Comparing PIRLS and PISA with NAEP in Reading, Mathematics, and Science

The purpose of this document is to provide background information that will be useful in interpreting the results from two key international assessments that are being released in November and December 2007 and in comparing these results with recent findings from the U.S. National Assessment of Educational Progress in similar subjects.

Background

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

NAEP measures fourth-, eighth-, and twelfth-grade students' performance in reading, mathematics and science, with assessments designed specifically for national and state information needs. Alternatively, the international assessments enable the United States to benchmark its performance to that of other countries—in fourth-grade reading literacy in PIRLS and in 15-year-old students' reading, mathematics and science literacy in PISA.[2] All three assessments are conducted regularly to allow the monitoring of student outcomes over time. [3]

While the international assessments may appear to have significant similarities with our national assessment program, such as the content areas studied or the age or grade of students, each was designed to serve a different purpose and each is based on a separate and unique framework and set of assessment items (or questions). Thus, not surprisingly, there may be differences in results for a given year or in trend estimates among the studies, each giving a slightly different view of U.S. students' performance in these subjects.

NCES is releasing the results from the 2006 administration of PIRLS and the 2006 administration of PISA in November and December of 2007, respectively. Also available are results from 2007 for fourth- and eighth-grade reading and mathematics in NAEP and from 2005 for fourth-, eighth- and twelfth-grade science and twelfth-grade mathematics and reading in NAEP (see Table 1).[4] This document is intended to provide information that will help the press and others understand the results across studies, grasp the similarities and differences in these results, and identify what each assessment contributes to the overall knowledge base on student performance.

Table 1. Scope of this briefing paper

	4th grade	8th grade	15 year olds	12th grade
Reading	**NAEP 2007 PIRLS 2006**	NAEP 2007	PISA 2006	NAEP 2005
Mathematics	NAEP 2007	**NAEP 2007**	**PISA 2006**	**NAEP 2005**
Science	NAEP 2005	**NAEP 2005**	**PISA 2006**	**NAEP 2005**

Comparing Features of the Assessments

PIRLS, PISA, and NAEP differ from one another on several key features, including purpose, partners, population, precision of estimates, and content.

Purpose and proximity to curriculum

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

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

PIRLS is the U.S. source for internationally comparative information on the reading achievement of students in the fourth grade and on related contextual aspects such as reading curricula and classroom practices across countries. The PIRLS framework and specifications are developed in a collaborative process involving international reading experts, as well as the national research coordinators from each participating country, and thus reflect recent developments and consensus in the international research community and the interests of a wide range of countries.

PISA is the U.S. source for internationally comparative information on the reading, mathematics and science literacy of students in the upper grades at an age that, for most countries, is near the end of compulsory schooling. The objective of PISA is to measure the "yield" of education

systems, or what skills and competencies students have acquired and can apply in these subjects to real-world contexts by age 15. The literacy concept emphasizes the mastery of processes, understanding of concepts, and application of knowledge and functioning in various situations within domains. By focusing on literacy, PISA draws not only from school curricula but also from learning that may occur outside of school.

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

Partners

The international assessments provide benchmarks with different groups of countries.

The PIRLS assessment and the PISA assessment differ in country composition. The sponsorship for PIRLS is the International Association for Evaluation of Educational Achievement (IEA), which includes in its assessments a diverse group of countries and jurisdictions. The Organization for Economic Cooperation and Development (OECD) sponsors PISA, with its 30 member countries representing the world's most industrialized nations.

Thirty-six countries and 9 jurisdictions within countries participated in PIRLS and are included individually in the country rankings of student performance (see Table 2). The 36 countries include 16 OECD countries and the jurisdictions include 5 Canadian provinces, England and Scotland (within the United Kingdom), and the Flemish and French communities of Belgium. This means that in PIRLS, in some cases, the United States is being compared not just with other countries but with jurisdictions within countries. In PISA, scores are reported only at the national level and a total of 57 countries participated in the 2006 administration, including all 30 OECD countries.[5] Students from 30 countries participated in both assessments. Also, the international average in PIRLS is based on all participating countries and jurisdictions; in PISA it is based only on the OECD countries' scores. Therefore, comparisons to the international averages in PISA and PIRLS involve comparisons with different sets of countries.

