40 percent). NAEP, however, had a much higher percentage in *algebra* (62 percent compared to 43 percent in TIMSS).

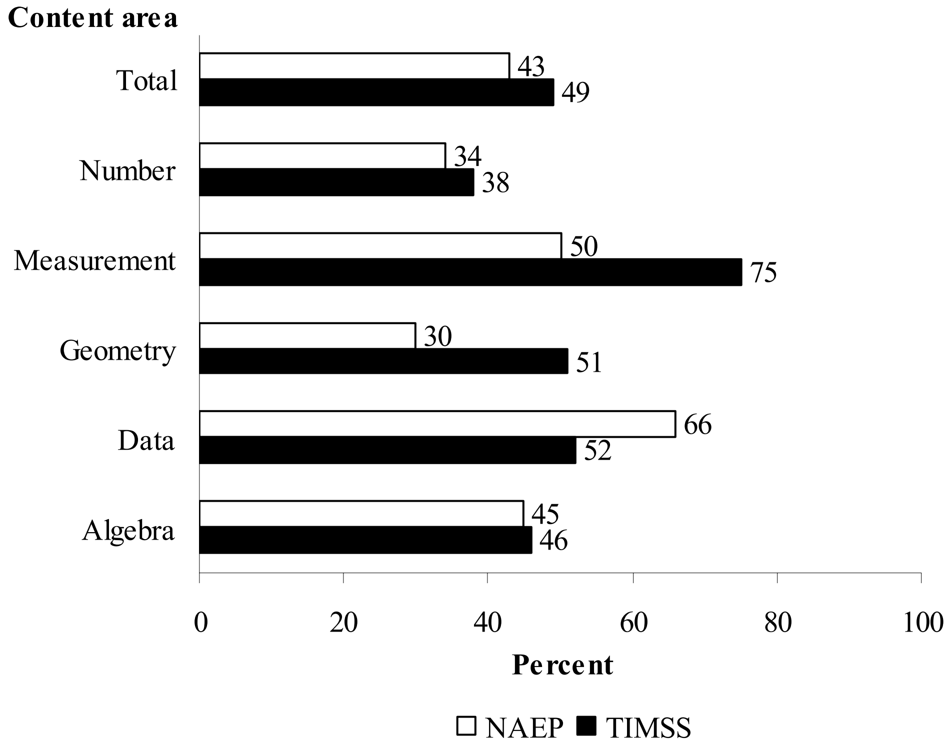
Figure 2. Percentage of NAEP and TIMSS fourth-grade items classified as moderate or high mathematical complexity level, by mathematics content area: 2003



[1] The total category refers to all items combined.
NOTE: Data reflect expert panel classifications of *mathematical complexity* level as defined in the NAEP 2005 mathematics framework.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment; International Association for the Evaluation of Educational Achievement, Trends in International Mathematics and Science Study (TIMSS) 2003 Assessment; and U.S. Department of Education, National Assessment Governing Board, *Mathematics Framework for the 2005 National Assessment of Educational Progress*, 2004.

At the eighth-grade level, relatively large differences between the NAEP and TIMSS items are found across the content areas of *measurement*, *geometry,* and *data* (figure 3). For TIMSS, there were a high proportion of items of moderate or high complexity level in *measurement*, with 75 percent of items compared to 50 percent of NAEP items. *Geometry* was also an area of higher complexity in TIMSS, with about half of items of moderate or high complexity level compared to only 30 percent of *geometry* items in NAEP. For NAEP, the content area with the highest complexity level was *data*, with 66 percent of NAEP items classified at the moderate or high level compared to 52 percent of TIMSS items. In none of the content areas did the percentage of items classified at the high complexity level exceed 10 percent for either assessment at fourth or eighth grade.

Figure 3. Percentage of NAEP and TIMSS eighth-grade items classified as moderate or high mathematical complexity level, by mathematics content area: 2003



[1] The total category refers to all items combined.
NOTE: Data reflect expert panel classifications of *mathematical complexity* level as defined in the NAEP 2005 mathematics framework.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment; International Association for the Evaluation of Educational Achievement, Trends in International Mathematics and Science Study (TIMSS) 2003 Assessment; and U.S. Department of Education, National Assessment Governing Board, *Mathematics Framework for the 2005 National Assessment of Educational Progress*, 2004.

