National Assessment Governing Board

Committee on Standards, Design and Methodology

November 18, 2016

AGENDA

10:15 – 11:30 am	**Joint Session with the Assessment Development Committee (ADC)**	
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, 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
11:30 am – 12:20 pm	2017 Grade 4 Writing Achievement Levels Setting Project Update and Design Document *Tim O'Neil, Pearson*	Attachment D
12:20 – 12:30 pm	Information Items • Uses of NAEP • Comparing NAEP with State Assessments	Attachment E Attachment F

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:

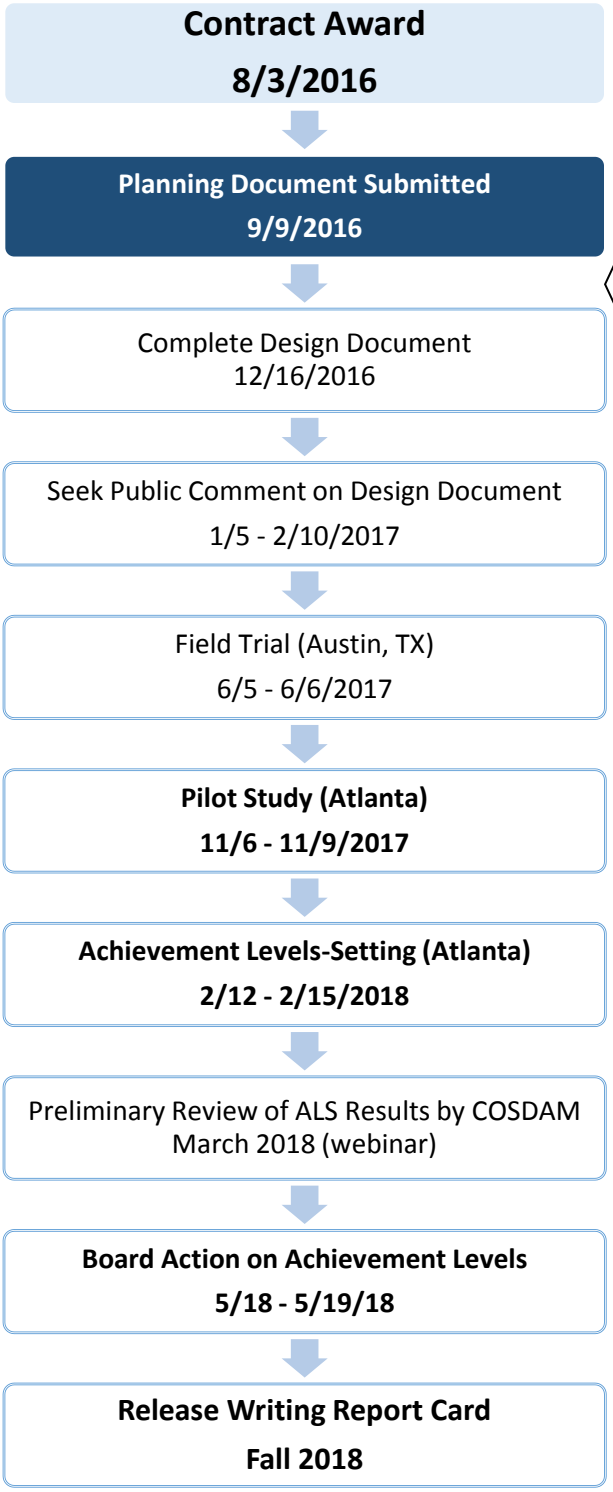
1. Are there differences in student performance at the item level according to items' coverage in the CCSS?
2. Are there differences in psychometric properties of items according to items' coverage in the CCSS?
3. How would state mean scores change if items student achievement is estimated using only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

Developing Achievement Levels for the National Assessment of Educational Progress Writing at Grade 4



Purpose: The purpose of this session is to provide an update to the Committee on Standards, Design and Methodology (COSDAM) regarding the development of achievement levels for the 2017 NAEP Grade 4 Writing and to present the plans for implementing the body of work standard setting methodology. In this session, Tim O'Neil, NAEP Grade 4 Writing Achievement Levels-Setting (ALS) Project Director for Pearson, will provide an update on the project and an overview of the Design Document.