Table 2. Countries participation in PIRLS and PISA (2006)

	Both PIRLS and PISA	PIRLS only	PISA only
OECD countries	Austria Belgium (as a single entity in PISA, as Flemish and French communities in PIRLS) Canada (as a single entity in PISA, as 5 individual provinces in PIRLS) Denmark France Germany Hungary Iceland Italy Luxemburg Netherlands New Zealand Norway Poland Slovak Republic Spain Sweden United Kingdom (as a single entity in PISA, as England and Scotland in PIRLS) United States	--	Australia Czech Republic Finland Greece Ireland Japan Korea Mexico Portugal Switzerland Turkey
Other countries	Bulgaria Chinese Taipei Hong Kong Indonesia Israel Latvia Lithuania Qatar Romania Russian Federation Slovenia	Georgia Iran Kuwait Macedonia Moldova Morocco Singapore South Africa Trinidad and Tobago	Argentina Azerbaijan Brazil Chile Colombia Croatia Estonia Jordan Kyrgyz Republic Macao-China Republic of Montenegro Republic of Serbia Thailand Tunisia Uruguay

Population

The students being studied may represent different groups.

NAEP, PIRLS, and PISA are all sample-based assessments—meaning that each program administers the assessment to a subgroup of U.S. students in such a way that the results can be generalized to the larger population. However, each assessment defines the population to which it is generalizing (and thus from which the sample is drawn) differently. One distinction between

NAEP and PIRLS, on the one hand, and PISA, on the other hand, is that the former use grade-based samples while PISA uses an age-based sample. These choices relate to the purposes of each program described earlier—NAEP and PIRLS to report on curricular achievement and PISA to describe the yield of systems toward the end of compulsory schooling.

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

- The PIRLS target population is all students in the grade corresponding to the fourth year of school, excluding kindergarten. For the United States this is fourth grade. Thus, the most recent PIRLS results reflect the performance of U.S. fourth-graders in 2006.

- The PISA target population is all 15-year-old students. Operationally in 2006, this included all students who were from 15 years and 3 months to 16 years and 2 months at the beginning of the testing period and who were enrolled in school, regardless of grade level or full- or part-time status. The most recent PISA results reflect the performance of U.S. 15-year-olds, who were in ninth, tenth, or sometimes another grade in 2006.

Thus for elementary grade reading, the most recent NAEP and PIRLS results are reporting on similar populations though in different academic years—NAEP with 2007 fourth-graders and PIRLS with 2006 fourth-graders. In the upper grades, the PISA population is uniformly older than NAEP eighth-graders and uniformly younger than the NAEP twelfth-graders. NAEP and PISA also are assessing different cohorts in different years. Taking this into account, perhaps the closest NAEP-PISA comparisons can be made between the NAEP 2005 eighth-grade and PISA 2006 15-year-old student cohorts, some of the former of whom theoretically could have been part of the latter. However, all side-by-side comparisons of NAEP and PISA results should be viewed with these population and cohort differences in mind.

Precision of estimates

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

NAEP, PIRLS, and PISA are all designed to provide valid and reliable measures of U.S. students' performance in the aggregate and for major subpopulations, and each study draws a sample sufficient for this purpose. NAEP and PIRLS and PISA differ, however, in the size of the differences in performance they are intended to detect. Student performance varies widely across countries and so PIRLS and PISA are designed to detect relatively large differences. NAEP is designed to detect smaller differences. This reflects smaller variations in student performance within the U.S. than across the many countries participating in PIRLS and PISA, as well as smaller variations in performance over time. It is important for NAEP to be sensitive to small changes in student performance over time, for the nation as a whole, and for individual states.

Sample sizes are calculated to balance needs for precision of estimates against burden to respondents. Because of NAEP's relatively higher need for precision, NAEP samples many more students than does PIRLS or PISA (see Table 3).

Table 3. Sample sizes in NAEP, PIRLS, and PISA

	No. of students sampled	No. of schools sampled
NAEP 2007 (4th grade)	191,000	7,830
NAEP 2007 (8th grade)	160,700	6,930
PIRLS 2006	5,190	183
PISA 2006	5,611	166

Content

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

As noted before, the assessments under discussion here are developed from frameworks that define the domain and specify the content and skills to be measured. Thus, a first task in comparing assessment programs is to compare how the frameworks and specifications are elaborated. A second task, which can provide a more in-depth view is to compare how the frameworks are operationalized through the actual assessment items and, in the case of reading, the passages in PIRLS and NAEP.

