

Committee on Standards, Design and Methodology

March 6, 2020

Discovery



AGENDA

10:30 – 10:45 am	Welcome, Introductions and Review of Agenda	
	Review of Strategic Vision Activities	<i>Attachment A</i>
	Upcoming COSDAM Activities	<i>Attachment B</i>
	<i>Andrew Ho, COSDAM Chair</i>	
10:45 – 10:55 am	ACTION: Intended Meaning of NAEP	<i>See plenary tab</i>
	<i>Andrew Ho</i>	
10:55 – 11:05 am	ACTION: Achievement Levels Work Plan	<i>Attachment C</i>
	<i>Andrew Ho</i>	
11:05 – 11:50 am	Update on Design of 2021 NAEP Reading and Mathematics Assessments	<i>Attachment D</i>
	<i>Enis Dogan, National Center for Education Statistics</i>	
11:50 am – 12:20 pm	Draft Achievement Levels Procedures Manual	<i>Attachment E</i>
	<i>Andrew Ho</i>	
	<i>Sharyn Rosenberg, Assistant Director for Psychometrics</i>	
12:20 – 12:30 pm	Information Item: NAEP Linking Studies	<i>Attachment F</i>

Strategic Vision Activities Led by COSDAM

During the November 2016 Board meeting, a [Strategic Vision](#) was formally adopted to guide the Board's work over the next several years. For each activity led by COSDAM, information is provided below to describe the current status and recent work, planned next steps, and the ultimate desired outcomes. Please note that many of the Strategic Vision activities require collaboration across committees and with NCES, but the specific opportunities for collaboration are not explicitly referenced in the table below. In addition, the activities that include contributions from COSDAM but are primarily assigned to another standing committee also have not been included below.

As of March 2020, most of the COSDAM-led activities have been accomplished or are directly referenced in new initiatives such as the Achievement Levels Work Plan.

Strategic Vision Activity	Current Status and Recent Work	Planned Next Steps	Desired Outcome
SV #2: Increase opportunities to connect NAEP to administrative data and state, national, and international student assessments <i>Incorporate ongoing linking studies to external measures of current and future achievement in order to evaluate the NAEP scale and add meaning to the NAEP achievement levels in reporting. Consider how additional work could be pursued across multiple subject areas, grades, national and international assessments, and longitudinal outcomes</i>	Informational update on current studies was provided in the March 2018 COSDAM materials Results from the national NAEP-ACT linking study were presented to COSDAM at the March 2019 Board meeting; the report is being finalized Related work is part of the Achievement Levels Work Plan slated for Board action at the upcoming March 2020 meeting	Complete ongoing studies Decide what new studies to take on Decide how to use and report existing and future results Complete additional studies	NAEP scale scores and achievement levels may be reported and are better understood in terms of how they relate to other important indicators of interest (i.e., other assessments and milestones)

Strategic Vision Activity	Current Status and Recent Work	Planned Next Steps	Desired Outcome
<p>SV #3: Expand the availability, utility, and use of NAEP resources, in part by creating new resources to inform education policy and practice</p> <p><i>Research when and how NAEP results are currently used (both appropriately and inappropriately) by researchers, think tanks, and local, state and national education leaders, policymakers, business leaders, and others, with the intent to support the appropriate use of NAEP results (COSDAM with R&D and ADC)</i></p> <p><i>Develop a statement of the intended and unintended uses of NAEP data using an anticipated NAEP Validity Studies Panel (NVS) paper and the Governing Board's research as a resource (COSDAM with NCES)</i></p> <p><i>Disseminate information on technical best practices and NAEP methodologies, such as training item writers and setting achievement levels</i></p>	<p>Ina Mullis of the NVS panel spoke with COSDAM at the March 2017 Board meeting and is working on a white paper about the history and uses of NAEP</p> <p>During the August 2018 Board meeting, COSDAM discussed how to use information from an ongoing study to inform a policy statement on intended and appropriate uses of NAEP</p> <p>During the March 2019 and May 2019 Board meetings (and via email), COSDAM discussed drafts of a statement on the intended meaning of NAEP (statement was discussed in plenary session at the November 2019 Board meeting and is slated for action at the upcoming March 2020 meeting)</p> <p>This idea was generated during the August 2017 COSDAM discussion of the Strategic Vision activities</p>	<p>Full Board adoption of statement on intended meaning of NAEP</p> <p>NCES produces documentation of validity evidence for intended uses of NAEP scale scores</p> <p>Governing Board produces documentation of validity evidence for intended uses of NAEP achievement levels</p> <p>(The remaining work described above has been incorporated into the Achievement Levels Work Plan slated for Board action at the upcoming March 2020 meeting)</p> <p>Work with NCES and R&D to refine list of technical topics for dissemination efforts</p>	<p>Board adopts formal statement about intended meaning of NAEP. The goal is to increase appropriate uses and decrease inappropriate uses.</p> <p>Stakeholders benefit from NAEP technical expertise</p>

