National Assessment Governing Board

Assessment Development Committee

November 17–18, 2016

AGENDA

Thursday, November 17		
12:30 – 4:00 pm	**Closed Session (12:30 – 2:30 pm)** Welcome and Introductions *Shannon Garrison, Chair* Review of NAEP Items in Mathematics and Science **Open Session (2:30 – 4:00 pm)** Review of NAEP Contextual Questions in Mathematics and Reading	Secure material provided under separate cover
Friday, November 18		
10:15 – 11:30 am	**Joint Session with the Committee on Standards, Design and Methodology (COSDAM)**	
10:15 – 10:20 am	Welcome and Session Overview *Andrew Ho, COSDAM Chair* *Shannon Garrison, ADC Chair*	
10:20 – 10:40 am	Background Information • Overview of Models for Framework Development and Update Processes *Mary Crovo, Governing Board Staff* • Dynamic Frameworks in the Future of NAEP *Dan McGrath and Dana Kelly, NCES*	Attachment A Attachment B
10:40 – 11:00 am	ADC and COSDAM Discussion/Q&A	
11:00 – 11:30 am	**Closed Joint Session** Alignment Between NAEP Math and Common Core State Standards at Grades 4 and 8 *Enis Dogan, NCES*	Attachment C

National Assessment Governing Board

Assessment Development Committee

November 18, 2016

AGENDA

Friday, November 18 (Continued)		
11:30 – 11:40 am	BREAK	
11:40 –11:45 am	Welcome and Introductions *Shannon Garrison, ADC Chair* ADC Staffing Announcement *Mary Crovo*	
11:45 am –12:30 pm	**Closed Session** History and Overview of NAEP Long-Term Trend Assessments in Reading and Mathematics *Mary Crovo, Governing Board Staff* *Andy Kolstad, NCES Consultant* *Eunice Greer, NCES* *Elvira Germino Hausken, NCES*	Attachment D
	Information Item: Item Review Schedule	Attachment E

Models for Framework Development and Update Processes

Overview

This joint ADC/COSDAM briefing and discussion will provide information on NAEP framework development processes, NCES' Future of NAEP recommendations on dynamic frameworks, and related activities from the international assessment arena.

According to the NAEP statute (P.L. 107-279), the Board is responsible for developing assessment objectives and test specifications for each NAEP subject area. Since 1989 the Governing Board has developed assessment frameworks and specifications in more than 10 subjects through comprehensive, inclusive, and deliberative framework projects. The Board's Framework Development Policy can be found here.

Three models have been used in the Board's framework development process over time:

1. New Framework/Start New Trend

In some cases, the Board has determined through research, outreach, content and policy input, and other means that a new framework is warranted in a subject area. In these subject area assessments, the new assessment framework defines a new construct, includes different content and skills, adds new item types, changes the assessment delivery mode (i.e., DBA), and other modifications. Examples of this model include 2009 Science and 2011 Writing. In these cases, the trend line was broken and results cannot be compared to previous years.

2. New Framework/Maintain Trend

In this model, the new framework is designed to be different in many ways from the previous framework; however, empirical investigation reveals that the construct does not differ substantially. The interest in maintaining trend prompts linking studies and other research to try to ensure trend lines can be maintained. The 2009 Reading Framework is an example, which resulted in trend remaining intact from 1992.

3. Updated Framework/Maintain Trend

This model is defined by gradual changes to a framework over time so that trend is maintained. For mathematics, the framework has been "tweaked" over time to more clearly define the objectives, shift content emphases, and refine the process dimension while not redefining the construct. NAEP has been able to maintain the mathematics trend line for grades 4 and 8 since 1990.

The Board's Strategic Vision, scheduled for action at the November 2016 quarterly meeting, includes the statement:

- Develop new approaches to update NAEP subject area frameworks to support the Board's responsibility to measure evolving expectations for students, while maintaining rigorous methods that support reporting student achievement trends.

The November 18th COSDAM and ADC discussion will provide the groundwork for further activities to address this Strategic Vision priority. One major challenge will be determining how much framework content can be changed and how quickly that can occur, without compromising the ability to maintain trend.

MAY 2012

NAEP:
LOOKING AHEAD

LEADING ASSESSMENT INTO THE FUTURE

Recommendations to the Commissioner
National Center for Education Statistics

NAEP: Looking Ahead
Leading Assessment into the Future

NCES Initiative on the Future of NAEP

The National Assessment of Educational Progress (NAEP) has undergone a series of notable changes in the past decade. The NAEP program has expanded to meet new demands. All 50 states, the District of Columbia, the Department of Defense schools, and (on a trial basis) 21 urban districts are now participating in the mathematics and reading assessments at grades 4 and 8. In addition, thirteen states are participating in trial state 12th-grade assessments in reading and mathematics. NAEP is also reporting in record time to ensure that the findings are highly relevant upon release. Technology has taken on a bigger role in the development and administration of NAEP, including computer-based tasks in the science and writing assessments. These are just a few of the major developments; the program has grown and matured in almost all respects.

There is also growing interest in linking NAEP to international assessments so that NAEP scores can also show how our nation's students measure up to their peers globally. Additionally, there is increasing interest in broadening assessments in the subject areas to incorporate college and career readiness, as well as what are often called "21st-century skills" (communication, collaboration, and problem-solving).