Fourth-grade reading: PIRLS 2006 and NAEP 2007

To date, there have been two studies undertaken to compare NAEP and PIRLS in the ways mentioned above. The first study compared NAEP 2002 and PIRLS 2001 at both the framework and item levels and was documented in an NCES technical report.[6] The second study updates with analysis of the passages and item sets added in NAEP 2007 and PIRLS 2006.[7] These studies form the basis of the text that follows.

Definitions and organization

In terms of how the domain is defined, there is considerable overlap between the NAEP and PIRLS concepts of reading literacy. For example, the frameworks for both assessments: (1) identify reading as a constructive process that involves interaction between the reader and the text; (2) suggest that the context for reading is an important element in how readers make meaning of text and in the skills and strategies they select; and (3) note that the structural elements of text influence readers' strategies. The differences are relatively minor: the PIRLS framework is more explicit about its targeting to young readers and acknowledges a more diverse set of reading contexts such as for personal enjoyment (versus the NAEP framework, which focuses more on school-based reading and is intended to be generally applicable across younger to older grades).

In terms of the organization of the frameworks, both NAEP and PIRLS are organized around two-dimensional matrices, which specify *processes* (i.e., the cognitive element) and the *purposes* or contexts for which students read. In particular, there are some notable differences at the framework level in how the processes (called aspects in NAEP) are broken out and elaborated. NAEP's four categories include: forming a general understanding, developing an interpretation, making reader-text connections, and examining content and structure. PIRLS' four categories include: locating and retrieving explicitly stated information, making straightforward inferences, interpreting and integrating ideas and information, and examining and evaluating content,

language and textual elements. The key areas of difference are that there is no apparent counterpart in the NAEP framework to the PIRLS locate and retrieve category, and there is no explicit counterpart in the PIRLS framework to the NAEP category that requires readers to think beyond the text and apply it to the real world (i.e., make reader-text connections). This suggests that there may be certain NAEP and PIRLS items that are unique to the respective programs.

In terms of the purposes for which students read, both frameworks specify a literary purpose and an information-related purpose. While the literary purposes seem to be defined in a similar way across the assessments, the information-related purposes suggest slight differences. PIRLS assesses not just reading to *acquire* information, but also to *use* information, in a way that goes beyond NAEP's definition. At the older grades, the NAEP framework includes a "reading to perform a task" purpose, which focuses on reading to learn how to do something, which is more similar to the use information aspect of PIRLS' "reading to acquire and use information purpose.

Passage and item analyses

The types of passages included in NAEP and PIRLS reflect the purposes that are assessed. In NAEP, students are presented with short stories, legends, biographies, and folktales, as well as magazine articles that focus on people, places, and events of interest to children—to cover both its literary experience and information purposes. Similarly, PIRLS also presents narrative fiction, usually in the form of short stories, as well as informational articles and, distinct from NAEP, brochures to cover its two similar purposes. Both NAEP and PIRLS strive to be "authentic" in that they try to present passages and items that would be encountered in and out of school. NAEP specifically calls for the use of authentic texts, and all passages are shown as previously published and generally are not edited at all (in terms of content or formatting) for use in NAEP. PIRLS also strives to use previously published texts, but has a more liberal policy on editing and changing the format of the texts used—which is sometimes necessary in an international context in order to meet constraints of translation to multiple languages and for culturally diverse participants. U.S. experts who have examined the PIRLS passages have noted the more edited, and sometimes less continuous, nature of some of these than the NAEP passages, particularly among passages for information purpose.

Altogether, the NAEP and PIRLS fourth-grade assessments each include 10 reading passages, although each student receives only a subset of those passages. In terms of length, the PIRLS passages tend to be shorter than the NAEP passages, averaging 707 words per passage compared to NAEP's 823 words per passage. The PIRLS passages range from 403 to 855 words; NAEP passages range from 644 to 1,361 words.[8]

Table 4. Results of readability analyses of PIRLS 2006 and NAEP 2007 passages

	PIRLS	NAEP
No. of sentences / 100 words	8.16	7.16
No. of syllables / 100 words	132.0	132.6
Fry average age	10.30	11.07
Fry average grade level	5th	6th
Flesch average reading ease	Easy (82.3)	Fairly Easy (79.7)
Flesch average grade level	5th-6th	7th
Lexile score	819.0	936.7
Corresponding grade level	4th-5th	6th-7th