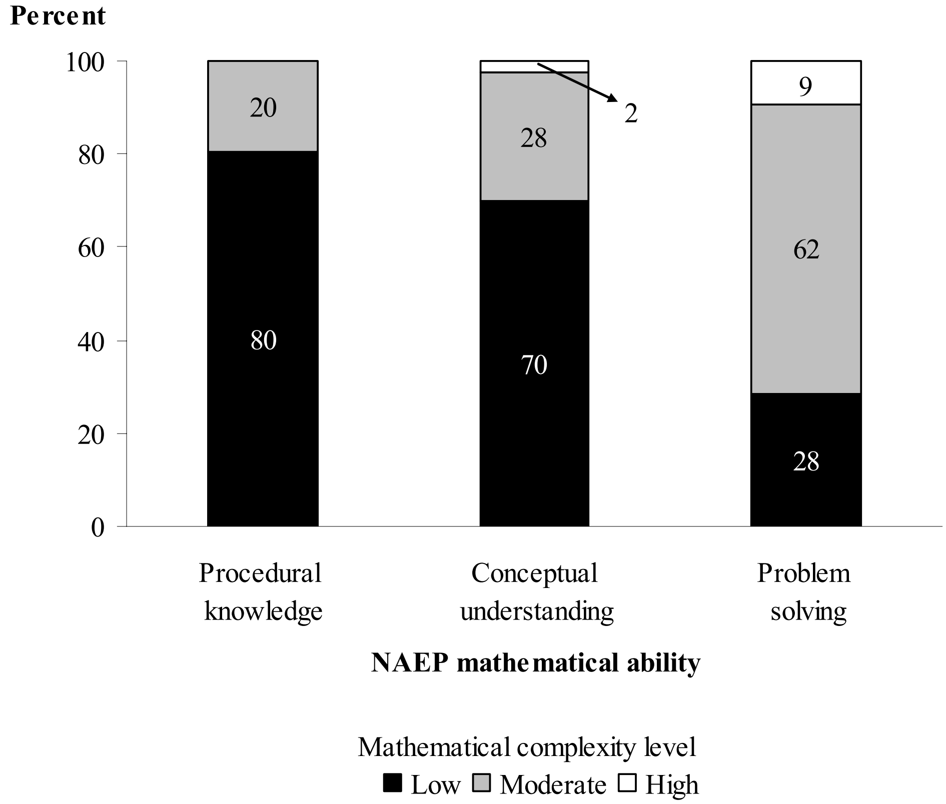
The frameworks for all three assessments have dimensions that describe and classify items according to cognitive abilities they require of students. Although a great deal can be learned from the framework documents about the meanings and intentions of each category and the ways in which they might influence item development, the classifications of all items to a common system of describing cognitive demand, the level of *mathematical complexity* in the NAEP 2005 framework, provides an additional means of examining these systems. Comparing the classification of items on the cognitive dimensions of each framework to the classifications for levels of complexity may reveal, for example, the degree to which each system is related to *mathematical complexity* level and the degree to which the systems represent a hierarchy of cognitive skills. To relate the expert panel classifications of *mathematical complexity* level to the intentions of the original framework, the NAEP, TIMSS and PISA items developed for each of the original cognitive categories were compared with respect to the proportion of low, moderate, and high complexity level. The results of these cross-classifications by cognitive categories are shown in the following sections.

Cross-classification by NAEP Mathematical Abilities Dimension

Figure 4 displays the percentage of items from each of the NAEP 2003 *mathematical abilities* categories (*procedural knowledge*, *conceptual understanding*, and *problem solving*) classified at each level of *mathematical complexity* according to the NAEP 2005 framework.[28] Not surprisingly, most of the *procedural knowledge* items were classified at the low complexity level (80 percent). The majority of items in the *conceptual understanding* category (70 percent) also were classified at the low complexity level. While a higher proportion of items were classified at the moderate complexity level in the *conceptual understanding* category, these first two *mathematical ability* categories do not reflect a clear hierarchy of complexity. Example 1 in appendix E shows a NAEP *conceptual understanding* item classified at the low *mathematical complexity* level. Many items from the *conceptual understanding* category tested student knowledge of a definition or concept with no additional steps or analysis required. These types of items are consistent with the definition of low complexity level. In comparison, most *problem solving* items were classified at the moderate complexity level (62 percent), although 28 percent were classified at the low complexity level and a number of these were word problems (figure 4). Example 2 in appendix E shows a NAEP *problem solving* item classified at the low *mathematical complexity* level. Most of the NAEP word problems were from the *problem solving* category, but many of these items were classified by the panel at the low complexity level because the underlying mathematics problem was fairly obvious and the solution required little abstract or flexible thought. Few items even in the *problem solving* category were classified at the high complexity level (9 percent).