The National Center for Education Statistics (NCES), which administers NAEP, is dedicated to moving the program forward with its upcoming procurement cycle which will take the program to 2017. Under the leadership of NCES Commissioner Jack Buckley, NCES convened a diverse group of experts in assessment, measurement, and technology for a summit in August 2011. These experts discussed and debated ideas for the future of NAEP. NCES convened a second summit of state and local stakeholders in January 2012. Participants at both gatherings were encouraged to "think big" about the role that NAEP should play in the decades ahead.

NCES assembled a panel of experts from the first summit, chaired by Edward Haertel, an expert in educational assessment, to consider and further develop the ideas from the two discussions and make recommendations on the role of NAEP in the future—10 years ahead and beyond. Based on summit deliberations and their own extensive expertise, the panel developed a high-level vision for the future of the NAEP program, as well as a plan for moving toward that vision.

This paper contains the panel's recommendations to the NCES Commissioner. NCES will consider these recommendations in their mid- and long-range planning for the program.

Panel Members

Edward Haertel (chair)
Jacks Family Professor of Education
School of Education
Stanford University

Russell Beauregard
Research Scientist & Director of Design
Education Market Platforms Group
Intel Corporation

Jere Confrey
Joseph D. Moore Distinguished University Professor
College of Education
North Carolina State University

Louis Gomez
MacArthur Chair in Digital Media and Learning
Graduate School of Education and Information Sciences
University of California, Los Angeles

Brian Gong
Executive Director
National Center for the Improvement of Educational Assessment

Andrew Ho
Assistant Professor
Graduate School of Education
Harvard University

Paul Horwitz
Senior Scientist and Director
Concord Consortium Modeling Center
Concord Consortium

Brian Junker
Professor
Department of Statistics
Carnegie Mellon University

Roy Pea
David Jacks Professor of Education
School of Education
Stanford University

Robert Rothman
Senior Fellow
Alliance for Excellent Education

Lorrie Shepard
Dean and Distinguished Professor
School of Education
University of Colorado, Boulder

3. NAEP's Assessment Frameworks and Learning Outcomes

3.1 Background and History

Assessment frameworks are conceptual, overview documents that lay out the basic structure and content of a domain of knowledge and thereby serve as a blueprint for assessment development. Typically, assessment frameworks, for NAEP and for other large-scale assessments, are constructed as two-dimensional matrices of content strands and cognitive processes. For example, the current NAEP mathematics framework includes five content areas: number properties and operations; measurement; geometry; algebra; and data analysis, statistics and probability. These are assessed at different levels of cognitive complexity, which include mathematical abilities such as conceptual understanding, procedural knowledge, and problem-solving. In geography, the content areas include: space and Earth places; environment and society; and spatial dynamics and connections. The levels of the cognitive dimension consist of knowing, understanding, and applying.

NAEP Assessment Frameworks are developed under the auspices of the Governing Board through an extensive process involving subject matter experts, who consider how research in the discipline and curricular reforms may have shifted the conceptualization of proficiency in a given knowledge domain. The development process also requires multiple rounds of reviews by educators, policy leaders, members of the public, and scholars. It is expected that assessment frameworks will need to be changed over time. However, the decision to develop new frameworks is approached with great caution because measuring change requires holding the instrument constant. Introducing new frameworks—while providing a more valid basis for the assessment—could threaten one core purpose of NAEP, which is to monitor "progress." In the past, when relatively minor changes have been made in assessment frameworks, as judged by content experts, trend comparisons over time have been continued and bridge validity studies have been conducted to verify that conclusions about gains have not been conflated with changes in the measuring instrument or redefinition of the construct being assessed.

When more profound changes occur in the conceptualization of an achievement domain, then a new framework is essential, and correspondingly the beginning of a new trend line. The adoption by nearly all states of the CCSS in English language arts and literacy and mathematics and the new Science Education Framework developed by the National Research Council (NRC) could be the occasion for a substantial enough change in conceptualization of these domains that new NAEP frameworks and new trend comparisons are warranted. Still, the future of NAEP—as a statistical indicator and as an exemplar of leading-edge assessment technology—requires great care and attention to the implications of new trend comparisons rather than merely acceding to the hoopla surrounding the new standards.

In the history of NAEP, few changes have been made in the assessment frameworks for reading and for mathematics. The old frameworks in these two core subjects, begun in 1971 and 1973 respectively, were replaced in the early 1990s, and then again in 2009 for reading. The old assessments have been continued on a less frequent cycle and are referred to as long-term trend NAEP. The 1990's mathematics framework and 2009 reading framework guide the present-day assessments, referred to as main NAEP. While NCES has been careful to insist that the old and new frameworks measure different things and therefore cannot be compared, the existence of the two trends provides a critically important example to illustrate how changing the measure can change interpretations about educational progress (e.g., see Beaton & Chromy, 2010). The earlier assessments focused much more on basic skills. Reading passages were generally shorter compared to today's NAEP and did not require students to demonstrate so wide a range of reading skills or answer extended-response questions. In mathematics, long-term trend NAEP had a greater proportion of computational questions and items asking for recall of definitions, and no problems where students had to show or explain their work. In a 2003 study, researcher Tom Loveless complained that the new NAEP mathematics assessment exaggerated progress in mathematics during the 1990s because gains on the basic skills test over the same period were much

smaller (when compared in standard deviation units of the respective tests). Because the two assessments are administered entirely separately, Loveless then had to rely on comparisons based on the less than satisfactory item-percent-correct metric to try to track progress in subdomains of the test. A more recent study using more sophisticated methods has largely confirmed his general conclusions, but that same study has highlighted the technical challenges of comparing trends for two assessments administered under such different conditions (Beaton & Chromy, 2010).