Readability analyses also suggest that the PIRLS passages may be slightly easier than NAEP (see Table 4). On a very simple measure, for example, sentence counts show that the PIRLS passages, with a higher number of sentences per 100 word sample, consist of shorter sentences on average than do the NAEP passages. On other more elaborate measures, such as Fry and Flesch analyses, which use sentence count along with syllable count to determine a corresponding age and grade level for each text, PIRLS passages are calculated to be about one grade level below the NAEP passages. Finally, a Lexile measure, which indicates the reading demand of the text in terms of semantic difficulty (vocabulary) and syntactic complexity (sentence length) and which is more recently developed and normed than the other measures, also suggests that the PIRLS passages are suitable for one to two grades below those from NAEP. It should be noted, however, that both assessments do include a range of passages suited below and above the targeted grade level to capture the range of reading ability.

Each of these passages of course has items associated with it—approximately 12-13 per passage in PIRLS and 10 per passage in NAEP. Mapping the PIRLS items onto NAEP's cognitive processes, or aspects, and comparing these classifications with those for the NAEP items confirms some of the similarities and differences suggested by the forgoing framework analysis. The two assessments are similar in that the majority of items on both assessments require students to develop an interpretation about what they have read, although there is a greater emphasis on this in NAEP, with 69 percent of items classified as such compared to 60 percent of the PIRLS items. PIRLS also has a notably smaller percentage of items classified as forming a general understanding or making reader text connections, having half or less the percentage NAEP has in those categories. One of the major differences between the two assessments, however, is that there are a number of PIRLS items (21 percent) that do not fit on the NAEP framework at all. In nearly all cases, these are items that ask the reader to retrieve explicitly stated information, which is not a skill delineated in the NAEP framework or found in its items.

Upper grades science: PISA 2006 and NAEP 2005

There has not yet been an extensive study comparing how PISA and NAEP each define and assess science for the upper grades. However, we can look to the respective frameworks for some insight as to similarities and differences. By necessity, we examine NAEP's framework for the 2005 assessment, even though the next administration in 2009 will be grounded in a revised science framework. The PISA framework examined here is that which was elaborated for the most recent assessment of scientific literacy in 2006.

In these documents, both PISA and NAEP emphasize that science extends beyond knowledge of scientific facts to include broad understanding of science concepts and knowledge of how to apply and use scientific concepts and skills. There is recognition in both documents that it is important for students to demonstrate a knowledge *about science itself*—what distinguishes science from other ways of knowing (or understanding what it can and cannot answer), how to approach issues and evaluate information scientifically, and what is its role and impact and interaction with man and society. However, each framework is organized differently and therefore, until there is a more comprehensive study of items, it will be difficult to say how similar or different the actual assessments might be.

The science framework for NAEP 2005 is organized around a matrix, with content and cognitive dimensions, as well as two overarching dimensions. The content dimension includes the *fields of science* (life, physical, and Earth sciences) in which students demonstrate their cognitive skills of conceptual understanding, scientific investigation, and practical reasoning. The overarching dimensions—meaning that a certain number of items within the assessment will also meet additional characteristics—include items on the *nature of science* (such as understanding the scientific process, the interaction of man with the world, the role of technology in science) and on interdisciplinary *themes*, such as the idea of scientific models, the notion of systems, and patterns of change. The NAEP framework also designates a number of items to be "hands-on" tasks involving the use of materials to conduct scientific investigations, which represents an item format unique to the NAEP assessment and one relevant to science instruction in the United States.

The PISA framework also has content and cognitive dimensions, although as an immediate organizational difference, it seems that they are broader, capturing some of what likely is included in NAEP's separately titled overarching dimensions. PISA's content dimension, for example includes both *knowledge of the natural world* (in the fields of life systems, physical systems, Earth and space systems, and technology systems) and *knowledge about science* (scientific inquiry and scientific explanations). PISA's *competencies* also are specified somewhat differently and seem to be centered around an explicit dissection of scientific inquiry: identifying scientific issues, explaining scientific phenomena, and using science evidence. In the PISA model, the competencies are prominent—they form the subscales for reporting, for instance—and are shown as influenced by the content/knowledge dimension. The PISA framework also is explicit about the situationally based nature of science literacy and thus has a *context* dimension that describes a range of situations in which individuals deploy their competencies. It also puts an *attitudinal* dimension alongside the content/knowledge dimension and embeds attitudinal items (normally placed in a background questionnaire) within the actual assessment. Like students' knowledge, their attitudes (interest and motivation, sense of responsibility, and support for inquiry) are seen in the PISA model as influences on competencies. Although the responses to the attitudinal items are not part of the PISA score—they are reported separately—their presence in the assessment is a feature unique to PISA.