[28] Because not all NAEP items are given classifications for mathematical power, this section examines only classifications for mathematical ability.

Figure 4. Percentage distribution of fourth- and eighth-grade NAEP mathematics items across mathematical complexity levels, by NAEP mathematical ability category: 2003

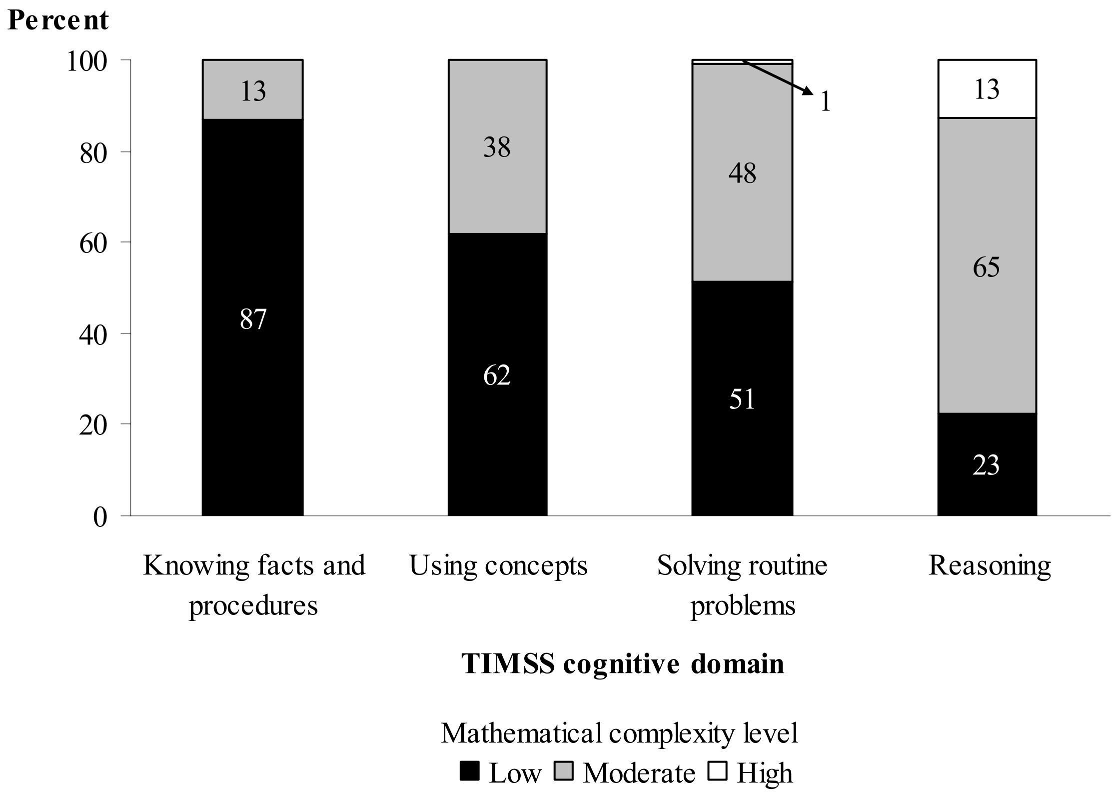


NOTE: Data reflect both fourth- and eighth-grade items combined. Classifications by *mathematical abilities* were provided by NAEP assessment developers; classifications by *mathematical complexity* level made by expert panel according to definitions in the NAEP 2005 framework. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment; and U.S. Department of Education, National Assessment Governing Board, *Mathematics Framework for the 2005 National Assessment of Educational Progress*, 2004.