3.2 New Approaches for Assessment Frameworks

3.2.1 Designing frameworks and assessments to evaluate directly the effects of changing domain definitions

NAEP cannot be a research program and in particular cannot be structured to investigate the effectiveness of various instructional interventions. However, it can and should be attentive to the ways that shifting definitions of subject matter competence can affect claims about progress or lack of progress (cf. Section 3.2.3). In the CCSS context, it will be especially important to pay attention directly to potential differences between consortium-based conclusions and NAEP trends. Taking this on as a role for NAEP continues its important function as a kind of monitoring instrument. For example, when some state assessment results have shown remarkable achievement gains and closing of achievement gaps, achievement trends for the same states on NAEP have helped to identify inflated claims. These disparities might exist because of teaching-the-test practices on state tests (Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998), state content or achievement standards that do not rise to NAEP levels (Bandeira de Mello, Blankenship, & McLaughlin, 2009), exclusion of low-performing students on NAEP, or lower motivation on NAEP. More direct linking by carefully accounting for the consortium frameworks within new NAEP frameworks, would allow NAEP to act somewhat like an external monitor for CCSS assessment results. While the current NAEP frameworks do cover many of the same skills as the CCSS, they can be enhanced with some shifts in content.

"21st-century skills" aren't actually new in this century, but it is a relatively new idea (beginning in the 1990s) that these reasoning skills should be more broadly attained and expected of all students. More importantly, it is indeed new that policy leaders would move toward a view of learning that calls for reasoning and explaining one's thinking from the earliest grades, in contrast to outmoded theories of learning predominant in the 20th century that postponed thinking until after the "basics" had been mastered by rote. In addition, the CCSS firmly ground reasoning, problem-solving, and modeling in relation to specific content, not as nebulous generalized abilities. While there is widespread enthusiasm for designing new assessments that capture these more rigorous learning goals, we should note that promises like this have been made before. In the case of the current NAEP mathematics assessment, item developers acknowledge that the proportion of high complexity items actually surviving to the operational assessment is much smaller than is called for in the NAEP Mathematics Framework, and a validity study at both grades 4 and 8 found that the representation of high-complexity problems was seriously inadequate at grade 8, especially in the Algebra and Measurement strands (Daro, Stancavage, Ortega, DeStefano, & Linn, 2007).

Good intentions to measure "higher order thinking skills" are often undermined for three interrelated reasons. First, test questions at higher levels of cognitive complexity are inherently more difficult to develop. Because the dimensions of the task are intended to be ill-specified, such problems are often perceived to be ambiguous. But as soon as the item developer provides clarifying parameters, the challenge of the problem is diminished. Second, because "21st-century skills" involve applying one's knowledge in real world contexts, prior experience with particular contexts (or lack thereof) can create very large differences in performance simply because students unfamiliar with the context are unable to demonstrate the intended content and reasoning skills. In fact, application or generalization can only be defined in relation to what is known to have been taught. This is the curriculum problem that haunts large-scale assessments like NAEP that seek to be curriculum independent. Finally, well

designed items can fail on statistical criteria if too few students can do them.

These are all cautionary tales. They do not imply that NAEP should be less ambitious in developing new assessment frameworks that reach as far as possible in representing these higher levels of subject matter proficiency. But they do suggest a hedging-one's-bets approach that does not discard old frameworks wholesale in favor of the new. Rather, as mentioned previously, some conscious combination of old and new would create an assessment better equipped to track progress over time. Later we discuss Innovations Laboratory studies like those NAEP has used historically to explore the feasibility of new assessment strategies. However, we should emphasize that studies of innovative assessment strategies that tap complex skills should not merely be new assessment formats administered to random samples of students. Rather, in recognition of the fact that opportunities to learn particular content and skills may affect whether an assessment looks psychometrically sound, studies should be undertaken with carefully selected populations where relevant opportunities to learn can be established. This will help determine whether more advanced performance can be accurately documented to exist within the parameters of the new standards.

3.2.2 Standing subject-matter panels

To aid in this process, provide substantive oversight, and ensure meaningful interpretation of trends, we elaborate a recommendation for the future of NAEP previously made by a National Academy of Education Panel, which called for standing subject-matter committees. We recommend an expanded role whereby standing committees of subject matter specialists would review field test data, for example, and call attention to instances when after-the-fact distortions of the intended domain occur because more ambitious item types fail to meet statistical criteria. These committees would also have a role in ongoing incremental updates to content frameworks. They might include at least one member with psychometric expertise to aid in formulating technical specifications. The role of these committees is further described in Section 6.1.3.