Legend:
Light shading: Completed
Dark shading: Current status
No shading: To be completed after 11/17/2016

Purpose: The purpose of this session is to provide an update to the Committee on Standards, Design and Methodology (COSDAM) regarding the development of achievement levels for 2017 NAEP Grade 4 Writing and to present the plans for implementing the body of work standard setting methodology. In this session, Dr. Tim O'Neil, NAEP Grade 4 Writing Achievement Levels-Setting (ALS) Project Director for Pearson, will provide an update on the project and an overview of the Design Document.

Project Overview: On August 3, 2016, the National Assessment Governing Board (Governing Board) awarded a contract to Pearson (as a result of a competitive bidding process) for developing achievement levels for the National Assessment of Educational Progress (NAEP) for grade 4 writing. The 2017 Grade 4 NAEP Writing assessment is the first administration of the grade 4 assessment developed to meet the design specifications described in the current computer-based Writing Framework. The assessment is a digital-based assessment, comprised of constructed response items, for which students compose and construct their responses using word processing software on a tablet. The assessment is to be administered to a nationally representative sample of approximately 22,000 grade 4 students in the spring of 2017.[1]

Dr. Tim O'Neil is the grade 4 writing ALS project director at Pearson and Dr. Marc Johnson is the assistant project director at Pearson. Pearson will conduct a field trial, a pilot study, and an achievement levels-setting (ALS) meeting and produce a set of recommendations for the Governing Board to consider in establishing achievement levels for the grade 4 NAEP writing assessment. The Governing Board is expected to take action on the writing grade 4 achievement levels during the May 2018 meeting. Pearson will utilize a body of work methodology using Moodle software to collect panelist ratings and present feedback. Dr. Lori Nebelsick-Gullet will serve as the process facilitator for the pilot and operational ALS meetings; Victoria Young will serve as the content facilitator for the pilot and operational ALS meetings; and Drs. Susan Cooper Loomis and Steven Fitzpatrick will serve as consultants.

For setting standards, Pearson will use a body of work approach in which panelists will make content-based cut score recommendations. The body of work methodology is a holistic standard setting method for which panelists evaluate sets of examinee work (i.e., bodies of work) and provide a holistic judgment about each student set. These content-based judgments will be made over three rounds. The process to be implemented for the standard setting meeting follows body of work procedures used in previous NAEP standard setting studies. In addition, a field trial will be conducted prior to the pilot study which will provide an opportunity to try out a number of key aspects of the ALS plan, including the logistical design of the ALS studies such as the use of tablets and laptop computers, the ease with which the panelists can enter judgments and questionnaire responses, and the arrangement of tables and panelists.

The Governing Board policy on Developing Student Performance Levels for NAEP (https://www.nagb.org/content/nagb/assets/documents/policies/developing-student-performance.pdf) requires appointment of a committee of technical advisors who have expertise in standard setting and psychometrics in general, as well as issues specific to NAEP. These

[1] Achievement levels were set for Writing grades 8 and 12 with the 2011 administration of those assessments. The grade 4 assessment initially was scheduled to be administered in 2013 but the Governing Board postponed it to 2017 due to budgetary constraints.

advisors will be convened for 8 in-person meetings and up to 6 webinars to provide advice at every key point in the process. They provide feedback on plans and materials before activities are implemented and review results of the process and analyses. Six external experts in standard setting are serving on the Technical Advisory Committee on Standard Setting (TACSS):

Dr. Gregory Cizek
Professor of Educational Measurement, University of North Carolina at Chapel Hill

Dr. Barbara Dodd
Professor of Professor of Quantitative Methods, University of Texas at Austin

Dr. Steve Ferrara
Independent Consultant

Dr. Matthew Johnson
Associate Professor of Statistics and Education, Teachers College, Columbia University

Dr. Vaughn G. Rhudy
Executive Director, Office of Assessment, West Virginia Department of Education

Dr. Mary Pitoniak
Senior Strategic Advisor for Statistical Analysis, Data Analysis, and Psychometric Research, Educational Testing Service (NAEP Design, Analysis, and Reporting Contractor)

November 2016 Update:

Kickoff Meeting

On Monday, August 8, 2016, Pearson staff met with members of the Governing Board staff to initiate work on the grade 4 writing ALS project. The purposes of the meeting were to identify the roles and responsibilities of Governing Board staff and contractor staff, to review and discuss proposed contract work, to discuss aspects of contract management, such as submission of reports, deliverables, and invoices, and to establish communication procedures.