Given how differently the frameworks are organized, comparing intended item distributions is difficult. Looking at the content/knowledge dimension, NAEP calls for a roughly even distribution of items across the fields of science (physical, life, Earth), with a slightly heavier emphasis on life science at the eighth-grade level. PISA, on the other hand, divides its knowledge dimension into more categories and thus the three fields of science it has in common with NAEP each represent no more than a quarter of the items. While NAEP items may have additional classifications (such as being a "nature of science" or "themes" item), each one ties back to a specific scientific discipline, whereas the PISA framework suggests that it has some items that are

solely on knowledge *about* science. Looking at the cognitive dimension, NAEP emphasizes conceptual understanding, with nearly half of all items meeting this definition. PISA aims for a slightly more even distribution across its competency clusters. But again, because the categories are quite different, comparisons are limited.

The additional elements of the PISA science framework and other apparent differences likely reflect its overall purpose to study the application of knowledge and skill in real-world contexts in which individuals are interacting and participating in society. On the other hand, the organization of the NAEP framework and its relatively more content-oriented nature is by definition grounded in assessing the school-based learning of students in the context of U.S. standards and instruction.

Upper grades mathematics: PISA 2006 and NAEP 2005/7

At first glance, there are some noticeable similarities in the structure of the NAEP and PISA mathematics frameworks, as summarized in an earlier comparison study.[9] For instance—although differently titled—both NAEP and PISA are organized (primarily) along a content dimension and a cognitive dimension. In NAEP, these are the *content strands* and *levels of mathematical complexity*, respectively, and in PISA, these are the *overarching ideas* and *competency clusters*. However, the manner in which the content and cognitive dimensions are further specified within these dimensions differs between the two—reflecting NAEP's close ties to the organizational structures used in traditional school curricula and, by contrast, PISA's focus on the application of mathematics in real-world situations. Thus, the PISA framework also includes a third dimension, which is the *situation* or *context* (e.g., educational, personal, …) of an item.

NAEP's *content strands* include five major areas of mathematics, including number sense, properties, and operations; measurement; geometry and spatial sense; data analysis, statistics, and probability; and algebra and functions—within which specific topics and subtopics are further identified. In PISA, the content areas are described in terms of *overarching ideas*, which include change and relationships; quantity; space and shape; and uncertainty. NAEP's *levels of mathematical complexity* (denoted as low, moderate and high) identify where along a continuum of cognitive demand an item falls, with skills like the ability to perform a stated procedure at one end and the ability to engage in abstract reasoning or generalize a pattern at the other end. PISA's *competency clusters*, while not an explicit or exact hierarchy, similarly denote three sets of gradually increasing cognitive demand, from reproduction to connection to reflection.

These structural and terminological differences suggest that there may be differences in the ways in which NAEP and PISA are operationalized and, indeed, the previous comparison study of NAEP and PISA mathematics items has shed light on the degree to which this is true.[10] In terms of content similarities, the past study showed that there is a considerable overlap between PISA's uncertainty overarching idea and NAEP's data analysis, statistics, and probability content strand, as well as the space and space idea and the geometry and spatial sense strands. In terms of content differences, however, PISA was shown to have a relatively greater focus on data analysis, statistics, and probability and lesser focus on algebra than the NAEP eighth-grade assessment.[11]

In terms of the cognitive dimensions, PISA had more items requiring "reflection," which includes less familiar or more complex problem settings and a higher demand for thinking and reasoning and developing and communicating an argument, than did NAEP at both eighth- and twelfth-grade. NAEP, on the other hand, had more items that fell into the reproduction category, which includes reproducing practiced material and performing routine operations.[12] However, while the

past study concluded that PISA items classified to a higher level of cognitive complexity and demand than NAEP items, the content covered was most consistent with the topics specified in the NAEP eighth-grade framework.