3.2.3 Dynamic assessment frameworks and reporting scales

As just explained in Section 3.1, NAEP assessment frameworks have historically been held fixed for a period of years and then changed. It might be added that historically, NAEP item pools have been constructed according to test specifications derived from assessment frameworks. NAEP reporting scales, in turn, have reflected the resulting mix of NAEP items. Periodic small revisions to assessment frameworks have been made while maintaining trend lines; major breaks requiring new trend lines have occurred only rarely. With standing subject-matter panels, assessment frameworks for each subject-grade combination might be adjusted more frequently, defining a gradually changing mix of knowledge and skills, analogous to the Consumer Price Index (cf. Section 5.3). At the same time, item pools might be expanded somewhat, including everything in the assessment framework but also covering some additional material. Assessment frameworks would still define the intended construct underlying NAEP reporting scales, but not all items in the NAEP exercise pool would be included in the NAEP reporting scales. For example, content required to maintain long-term trend NAEP, to assure sufficient representation of the CCSS, or to improve the linkage to some other assessment could be introduced into the pool without affecting NAEP reporting scales. With somewhat broader exercise pools, alternative construct definitions could be investigated in special studies. The panel assumes that broader exercise pools, supporting modestly different construct definitions, will increase the value of NAEP by highlighting distinctions among achievement patterns under different construct definitions. Of course, there would still be one main NAEP reporting scale for each subject/grade combination. Clarity in communicating NAEP findings would remain a priority.

Different assessment frameworks may imply different definitions of the same broad subject area achievement construct (e.g., "reading" or "mathematics"), and achievement trends may differ depending on the construct definition chosen. Incremental changes in assessment frameworks and the corresponding set of items on which NAEP reporting scales were based would afford local (i.e., near-term) continuity in the meaning of those scales, but over a period of decades, constructs

might change substantially. This was seen by the panel as a potential strength, but also a potential risk. Policymakers and the public should be aware of how and when the construct NAEP defines as "reading," for example, is changed. Not every small, incremental change would need to be announced, but it would be important to establish and to enforce clear policies concerning the reporting of significant changes in assessment frameworks, so as to alert stakeholders when constructs change and to reinforce the crucially important message that not all tests with the same broad content label are measuring the same thing. As small content framework adjustments accumulate over time, standing committees, using empirical studies, would need to determine when the constructs measured have changed enough to require establishing new trend lines.

Dynamic frameworks would balance dual priorities of trend integrity and trend relevance. As an analogy, the Consumer Price Index (CPI) tracks inflation by deliberately conflating two concepts: change in the cost of a fixed basket of goods and change in the composition of the basket itself. As time passes, an increase in the cost of a product that is no longer relevant should contribute less to estimated inflation. By adopting dynamic frameworks, NAEP would similarly conflate increases in student proficiency with a change in the definition of proficiency itself. Although this conflation may seem undesirable, it may be the best way to balance desires for both an interpretable trend and a relevant trend.

3.2.4 Learning progressions as possible guides to assessment frameworks

Learning progressions or trajectories represent descriptions of how students' knowledge, skills, and beliefs about the domain evolve from naïve conceptions through gradual transformations to reach proficiency with target ideas at high levels of expertise over a period of years (Heritage, 2008). They entail the articulation of intermediate proficiency levels that students are likely to pass through, obstacles and misconceptions, and landmarks, of predictable importance as students' knowledge evolves over time. Empirical study of learning progressions highlights the key roles of instruction, use of tools, and peer interactions in supporting learning. Because the process of evolving understanding can take multiple years, learning progressions bridge formative and summative assessment.

A learning progression can provide much more information than a typical assessment framework. A learning progression ideally specifies both what is to be learned as well as how that learning can take place developmentally over time. It often integrates content and cognition. It includes not only the learning targets but also common less-than-ideal states that many students pass through. It is ordered developmentally. It provides a domain-based interpretation of development or growth that is useful to educators. The 2009 NAEP Science Framework already contains a section on learning progressions; however, learning progressions may offer guidance for the development of future NAEP assessment frameworks, especially in mathematics.

Learning progressions are closely entwined with instructional decisions regarding the sequencing of key concepts and skills. In the Netherlands, for example, the related constructions are referred to as "learning-teaching trajectories." However, few empirically supported "learning progressions" as yet exist, and developing more has proven challenging. In addition, because of NAEP's role as a curriculum-independent monitor, it may be more difficult to develop assessment frameworks that are entirely built as a collection of learning progressions. More likely some particular sequences, if proven to be valid across curricula, could be embedded within more general assessment frameworks.

Alignment between NAEP items and the CCSS and student performance in 2015 grade 4 and grade 8 Mathematics assessments

In 2015, Daro, Hughes, and Stancavage of the NAEP Validity Studies Panel conducted a study to evaluate the degree of alignment between 2015 NAEP grade 4 and grade 8 mathematics assessments and the CCSS in mathematics. They had a panel of experts classify these items into one of three categories: "in the standards at or below the NAEP grade level," "not in the standards at or below the NAEP grade level," and "uncertain." Seventy-nine percent of the grade 4 and 87% of the grade 8 items were classified as "in the standards". The degree of alignment was uneven across the subscales. At both grades, lowest level of alignment was observed in data analysis, statistics, and probability subscale with 47% and 74% alignment at grade 4 and 8, respectively.

In this study we use the classification of the items from the abovementioned study **to investigate the student performance in 2015 NAEP grade 4 and grade 8 mathematics assessments in relation to the alignment of the items to the CCSS.** The research questions are as follows:

1. Are there differences in student performance at the item level according to items' coverage in the CCSS?
2. Are there differences in psychometric properties of items according to items' coverage in the CCSS?
3. How would state mean scores change if items student achievement is estimated using only the items that are covered in the CCSS?