Cross-classification by TIMSS Cognitive Domains

Figure 5 displays the percentage of items from each of the TIMSS 2003 *cognitive domain* categories (*knowing facts and procedures*, *understanding concepts*, *solving routine problems*, and *reasoning*) classified at each level of *mathematical complexity* according to the NAEP 2005 framework. Similar to the NAEP *procedural knowledge* category, a very high percentage of TIMSS *knowing facts and procedures* items were classified at the low complexity level (87 percent). Also similar to the classification of NAEP *conceptual understanding* items, 62 percent of TIMSS *understanding concepts* items were classified at the low complexity level for similar reasons, most notably that recalling or using a concept does not necessarily require abstract or flexible thought. Items developed for the category of *solving routine problems* in TIMSS were about evenly distributed across the low and moderate complexity levels (51 and 48 percent, respectively), with only one item at the high complexity level (1 percent). The *mathematical complexity* level distribution for TIMSS *reasoning* items is similar to that for the NAEP *problem solving* items, with 23 percent low, 65 percent moderate, and 13 percent at the high complexity level. A TIMSS *reasoning* item classified at the low *mathematical complexity* level is illustrated by Example 3 in appendix E. A review of items indicates that most items representing non-routine situations were classified as *reasoning* in TIMSS. Using the different classification system, however, the panel judged some of these items as not requiring much abstract or flexible thinking and classified them as low or moderate with respect to the level of *mathematical complexity*.

Figure 5. Percentage distribution of fourth- and eighth-grade TIMSS mathematics items across mathematical complexity levels, by TIMSS cognitive domain category: 2003

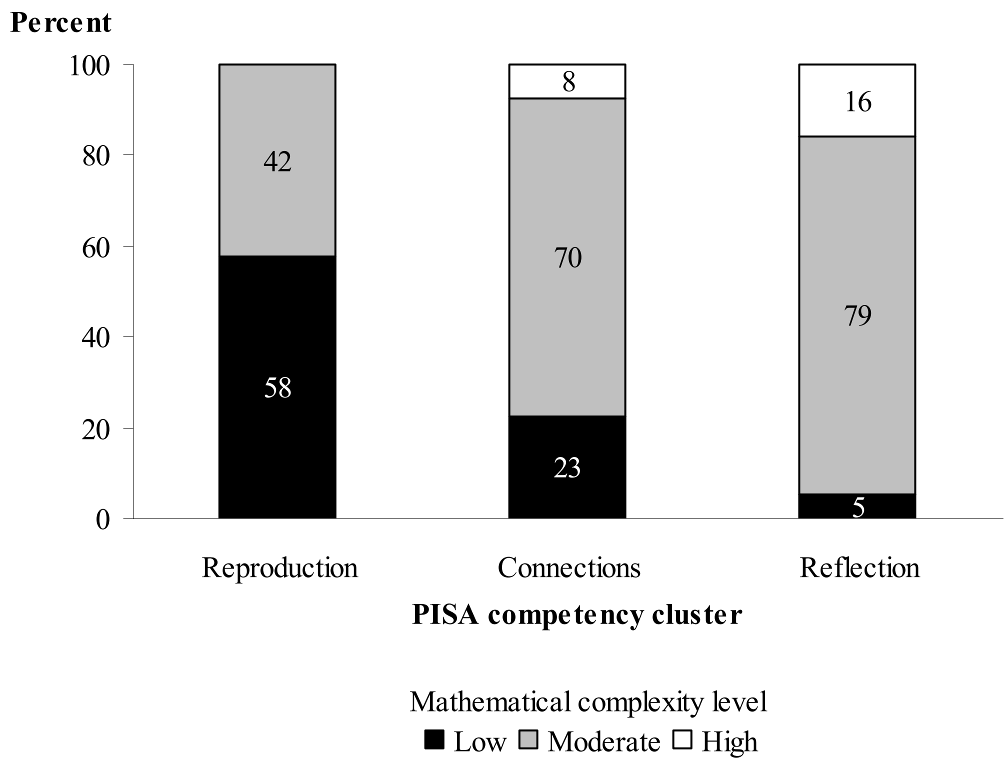


NOTE: Data reflect fourth- and eighth-grade items combined. Classifications by *cognitive domains* were provided by TIMSS assessment developers; classifications by *mathematical complexity* level were made by expert panel according to definitions in the NAEP 2005 framework. Multi-part items were counted in the reasoning category if any of the item parts were classified as reasoning; otherwise, multi-part items were counted in the category corresponding to the most frequent category across item subparts. Detail may not sum to totals because of rounding.
SOURCE: International Association for the Evaluation of Educational Achievement, Trends in International Mathematics and Science Study (TIMSS) 2003 Assessment; U.S. Department of Education, National Assessment Governing Board, *Mathematics Framework for the 2005 National Assessment of Educational Progress*, 2004; and International Study Center, Lynch School of Education, Boston College, *TIMSS Assessment Frameworks and Specifications 2003, 2nd ed.*, 2003.