Planning Document

On September 9, 2016, Pearson submitted the Planning Document to the Governing Board that provides details and timelines for each task conducted as part of the ALS process, to enable the Board and Board staff to complete long-range planning for the grade 4 writing ALS. The Planning Document included a Gantt chart project schedule, for use in monitoring contract deliverables.

Technical Advisory Committee on Standard Setting (TACSS) webinar

On September 22, 2016 the first webinar meeting of the TACSS for the 2017 Grade 4 writing ALS was convened. Topics of discussion included an introduction to the Grade 4 NAEP Writing

Framework, an overview of the Planning Document (to include high level plans for the field trial, pilot study, and operational ALS meeting, panelist recruitment and external validity studies), consideration of the inclusion of a borderline achievement level descriptors task (which had been part of Pearson's initial proposal), and a description of computers and software to be used in the ALS.

The overall body of work design closely follows the design implemented for the 2011 Grade 8 and 12 NAEP Writing ALS in that the third round of panelist ratings is conducted with a new/comparable set of bodies of work instead of a pinpointing round. This had been vetted through the 2011 TACSS and recommended as the best course. Additionally, one TACSS member noted that the ordering of booklets for the classification by panelists using the body of work method was an important issue for the 2011 ALS. He recalled that the booklets were originally presented in order from lowest performance to highest, but the decision was made to change the ordering to be from highest performance to lowest for the ALS. The current TACSS requested further information on these points which will be discussed at the first TACSS meeting on November 2nd and 3rd.

External Validity Study Design Meeting

On October 5, 2016 Tim O'Neil, and Marc Johnson (Assistant Project Director) met with Victoria Young (content facilitator) to discuss the viability of proposed data sources and design of external validity studies in support of the ALS outcome. Details will be included in the Design Document and discussed at the first TACSS meeting on November 2nd and 3rd.

Design Document

The first draft of the Design Document was submitted to Governing Board staff on October 4, 2016. The Design Document is intended to provide the foundation for all achievement levels-setting activities. The Design Document for the grade 4 achievement levels-setting process includes discussion of the methodology, procedures, and documentation of the entire project. During the November 2016 COSDAM session, ALS Project Director Tim O'Neil will provide an overview of the Design Document. A draft of the Design Document will be sent to COSDAM members by email no later than Friday, November 11th.

Appropriate uses of NAEP data

Since its inception, the NAEP Validity Study (NVS) Panel has been engaged in research on various aspects of the validity of the NAEP assessment program. The choice of topics was informed by the judgments of both panel members and the National Center for Education Statistics (NCES) regarding the most pressing validity research needs at any given point in time. In October 2002, NCES asked the panel to put together a framework for their work and also asked the panel to be more forward looking in generating possible research topics to be studied. As a result of this request, in 2002, the panel developed a research agenda that was based on a framework defined by categories:

1. The constructs measured within each of NAEP's subject domains
2. The manner in which these constructs are measured
3. The representation of the population to be assessed
4. The analysis of data
5. The reporting and use of NAEP results
6. The assessment of trends

This framework, which was published as an NVS report, continued to be used as an organizing tool for the panel for several subsequent annual updates to the validity research agenda until the recent past.

However, by the start of the current five-year contract (2013-2018), it was time to update the NVS framework in light of more recent developments. The most notable of these was criticism from a Congressionally-mandated evaluation of the NAEP program that was completed in 2009 by of scholars from the Buros Center for Testing at the University of Nebraska–Lincoln and the Center for Educational Assessment at the University of Massachusetts–Amherst. The evaluators argued that the then-current approach to NAEP validity research seemed to imply that the validity of NAEP was in the instrument rather than in the uses to which NAEP has been put. Instead, validity must be established for each purpose or use. More specifically the evaluation said: "Validation is an ongoing process because it is the interpretation or use of assessment results that are supported (validated), not the assessment instrument itself." (Buckendahl, Davis, Plake, Sireci, Hambleton, Zenisky and Wells, 2009, p.xvii). They also noted that, in their view, much of the validity research that NCES had done to this point in time was piecemeal and without the benefit of a comprehensive framework. The specific language the evaluators used is: "NAEP has not had the benefit of a comprehensive framework to guide the *systematic* accumulation of evidence in order to substantiate the ways in which its assessment results may be reasonably interpreted and applied." (Buckendahl et al., p .xi).