In relation to the first research question, we examined the changes in average p+ values for trend items at state level by item alignment. In addition, we computed an item residual for each item for each state based on the difficulty of the given item across states and based on the performance of the given state across all items. In answering the second research question, we first compared the estimates for the discrimination parameter between CCSS-aligned and other items. Next, we conducted differential item functioning (DIF) analyses to examine whether items function differently in CCSS states versus other states.

In order to answer the final research question, mean state scores were-recomputed based on only the items judged to be aligned to the CCSS. Dependent sample t-tests were run to compare the reported and re-estimated means for 2015 for each state separately, one scale at a time. We also investigated if the directionality (i.e. increase, no change, decrease) of the trend results between 2013 and 2015 would have changed with the re-estimated state means. Independent sample t-tests were conducted to compare the reported mean for 2013 to the 2015 reported and the 2015 re-estimated means for the composite scale and subscales for each state separately.

This session will be closed because the study results have not yet been released.

History and Overview of NAEP Long-Term Trend Assessments in Reading and Mathematics

Overview

As stated in the NAEP statute (P.L. 107-279), the Commissioner for Education Statistics shall "continue to conduct the trend assessment of academic achievement at ages 9, 13, and 17 for the purpose of maintaining data on long-term trends in reading and mathematics."

The Governing Board has been exploring issues related to NAEP Long-Term Trend (LTT) assessments for several decades. The Board's draft Strategic Vision, slated for action at this November 2016 Board meeting, includes a specific reference to the NAEP LTT calling for the Board to:

> Research policy and technical implications related to the future of NAEP Long-Term Trend in reading and mathematics.

The purpose of this closed ADC briefing and discussion is to familiarize the Committee with details of the LTT history, design, and content. The session will focus in particular on the LTT content, including secure reading and math test items. The content of the LTT assessments is an important consideration in the upcoming discussions on how to implement the Board's Strategic Vision for LTT. These discussions will include the Board's planned LTT symposium in March 2017. The ADC will be providing content guidance in these upcoming Board deliberations on the future of the LTT assessments.

Reference materials:

- Long-Term Trend history and next steps
- Table of Long-Term Trend assessments
- Comparison chart of Long-Term Trend and Main NAEP
- Executive summary from the 2012 Long-Term Trend report

Long-Term Trend Overview and Update

Background

NAEP includes two national assessment programs—Long-Term Trend (LTT) NAEP and Main NAEP. While both assessments enable NAEP to measure student progress over time, there are similarities and differences between the two assessments. Both assessments measure reading and mathematics. The NAEP LLT assessment measures national educational performance in the United States at ages 9, 13 and 17. In contrast, the Main NAEP assessments focus on populations of students defined by grade, rather than age, and go beyond the national level to provide results at the state and district level. LTT trend lines date back to the early 1970s and Main NAEP trend lines start in the early 1990s. The content differs as well—for example, LTT math measures more "traditional" mathematics than the current Main NAEP math content.

The Main NAEP assessments in reading and mathematics are administered every two years, as required by law. The administration of NAEP LTT assessments in reading and mathematics at ages 9, 13, and 17 is also required by law, but the periodicity is not specified. The NAEP LTT assessments had been administered approximately every four years over the past two decades (and more frequently prior to that), but were last administered in 2012. The Governing Board postponed the NAEP LTT planned administration for 2016 to 2020, and then to 2024 due to budgetary constraints. Some stakeholders have expressed concern with the gap of 12 years between assessment administrations, which represents a cohort's entire length of schooling. Other stakeholders argue that the NAEP LTT is not very useful now that Main NAEP provides trend information back to the early 1990s, and that it should be eliminated altogether.

Next Steps

In 2012, the Future of NAEP panel recommended exploring ways of consolidating or combining Long-Term Trend and Main NAEP data collections. This is a complex challenge due to the many differences in content, sampling, and administration of the assessments. To explore the feasibility of combining the data collection efforts, and to debate the relative merits of NAEP LTT, the Governing Board is organizing a symposium on the future of NAEP Long-Term Trend. The symposium will take place on the morning of March 2, 2017, immediately preceding the quarterly Governing Board meeting.

In advance of the symposium, Edward Haertel of Stanford University (who previously served as Chair of the Future of NAEP panel and Chair of COSDAM) is preparing a white paper of approximately 30 pages on the history of NAEP Long-Term Trend and a consideration of current issues. The white paper will be distributed to four additional participants, who will each prepare a shorter response (8-10 pages) on their perspective of the future of NAEP LTT. The papers will be disseminated in advance of the symposium and will serve as the basis for discussion during the March 2nd event. In addition, the participants will also discuss their perspectives and solicit external input at a planned session during the annual American Educational Research Association (AERA) conference in April, 2017.

During the May 2017 quarterly meeting, the Governing Board will discuss key takeaways and potential next steps regarding the future of the NAEP Long-Term Trend assessments.

NAEP Long Term Trend Assessments over the Years

The longest running NAEP LTT assessments are the LTT Reading and Mathematics assessments, followed by the LTT Science and Writing assessments. The LTT Science and Writing assessments were discontinued after 1996 (Writing for technical reasons and Science for outdated content). Administration years for each of these subject areas are shaded below, showing the years for which trends over time were reported.