Cross-classification by PISA Competency Clusters

Figure 6 displays the percentage of items from each PISA 2003 *competency cluster* (*reproduction*, *connections*, and *reflections*) classified at each level of *mathematical complexity* according to the NAEP 2005 framework. Many items in all PISA *competency clusters* were found to be at the moderate complexity level. That 42 percent of PISA *reproduction* items were classified at the moderate complexity level may at first appear inconsistent with the description of the *reproductions* cluster in the PISA framework: “The competencies in this cluster essentially involve reproduction of practiced knowledge….” (OECD 2003). However, the framework presents a fairly broad interpretation of this phrase, providing illustrations of “the reproduction of practiced knowledge” in settings such as “thinking and reasoning,” “modeling,” and “problem posing and solving.” Example 4 in appendix E presents a PISA *reproduction* item classified at the moderate *mathematical complexity* level. A high degree of correspondence is seen between the definitions provided in the PISA framework and the percentage of low complexity level items found across the series of competency clusters from *reproduction* to *connections* to *reflection*. The trend is less compelling when considering the differentiation between the moderate and high complexity levels. While there is the highest proportion of high complexity level items in the PISA *reflections* category than for any other set of items across the three assessments, this still represents less than 20 percent of items.

Figure 6. Percentage distribution of PISA mathematics items across mathematical complexity levels, by PISA competency cluster: 2003



NOTE: Classifications by *competency cluster* were provided by PISA assessment developers; classifications by *mathematical complexity* level were made by expert panel according to definitions in the NAEP 2005 framework. Detail may not sum to totals because of rounding.
SOURCE: Organization for Economic Cooperation and Development, Program for International Student Assessment (PISA) 2003 Mathematical Literacy Assessment; U.S. Department of Education, National Assessment Governing Board, *Mathematics Framework for the 2005 National Assessment of Educational Progress*, 2004; and Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), *The PISA 2003 Assessment Framework: Mathematics, Reading, Science and Problem Solving Knowledge and Skills*, 2003.

4.4. Item Format

The items in the NAEP, TIMSS, and PISA assessments were also compared with respect to the types of item formats used and their proportion on the assessments. Table 9 shows the percentage distribution of NAEP, TIMSS, and PISA items by item format, including the main formats of multiple choice and constructed response as well as different types of constructed-response items defined by each assessment. Items vary in difficulty level and cognitive demand regardless of format, though constructed-response items can be particularly important in assessing students' abilities to communicate their mathematical understanding and explain their solutions. Including a variety of item types ensures that a range of knowledge and skills is being assessed. Examples are included in appendix E to illustrate some of the item formats across the three assessments: a short constructed-response and an extended constructed-response item from the NAEP assessment (Examples 5 and 6), an extended constructed-response item from the TIMSS assessment (Example 7), a complex multiple-choice item and an open constructed-response item from the PISA assessment (Examples 8 and 9).

At both grades, the distribution of NAEP and TIMSS items by item format is similar, with more than 60 percent multiple-choice items, although at eighth grade TIMSS has a slightly higher proportion of multiple choice (71 percent). At both the fourth- and eighth-grade levels, the TIMSS assessment identified a relatively higher proportion of items as extended constructed response than the NAEP assessments. However, the definition and nature of extended-response items may not be the same across the two assessments. In both assessments, the constructed-response items are scored with rubrics that are customized for each item. In TIMSS, the short-answer items are scored with a 2-level rubric (correct/incorrect) and extended-response items with a 3-level rubric (correct/partial/incorrect). In NAEP, the short-answer items may be scored with either a 2-level or 3-level rubric, while extended-response items may have up to five score levels (extended/satisfactory/partial/minimal/incorrect). The criteria that differentiate each response level vary by item within and across assessments.