Finally, they argued that "there is a need for an ongoing, systematic appraisal of the validity of the interpretations and uses being built on the NAEP assessments." (Buckendahl et al., p.14).

In response to the criticism of Buckendahl et al. (2009), NCES requested AIR's NAEP Statistical Services Institute (NESSI) to construct a comprehensive NAEP validity framework based on the uses to which NAEP is put. In order to keep the task a manageable one, the NESSI team decided to focus only on uses designated by the federal government. That is, the framework does not include the various non-official uses to which stakeholders might employ NAEP.

The NESSI staff identified five such official uses:

1. Monitoring student performance at a given point in time in mathematics, reading and other subjects at grades 4 and 8 (and at grade 12) at the national, state and selected district levels using both scale scores and achievement levels
2. Monitoring trends in mathematics, reading and other subjects (and at grade 12) at the national, state and district levels and reported both by scale scores and achievement levels
3. Comparing the performance of achievement across states and districts as well as internationally
4. Disaggregating and reporting results by race, ethnicity, socioeconomic status, gender, disability and limited English proficiency
5. Using NAEP results to inform and evaluate federal educational policies

The team then asked what validity questions would have to be answered to be able to assess the validity of a particular use. The crossing of the various uses of NAEP by its related validity questions resulted in the validity framework.

By agreement with NCES, NVS used the NESSI framework as a starting point for the new framework, which was primarily intended to provide structure for an NVS review of prior research on NAEP validity and to guide the choice of topics for future NVS validity studies.

After conversations with NCES and NAGB, the framework will be expanded to include a general discussion of how test scores and NAEP data are intended to be interpreted and used. This discussion will include evidence as to why specific uses and analyses are not an appropriate use of these data.

Comparing NAEP with State Assessments

The NAEP Validity Studies (NVS) panel is undertaking a suite of three interrelated studies to examine the alignment between current-generation state assessments and NAEP, with the goal of informing the following validity question:

- At grades 4 and 8, does NAEP remain sufficiently aligned with what students are learning in the classroom to continue to serve as a valid measure of what students know and can do across the nation?

Given the move by the vast majority of states to adopt either the Common Core or other college- and career-ready standards, it is important for policymakers and practitioners to understand the degree to which NAEP's assessments continue to measure what is in the various curricula being taught by the states. Three previous NVS studies explored the same validity question by comparing NAEP frameworks, and then NAEP 2015 math items, to the Common Core Standards. These studies found substantial overlap, but also some major areas of difference. The 2015 study also found some correspondence between NAEP subscale results and the extent to which NAEP content within subscales was aligned with CCSS content standards at or below the grade tested by NAEP.

The three new studies, each led by different NVS panel members are:

- Math item comparison study
    - PIs: Phil Daro and Gerunda Hughes
- English Language Arts (ELA) item comparison study
    - PIs: Sheila Valencia and Karen Wixson (former panel member)
- Linking/joint scaling study
    - PI: David Thissen

The research studies plan on using items and test data from the 2017 assessments. Working with the Council of Chief State School Officers, NVS has identified several states interested in participating. State affiliation includes Smarter Balanced, PARCC, and states using other assessments not affiliated with these two Consortia.

The item comparison studies will use expert panels to compare the content and skills addressed by NAEP to those covered in each of the state assessments.

The linking/joint scaling study will place NAEP items on a common scale with each of the other assessments in order to examine, empirically, the relationship of NAEP items and other assessment items. (The resulting displays will be similar to those obtained from NAEP item mapping exercises.)

We plan to conduct the item comparison studies in spring 2017; the linking/joint scaling study will start when 2017 assessment data are available.