	’69	’70	’71	’73	’75	’77	’78	’80	’82	’84	’86	’88	’90	’92	’94	’96	’99	’04	’08	’12
Reading			■		■		■			■		■	■	■	■	■	■	■	■	■
Mathematics				■			■		■		■		■	■	■	■	■	■	■	■
Science	■	■		■		■			■		■		■	■	■					
Writing										■		■	■	■	■	■				

What Are the Differences Between Long-Term Trend NAEP and Main NAEP?

Although long-term trend and main NAEP both assess mathematics and reading, there are several differences, particularly in the content assessed, how often the assessment is administered, and how the results are reported. These and other differences mean that results from long-term trend and main NAEP cannot be compared directly.

	Long-Term Trend Assessment	Main NAEP Assessment
Origin	Reading series began in 1971. Mathematics series began in 1973.	Reading series began in 1992. Mathematics series began in 1990.
Frequency	Since 2004, long-term trend NAEP has measured student performance in mathematics and reading every four years. Last reported for 2008, it will be reported next for 2012.	Main NAEP assessments measure student performance in mathematics and reading every two years.
Content Assessed	Long-term trend NAEP has remained relatively unchanged since 1990. In the 1970s and '80s, the assessments changed to reflect changes in curriculum in the nation's schools. Continuity of assessment content was sufficient not to require a break in trends. Mathematics focuses on numbers and numeration, variables and relationships, shape and size and position, measurement, and probability and statistics. Basic skills and recall of definitions are assessed. Reading features short narrative, expository, or document passages, and focuses on locating specific information, making inferences, and identifying the main idea of a passage. On average, passages are shorter in long-term trend reading than in main NAEP reading.	Main NAEP assessments change about every decade to reflect changes in curriculum in the nation's schools; new frameworks reflect these changes. Continuity of assessment content was sufficient not to require a break in trends, except in grade 12 mathematics in 2005. Mathematics focuses on numbers, measurement, geometry, probability and statistics, and algebra. In addition to basic skills and recall of definitions, students are assessed on problem solving and reasoning in all topic areas. Reading features fiction, literary nonfiction, poetry, exposition, document, and procedural texts or pairs of texts, and focuses on identifying explicitly stated information, making complex inferences about themes, and comparing multiple texts on a variety of dimensions.
Question formats	Students respond to questions in multiple-choice format; there are also a few short answer questions (scored on a two-point scale). In reading, there are also a few questions requiring an extended answer (usually scored on a five-point scale).	Students respond to questions of several possible types: multiple choice, short answer, and extended answer. Constructed-response questions may be scored as correct or incorrect, or they may be scored on a multi-level scale that awards partial credit.

	Long-Term Trend Assessment	Main NAEP Assessment
Students Sampled	Students are selected by age (9, 13, and 17) to represent the nation and to provide results for student groups such as Black, Hispanic, White, and sometimes others, by gender, family income, school location, and school type (public or private). Students with disabilities (SD) and English language learner (ELL) students are included using the same participation guidelines and with the same accommodations (as needed) in main NAEP. Since 2004, accommodations have been provided to enable participation of more SD and ELL students.	Students are selected by grade (4, 8, and 12). Students represent the nation and provide results for student groups such as Black, Hispanic, White, and sometimes others, by gender, family income, and school location and school type. In some assessments, samples are chosen to report on states or selected large urban districts and as a result, more students must participate. The inclusion and accommodation treatment is the same for main and for long-term trend assessments.
Administration	Long-term trend is assessed every four years, throughout the school year: in October through December for 13-year-olds, January through March for 9-year-olds, and March through May for 17-year-olds. See the schedule for all assessments (long-term trend as well as main NAEP). Test booklets contain three 15-minute blocks of questions, plus one section of student questions concerning academic experiences and demographics. There are no ancillary materials, such as calculators or manipulatives, provided.	Main NAEP mathematics and reading are assessed every two years (the odd-numbered years) at grades 4, 8, and 12. The administration takes place from late January through early March. Test booklets contain two 25-minute blocks, plus student questions concerning academic experiences and demographics. There may be ancillary materials provided with the test booklets.

	Long-Term Trend Assessment	Main NAEP Assessment
Results Reported	National-level performance and how it has changed since the 1970s is reported using scores on a 0-500 scale. Long-term trend also reports descriptive performance levels (150, 200, 250, 300, and 350) that have the same meaning across the three age levels. There are no achievement levels to correspond with those used in main NAEP. There are student questionnaires, but no teacher or school questionnaires.	Main NAEP has been reported since the 1990s for the nation and participating states and other jurisdictions, and since 2002 for selected urban districts. Performance and how it has changed over the past several years is reported using scale scores and achievement levels. Scores are reported using either a 0-300 or 0-500 scale, depending on the subject. The achievement levels reported are *Basic*, *Proficient*, and *Advanced.* Student results are reported in the context of the questionnaires given to the students' teachers and principals.

Source: https://nces.ed.gov/nationsreportcard/about/ltt_main_diff.aspx

Contents

What Is The Nation's Report Card™?

The Nation's Report Card™ informs the public about the academic achievement of elementary and secondary students in the United States. Report cards communicate the findings of the National Assessment of Educational Progress (NAEP), based on assessments conducted periodically in reading, mathematics, science, writing, U.S. history, civics, geography, and other subjects.