Strategic Vision Activity	Current Status and Recent Work	Planned Next Steps	Desired Outcome
<p>SV# 5: 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</p> <p><i>Consider new approaches to creating and updating the achievement level descriptors and update the Board policy on achievement levels</i></p>	<p>A revised policy on Developing Achievement Levels for NAEP was unanimously adopted during the November 2018 Board meeting, following a year-long process of soliciting expert input and public comment</p> <p>The Achievement Levels Working Group was formed in March 2019 to develop a comprehensive plan for responding to the remaining recommendations from the evaluation</p> <p>COSDAM and R&D discussed preliminary ideas from Achievement Levels Working Group during the August 2019 Board meeting</p> <p>Full Board discussion of the Achievement Levels Work Plan took place at the November 2019 Board meeting; action is planned for the upcoming March 2020 meeting following a Board discussion call on February 25th</p>	<p>At the upcoming March 2020 Board meeting, COSDAM will have an initial discussion on the draft Achievement Levels Procedures manual</p> <p>Additional edits and comments on the Achievement Levels Procedures manual will be discussed on an optional COSDAM call in late March or early April</p> <p>COSDAM approval of the Achievement Levels Procedures Manual is planned for the May 2020 Board meeting</p> <p>Following adoption of the Achievement Levels Work Plan, Board staff will work on implementing the plan, including a new procurement to review and revise the achievement level descriptions</p>	<p>Board has updated policy on achievement levels that meets current best practices in standard setting and is useful for guiding the Board's achievement levels setting work</p>

Strategic Vision Activity	Current Status and Recent Work	Planned Next Steps	Desired Outcome
<p>SV# 7: Research policy and technical implications related to the future of NAEP Long-Term Trend assessments in reading and mathematics</p> <p><i>Support development and publication of multiple papers exploring policy and technical issues related to NAEP Long-Term Trend. In addition to the papers, support symposia to engage researchers and policymakers to provide stakeholder input into the Board's recommendation</i></p>	<p>Several Board discussions and discussions with external stakeholders took place during 2017 and 2018</p> <p>The NAEP budget in Fiscal Year 2019 was increased by \$2 million with the goal of moving up the next administration of LTT</p> <p>Following discussion at the November 2018 Board meeting, former Chair Bev Perdue sent a response to Congress indicating that the Board would add a paper-based 2020 LTT administration to the NAEP Assessment Schedule</p> <p>The Board took action on a NAEP Assessment Schedule during the May 2019 Board meeting, to include administration of the Long-Term Trend Assessments</p>	<p>NCES will present design considerations for LTT bridge studies at a future Board meeting (to inform the 2024 NAEP LTT)</p>	<p>Determine whether changes to the NAEP LTT schedule, design and administration are needed (led by Executive Committee and NCES)</p>