Finally, it is important to note the differing distributions of items of a particular format in the two programs—with NAEP aiming for a roughly even split between multiple-choice and constructed response (or, open-ended) items and PISA aiming for a greater emphasis on constructed response with about two-thirds of items of such a nature. While item type is not necessarily directly related to item difficulty, students' ease or familiarity with different item types may systematically differ and thus contribute to any observed differences in results.

Examining Results in the Context of the Distinctions among the Assessments

Both PIRLS and NAEP provide a measure of fourth-grade reading, and both PISA and NAEP provide measures of mathematics and science performance for older students (in grades 8 to 12). It is natural to compare them, but the distinctions described previously need to be kept in mind in understanding the converging or diverging results.

Comparing select results for fourth-grade reading

The most recent results from PIRLS and NAEP include information on trends over time in fourth-grade reading: in PIRLS between 2006 and 2001 and in NAEP between 2007 and several earlier time points going back to 1992. Here we describe the NAEP 2002 to 2007 period, since it provides a similar time interval to PIRLS.

PIRLS shows that statistically there is no change in U.S. fourth-grade students' average scores from 2001 to 2006. This contrasts with NAEP results for 2007, which show an upward tick (by 2 score points) in fourth-grade reading scores from 2002, all of which occurred since 2005. However, although the populations in PIRLS and NAEP are the same, as the previous sections highlighted, there are some differences in the nature of the reading passages and in the reading skills being measured, with about one-fifth of the PIRLS items not corresponding well to the NAEP framework. Additionally, because NAEP uses a much larger sample size, it is more sensitive to picking up small changes over short periods of time than is PIRLS, which is not designed primarily for that purpose but for detecting differences among countries.

Comparing select results for upper grades science

It is more difficult to compare the results from PISA and NAEP in science at the upper grades, not only for the population and framework differences, but also because PISA is not yet reporting a trend measure for science, which would be the most likely element to examine in the context of NAEP's trend measure. The last assessment of science in NAEP (2005), showed no statistically significant differences in the performance of eighth- or twelfth-graders since 2000 (although there was a slight decrease since 1996 among the older students).

Comparing select results for upper grades mathematics

In mathematics at the upper grades, PISA 2006 shows that statistically there is no change in the scores of U.S. 15-year-olds since 2003. On PISA 2006, the U.S. score for mathematics literacy is below the average for all OECD countries. As with fourth-grade reading, NAEP shows an increase in the scores of eighth-grade mathematics students between 2003 and 2007. Again, NAEP's design allows it to pick up small changes in the performance of U.S. students.

Summary

In sum, there appears to be an advantage in capitalizing on the complementary information presented in national and international assessments. NAEP measures in detail the reading, mathematics and science knowledge of U.S. students as a whole, and can also provide trend information for individual states, different geographic regions, and demographic population groups. International assessments like PIRLS and PISA add value by providing a method for comparing our performance in the United States to the performance of students in other nations. However, their differences need to be recognized when interpreting results. Some of the differences between NAEP, PIRLS, and PISA include:

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

- The two international assessments provide benchmarks with different groups of countries. Thirty-nine countries participated in PIRLS, 16 of which are the industrialized OECD countries. Fifty-seven countries participated in PISA, 30 of which are OECD countries. Thirty countries participated in both studies. One key difference between the two studies, however, is that the international average in PIRLS is based on all participating countries whereas in PISA it is based on OECD countries only.

- The students being studied represent different groups. Both NAEP and PIRLS use grade-based samples and both target fourth-grade students. However, the last NAEP assessment in fourth-grade reading was in 2007, whereas for PIRLS it was in 2006, so the results do not generalize to the same group of students. PISA uses an age-based sample, which targets 15-year-olds, who likely are between the NAEP target populations of eighth- and twelfth-graders.

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

- The reading skills, mathematics, and science being assessed can be different in terms of the ways in which the frameworks for assessment are organized and in terms of content coverage, item format, and other key features. Examinations of the frameworks for NAEP, PIRLS, and PISA in reading, mathematics, and science show areas of potential overlap and potential difference in terms of the content and skills being measured in the respective subject areas and grades. Further, additional analyses of the fourth-grade reading passages and items show that (1) PIRLS passages are slightly shorter and slightly easier than NAEP fourth-grade passages, and (2) PIRLS appears to have a subset of items that are distinct from the types of items found in NAEP.