NAEP collects and reports information on student performance at the national, regional, and—since 1990 for main NAEP—state levels. Main NAEP assessments track student performance in grades 4, 8, and 12. Since 1971, NAEP's long-term trend assessments have tracked student progress at ages 9, 13, and 17. These assessments are an integral part of our nation's evaluation of the condition and progress of education. Only academic achievement data and related contextual information are collected. The privacy of individual students and their families is protected.

NAEP is a congressionally authorized project of the National Center for Education Statistics (NCES) within the Institute of Education Sciences of the U.S. Department of Education. The Commissioner of Education Statistics is responsible for carrying out the NAEP project. The National Assessment Governing Board oversees and sets policy for NAEP.

Executive Summary

Since the 1970s, the National Assessment of Educational Progress (NAEP) has monitored the academic performance of 9-, 13-, and 17-year-old students with what have become known as the long-term trend assessments. Four decades of results offer an extended view of student achievement in reading and mathematics. Results in this report are based on the most recent performance of more than 50,000 public and private school students who, by their participation, have contributed to our understanding of the nation's academic achievement.

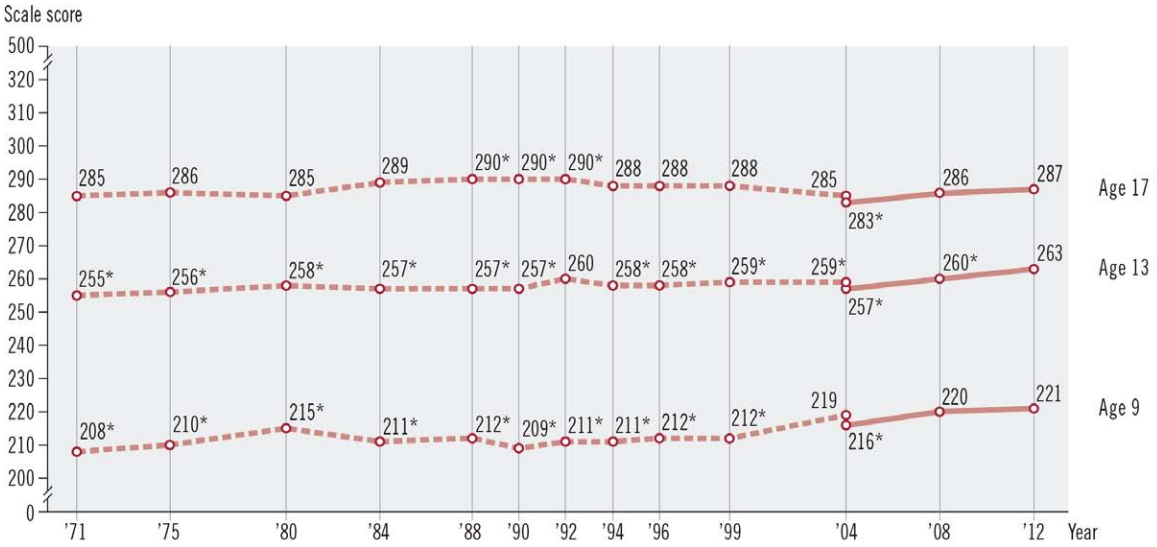
Nine- and 13-year-olds make gains

Both 9- and 13-year-olds scored higher in reading and mathematics in 2012 than students their age in the early 1970s (**figure A**). Scores were 8 to 25 points higher in 2012 than in the first assessment year. Seventeen-year-olds, however, did not show similar gains. Average reading and mathematics scores in 2012 for 17-year-olds were not significantly different from scores in the first assessment year.

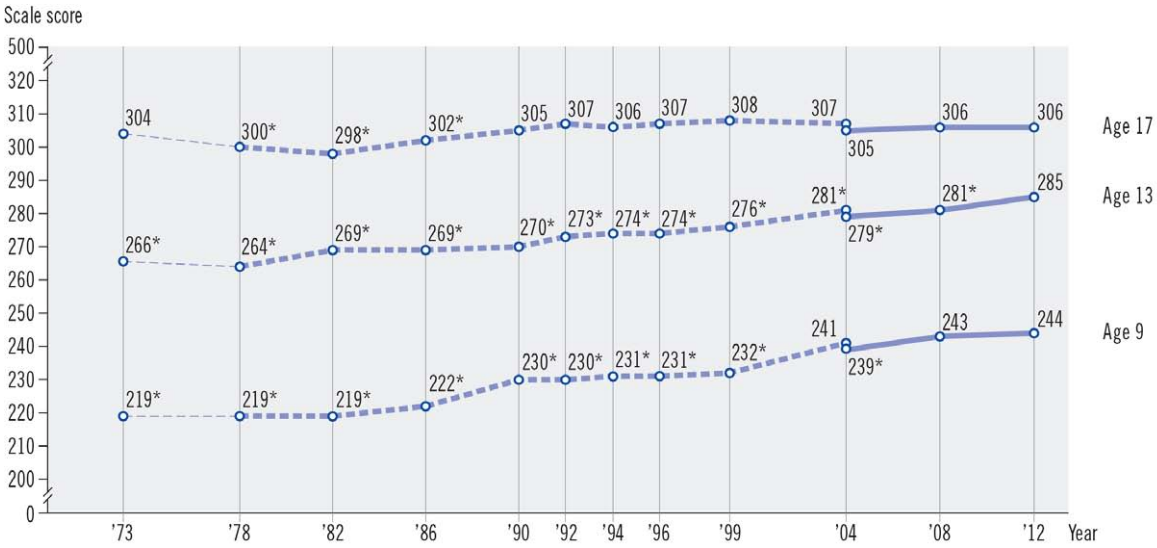
Since the last administration of the assessments in 2008, only 13-year-olds made gains—and they did so in both reading and mathematics.