Strategic Vision Activity	Current Status and Recent Work	Planned Next Steps	Desired Outcome
<p>SV# 9: Develop policy approaches to revise the NAEP assessment subjects and schedule based on the nation's evolving needs, the Board's priorities, and NAEP funding</p> <p><i>Pending outcomes of stakeholder input (ADC activity), evaluate the technical implications of combining assessments, including the impact on scaling and trends</i></p>	<p>COSDAM presentation and discussion on initial considerations for combining assessments</p> <p>During the past few years, there have been several full Board presentations and discussions on the NAEP Assessment Schedule</p> <p>Action on the NAEP Assessment Schedule took place during the May 2019 Board meeting</p>	<p>The planned design change in 2021 is based on coordination of two subjects; implications for analysis and reporting will be discussed with the Board as the design is implemented</p> <p>Revision of the 2027 NAEP Science Framework may consider whether Technology and Engineering Literacy should be consolidated with Science (white papers may be commissioned on this topic)</p>	<p>Determine whether NAEP Assessment Schedule should include any consolidated frameworks or coordinated administrations</p>
<p>SV# 10: Develop new approaches to measure the complex skills required for transition to postsecondary education and career</p> <p><i>Continue research to gather validity evidence for using 12th grade NAEP reading and math results to estimate the percentage of grade 12 students academically prepared for college</i></p>	<p>During the November 2018 Board meeting, the Board took action to explore the creation of a postsecondary preparedness conceptual framework and dashboard; this work was discussed by the full Board at the November 2019 Board meeting</p> <p>As part of the Strategic Vision 2025 discussions this year, the Board will decide whether and how preparedness research should be included in future efforts</p>	<p>Decide whether Board should make stronger statement and/or set "benchmarks" rather than using "plausible estimates" of NAEP for predicting academic preparedness for college</p>	<p>Statements about using NAEP as an indicator of academic preparedness for college continue to be defensible and to have appropriate validity evidence</p>

Upcoming COSDAM Activities

The Committee on Standards, Design and Methodology (COSDAM) “shall be responsible for recommending to the Board policies related to all technical aspects of NAEP, for developing the achievement levels, and for recommending to the Board the achievement levels for adoption for each grade and subject in the National Assessment. The areas this committee shall address include but are not limited to:

- (1) technical issues dealing with NAEP assessments;
- (2) overall issues related to the design, methodology, and structure of NAEP assessments;
- (3) maintaining the integrity of trend lines while encouraging reasonable experimentation and trials of new approaches;
- (4) maximizing utility of NAEP data;
- (5) receiving and reviewing NAEP evaluation and validity studies; and
- (6) developing a process for review of the technical quality of the assessment.”

([By-laws](#) of the National Assessment Governing Board, amended 11/20/2010)

During the November 2019 COSDAM meeting, Chair Andrew Ho led a brief discussion on the Committee’s responsibilities, recent activities, and upcoming priorities. During that discussion, there was a request for additional information about specific timelines associated with upcoming work for the Committee.

Throughout 2020 and 2021, the majority of time (for Committee staff and during quarterly meetings) is expected to be spent on the implementation of the Achievement Levels Work Plan. There are several different activities requiring ongoing work and COSDAM discussion. In addition, the process of wrapping up the existing Strategic Vision and planning and implementing the next Strategic Vision will also require ongoing Committee discussion. In addition to activities that can be anticipated well in advance, there are many issues that often arise with little notice (e.g., operational issues, Board policy discussions with technical implications) that also require COSDAM input.

A preliminary timeline of planned activities anticipated to be led (or co-led) by COSDAM between now and the end of next year is presented in a chart on the following page. These activities will be handled through a variety of means: discussions at quarterly Board meetings, informational updates at quarterly Board meetings, and through additional webinars or calls in between Board meetings (when necessary).

Anticipated Activities and Actions Led (or Co-Led) by COSDAM: 2020 – 2021

	March 2020	May 2020	August 2020	November 2020	March 2021	May 2021	August 2021	November 2021
Achievement Levels Work Plan								
Full Board Action on Work Plan								
Review and Approve ALS Procedures Manual								
Review Technical Memo on Synthesizing and Reporting Findings from NAEP Linking Studies								
Plans for Additional NAEP Linking Studies								
Review and Revise Math and Reading ALDs								
Full Board Action on Math and Reading Reporting ALDs								
Discuss Advisory Group Ideas for Communicating NAEP Achievement Levels (with R&D)								
Develop Interpretative Guides (with R&D)								
Full Board Approval of Interpretative Guides								
Collect and Document Validity Evidence								
Strategic Vision 2020 and 2025								
Close Out Existing Activities								
Plan New Activities								
Implement New Activities								
Other Known Activities								
Review of Specifications for 2025 Reading Framework (with ADC)								
Discuss Results from 2021 Design of Reading and Math								
Provide Input on Design of All 2023 Assessments								

Developing a Comprehensive Plan to Implement the Governing Board’s Response to the 2016 Evaluation of NAEP Achievement Levels

Public Law 107-279 states:

The achievement levels shall be used on a trial basis until the Commissioner for Education Statistics determines, as a result of an evaluation under subsection (f), that such levels are reasonable, valid, and informative to the public.