Suggested Citation

Stephens, M., and Coleman, M. (2007). *Comparing PIRLS and PISA with NAEP in Reading, Mathematics, and Science* (Working Paper). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Available at: http://nces.ed.gov/Surveys/PISA/pdf/comppaper12082004.pdf

Contact Information

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

Useful Websites

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

[1] PIRLS is conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA). PISA is sponsored by the Organization for Economic Cooperation and Development (OECD). The United States also participates in the Trends in International Mathematics and Science Study (TIMSS), conducted under the auspices of the IEA. See "Comparing NAEP, TIMSS, and PISA in Mathematics and Science," available at http://nces.ed.gov/Surveys/PISA/pdf/comppaper12082004.pdf, for a comparison of the most recent TIMSS (TIMSS 2003) and NAEP.

[2] The 2006 PISA reading literacy assessment is not included in the comparisons in this paper, however, because the results will not be reported for the United States. Because of a formatting error in the testing booklets and the small number of reading literacy items in the 2006 administration, the scores could not be recalibrated to exclude the affected items.

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

[4] See:

Baer, J., Baldí, S., Ayotte, K., Green, P., and McGrath, D. (2007). *The Reading Literacy of U.S. Fourth-Grade Students in an International Context: Results from the 2001 and 2006 Progress in Reading Literacy Study (PIRLS)* (NCES 2008-017). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

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

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

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

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

Lee, J., Grigg, W.S., and Donahue, P.L. (2007). *The Nation's Report Card: Reading 2007* (NCES 2007-496). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

[5] For example, in PISA, while Belgium administers the test in both Flemish and French, for reporting purposes, results are combined together.

[6] Binkley, M., and Kelly, D.L. (2003). *A Content Comparison of the NAEP and PIRLS Fourth-Grade Reading Assessments* (Working Paper No. 2003:10). U.S. Department of Education. Washington, DC: National Center for Education Statistics. This provides the information in this briefing paper on the theoretical comparisons of the fourth-grade reading definitions and frameworks, which is valid because the frameworks for each study have remained essentially the same.

[7] This study involved a five-person expert panel to externally verify the continuity of the NAEP 2002 and 2007 frameworks and to classify the new PIRLS items to the NAEP framework. This occurred in September 2007. The classifications for the new PIRLS items were combined with the classifications from the first study for the items used in both 2001 and 2006 to obtain a complete data set that could be compared with the NAEP assessment specifications. This informs the section on the item-level comparisons.

[8] This quantitative information in this paragraph and the two that follow was calculated as part of the passage and item analyses undertaken for this briefing paper.

[9] The information in this section is taken from the technical report on a comparison study that was undertaken for the NAEP 2003 and PISA 2003 mathematics assessments. See: Neidorf, T.S., Binkley, M., Gattis, K., and Nohara, D. (2005). *A Content Comparison of the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Mathematics Assessments* (NCES 2006-029). U.S. Department of Education. Washington, DC: National Center for Education Statistics. This source is used because there have been no changes to the frameworks (that were not already accounted for in the study) between the 2003 and most recent administrations of the two assessments.

[10] It should be noted that the percentages reported in the endnotes for this section are for the sets of items used in the 2003 administrations of NAEP and PISA; item-level comparisons were not redone with the NAEP 2005/7 and PISA 2006 items. Because the PISA 2006 items are a subset of the 2003 items and because the newer NAEP items will include some items repeated from 2005/07, as well as newer items designed to replicate the items

being replaced, the Neidorf et al. paper, the general conclusions remain valid. The percentages should be taken as illustrative rather than exact, however.

[11] Thirty-nine percent of PISA items were classified to the NAEP data category, compared with 10 percent of eighth-grade and 25 percent of twelfth-grade items in NAEP. In contrast, 9 percent of PISA items were classified as algebra, compared with 15 percent of eighth-grade and 35 percent of twelfth-grade items in NAEP.

[12] Thirty-one percent of PISA items were classified to the reproduction competency cluster, compared with 58 and 39 percent of NAEP eighth- and twelfth-grade problem solving items, respectively. On the other hand, 22 percent of PISA's items were deemed to fit the reflection competency cluster, compared to 5 percent or less of NAEP items at both grades.