Figure A. Trend in NAEP reading and mathematics average scores for 9-, 13-, and 17-year-old students

Reading



Mathematics



* Significantly different ($p < .05$) from 2012.

Extrapolated data adjusting for the limited number of questions from the 1973 mathematics assessment in common with the assessments that followed

Original assessment format using the same assessment procedures established for the first assessment year

Revised assessment format introducing more current assessment procedures and content

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), various years, 1971–2012 Long-Term Trend Reading and Mathematics Assessments.

Racial/ethnic and gender gaps narrow

Closing achievement gaps is a goal of both national and state education policy. The results from the 2012 NAEP long-term trend assessments show some progress toward meeting that goal. The narrowing of the White – Black and White – Hispanic score gaps in reading and mathematics from the 1970s is the result of larger gains by Black and Hispanic students than White students. Only the White – Hispanic gap in mathematics at age 9 has not shown a significant change from the early 1970s.

Female students scored higher in reading than male students at all three ages. The 2012 results show 9-year-old males making larger score gains than females. This has led to a narrowing of the gender gap at age 9 as compared to 1971.

In mathematics, male 17-year-old students scored higher than female students. The gender gap at age 17 narrowed because female students made gains from 1971 to 2012, but 17-year-old male students did not.

Reading

Characteristic	Score changes from 1971			Score changes from 2008		
	Age 9	Age 13	Age 17	Age 9	Age 13	Age 17
All students	↑ 13	↑ 8	↔	↔	↑ 3	↔
Race/ethnicity						
White	↑ 15	↑ 9	↑ 4	↔	↔	↔
Black	↑ 36	↑ 24	↑ 30	↔	↔	↔
Hispanic[1]	↑ 25	↑ 17	↑ 21	↔	↑ 7	↔
Gender						
Male	↑ 17	↑ 9	↑ 4	↔	↔	↔
Female	↑ 10	↑ 6	↔	↔	↑ 3	↔
Score gaps						
White – Black	Narrowed	Narrowed	Narrowed	↔	↔	↔
White – Hispanic	Narrowed	Narrowed	Narrowed	↔	Narrowed	↔
Female – Male	Narrowed	↔	↔	↔	↔	↔

Mathematics

Characteristic	Score changes from 1973			Score changes from 2008		
	Age 9	Age 13	Age 17	Age 9	Age 13	Age 17
All students	↑ 25	↑ 19	↔	↔	↑ 4	↔
Race/ethnicity						
White	↑ 27	↑ 19	↑ 4	↔	↔	↔
Black	↑ 36	↑ 36	↑ 18	↔	↔	↔
Hispanic	↑ 32	↑ 32	↑ 17	↔	↔	↔
Gender						
Male	↑ 26	↑ 21	↔	↔	↔	↔
Female	↑ 24	↑ 17	↑ 3	↔	↑ 5	↔
Score gaps						
White – Black	Narrowed	Narrowed	Narrowed	↔	↔	↔
White – Hispanic	↔	Narrowed	Narrowed	↔	↔	↔
Male – Female[2]	↔	↔	Narrowed	↔	↔	↔

[1] Reading results for Hispanic students were first available in 1975. Therefore, the results shown in the 1971 section for Hispanic students are from the 1975 assessment.

[2] Score differences between male and female students in mathematics were not found to be statistically significant ($p < .05$) at age 9 in 1973, 2008, or 2012, and at age 13 in 1973 and 2012.

NOTE: Black includes African American, and Hispanic includes Latino. Race categories exclude Hispanic origin.

↑ Indicates score was higher in 2012

↔ Indicates no significant change in 2012

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), various years, 1971–2012 Long-Term Trend Reading and Mathematics Assessments.

NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

Assessment Development Committee

Item Review Schedule
November 2016 - April 2017

October 13, 2016

Review Package to Board	Board Comments to NCES	Survey/ Cognitive	Review Task	Approx. Number Items	Status
11/9/16	11/29/16	Cognitive	2019 Math (12) Pilot (SBT) Draft builds	2 tasks	For review at November Board meeting
11/9/16	11/29/16	Cognitive	2019 Science (4, 8) Pilot (ICT) Draft Builds	2 tasks	
11/9/16	11/29/16	Survey	2019 Math (12) Pilot	10-20	
11/9/16	11/29/16	Survey	2019 Reading (12) Pilot	10-20	
1/5/17	1/26/17	Cognitive	2019 Reading (12) Pilot (SBT) Draft Builds	2 tasks	
2/3/17	2/20/17	Cognitive	2019 Science (12) Pilot (ICT) Draft Builds	2 tasks	
4/3/17	4/26/17	Cognitive	2019 Reading (12) Pilot (DI)	20	
4/20/17	05/02/17	Survey	2019 Science (4, 8, 12) Pilot	80-100	

NOTE: "SBT" indicates Scenario-Based Task
"DI" indicates Discrete Item