PISA relies on multiple-choice items far less than do NAEP or TIMSS. Only one-third of PISA items use a multiple-choice format, compared to approximately two thirds of items on NAEP and TIMSS. In PISA, this item format category includes traditional multiple-choice items as well as "complex multiple choice" items that require students to answer a series of multiple-choice or true-false questions based on the same information (see Example 8). In most cases, students must answer all questions correctly in order to receive credit for the item, while a few items allow partial credit for answering one or more, but not all, questions correctly. There also are three types of constructed response items identified in PISA: short response, closed response, and open response. Short-response items and closed constructed-response items require students to write or otherwise indicate the answer to the question but not show their work. These two item types are similar although the closed constructed-response items are more constrained to a specific set of possible answers. These types of items are scored dichotomously and do not allow partial credit. They are, thus, different from the short constructed-response items in NAEP, which sometimes allow for partial credit and often require students to write brief explanations or show their work. The PISA open constructed-response items may require students to show their work, and there may be numerous ways in which students may receive credit. Partial credit rubrics are often used for these items, but only with three levels (full/partial/no credit). In this respect, they are more similar to the extended constructed-response items in TIMSS and some of the short-answer items in NAEP.

Table 9. Percentage distribution of NAEP, TIMSS, and PISA mathematics items across item formats, by grade/age and survey: 2003

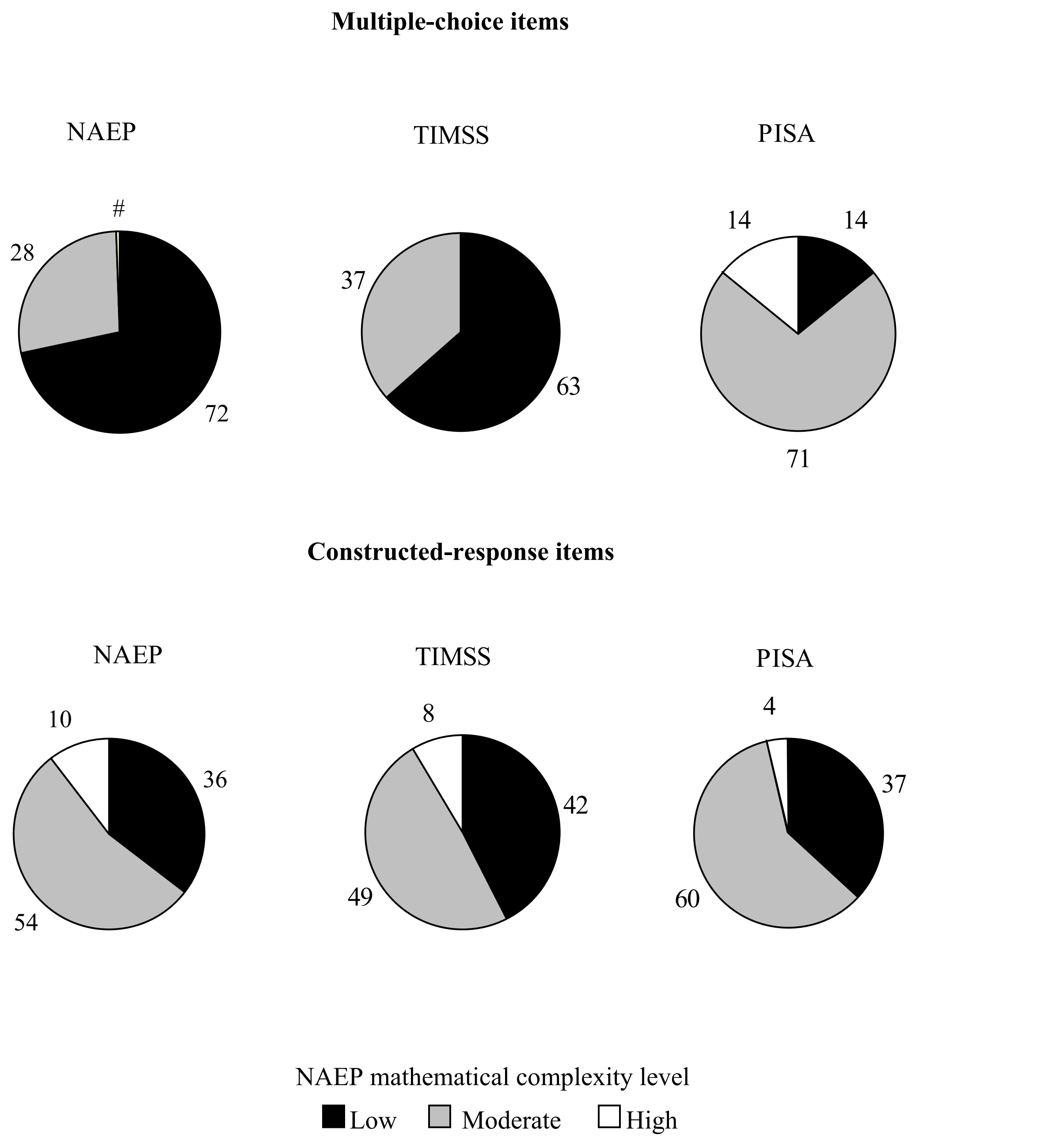
Item format	Grade 4		Grade 8		15-year -olds
	NAEP	TIMSS	NAEP	TIMSS	PISA
Multiple choice	63	63	65	71	33[1]
Constructed response	37	37	35	29	67
Short answer (short response)	33	28	30	16	27
Extended response	4	10	5	13	—
Closed response	—	—	—	—	15
Open response	—	—	—	—	25

— Not available. These constructed-response item format classifications were not included in the information provided by the assessment developers.
[1] PISA also includes "complex multiple-choice" items (13 percent), which are reflected in the percentage of multiple-choice items.
NOTE: The breakdown of constructed-response items was provided by the assessment developers for the NAEP and PISA items. For the TIMSS items, the assignment was based on examination of the items and the level of score points in the scoring guides in accordance with information provided by the TIMSS assessment developers—extended response items reflect multi-part items and items that were scored with 3-level scoring rubrics. Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment; International Association for the Evaluation of Educational Achievement, Trends in International Mathematics and Science Study (TIMSS) 2003 Assessment; and Organization for Economic Cooperation and Development, Program for International Student Assessment (PISA) 2003 Mathematical Literacy Assessment.

To compare the cognitive demand placed on students by the items of different formats, figure 7 shows the percentage of multiple-choice and constructed-response items in each assessment that were classified at the low, moderate, or high *mathematical complexity* level. NAEP and TIMSS show similar profiles, with multiple-choice items being composed primarily of low complexity level items (72 percent in NAEP and 63 percent in TIMSS) and essentially no high complexity level items. The complexity profiles for constructed-response items are also quite similar between NAEP and TIMSS, with about half of items classified at the moderate complexity level and 8 percent of TIMSS items and 10 percent of NAEP items classified at the high complexity level.

Once again, PISA stands alone in terms of the *mathematical complexity* level of its multiple-choice items. In PISA, the multiple-choice items (including both the more traditional multiple-choice items as well as the complex multiple-choice items) are predominantly moderate complexity level (71 percent) compared to about one-third in NAEP and TIMSS. In addition, this type of item format in PISA also includes several high complexity level items (14 percent), which were not found for NAEP or TIMSS. In fact, in PISA a larger proportion of multiple-choice items than constructed-response items are at the high complexity level. This reflects the fact that none of the short-response or closed constructed-response items were classified at the high complexity level. Among constructed-response items, the profile for PISA is similar to that for NAEP and TIMSS, but there are still a higher proportion of items at the moderate complexity level than found in NAEP and TIMSS (60 percent compared to about half in NAEP and TIMSS).

Figure 7. Percentage distribution of NAEP, TIMSS, and PISA mathematics items across levels of mathematical complexity, by item format: 2003



Rounds to zero.
NOTE: Data reflect expert panel classifications of *mathematical complexity* level according to definitions in the NAEP 2005 framework. Data for NAEP and TIMSS reflect both fourth- and eighth-grade items combined. The graphic for NAEP does not show one multiple-choice item classified at the high complexity level (< 0.5 percent). Detail may not sum to totals because of rounding.
SOURCE: U.S. Department of Education, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2003 Mathematics Assessment; International Association for the Evaluation of Educational Achievement, Trends in International Mathematics and Science Study (TIMSS) 2003 Assessment; Organization for Economic Cooperation and Development, Program for International Student Assessment (PISA) 2003 Mathematical Literacy Assessment; and U.S. Department of Education, National Assessment Governing Board, *Mathematics Framework for the 2005 National Assessment of Educational Progress*, 2004.

This section compared the assessments overall with respect to content coverage, grade level, levels of mathematic complexity, and item format. In the next section, more detailed comparisons of the content of the NAEP and TIMSS assessments are made in each of the main content areas of *number*, *measurement*, *geometry*, *data*, and *algebra*.