Even after being in use for over 25 years and undergoing several evaluations, the NAEP achievement levels are still considered to be on a trial basis. The [2016 evaluation of NAEP achievement levels](#), conducted by the National Academies of Sciences, Engineering, and Medicine, focused on the NAEP mathematics and reading achievement levels for grades 4, 8, and 12. This evaluation report, presented to the Governing Board at its November 2016 meeting, stated, “During their 24 years [the achievement levels] have acquired meaning for NAEP’s various audiences and stakeholders; they serve as stable benchmarks for monitoring achievement trends, and they are widely used to inform public discourse and policy decisions. Users regard them as a regular, permanent feature of the NAEP reports” (page Sum-8). This evaluation included several recommendations, and the Board issued a formal response noting its planned actions in December 2016.

One important aspect of the Board’s response to the 2016 evaluation was a commitment to update the guidance provided in the Board policy statement on NAEP achievement levels. The Committee on Standards, Design and Methodology (COSDAM) began working to update this policy guidance in March 2017, and the revised policy on [Developing Student Achievement Levels for NAEP](#) was adopted by the Board in November 2018.

During the March 2019 Board meeting, Governing Board Chair Beverly Perdue established an Achievement Levels Working Group¹ to develop a comprehensive plan (including a list of activities for the Governing Board to pursue in conjunction with the National Center for Education Statistics) to fully respond to the evaluation. Over the past year, the Working Group has convened by telephone and in person to produce the attached plan. Preliminary ideas were shared and discussed with NCES Commissioner Lynn Woodworth and Associate Commissioner Peggy Carr during the spring and summer of 2019, as well as with COSDAM and Reporting and Dissemination (R&D) Committee members during the August 2019 Board meeting.

During the November 2019 Board meeting, Achievement Levels Working Group Chair Gregory Cizek presented the proposed Achievement Levels Work Plan for input and discussion by the full Board. An optional full Board call to discuss any remaining questions or issues will be held on Tuesday, February 25th from 3:00 – 3:45 p.m. The Board is scheduled to take action on the plan during the upcoming March 2020 Board meeting.

¹ The Achievement Levels Working Group was comprised of the following members: Gregory Cizek (Chair), Father Joe O’Keefe, Fielding Rolston, Linda Rosen, and Joe Willhoft.



National Assessment Governing Board Achievement Levels Work Plan

Overview

The National Assessment Governing Board has developed a comprehensive work plan (the Plan) to fully respond to the [National Academies of Sciences, Engineering, and Medicine \(NAS\) evaluation of NAEP achievement levels](#). The ultimate aim of the Plan is to develop a body of evidence that provides a sound basis for removal of the trial status of the NAEP achievement levels. Other related goals are to develop, for Governing Board members and other interested stakeholders, a summary of the validity evidence supporting the interpretation of NAEP achievement levels and to facilitate clear, accurate, and informative reporting of NAEP achievement level results to the public. The Plan described here includes a list of activities (and associated timelines) to be pursued in conjunction with the National Center for Education Statistics (NCES). COSDAM will provide oversight for the Plan's implementation, in conjunction with other committees and NCES, as appropriate.

Background

The Governing Board issued an initial response to the NAS evaluation in December 2016 (see Appendix A) and adopted a revised policy on [Developing Student Achievement Levels for NAEP](#) in November 2018. This Plan provides detail about how each of the seven recommendations from the evaluation will be addressed (using guidance from the revised policy statement, where appropriate), including roles and priorities for accomplishing the work. Supplementing this Plan is a statement of intended purpose and meaning of NAEP (see Appendix B).

As indicated above, a primary aim of the Plan is to develop a body of evidence that provides a sound basis for removal of the trial status of the NAEP achievement levels. According to the NAEP legislation (PL 107-279), "The achievement levels shall be used on a trial basis until the Commissioner for Education Statistics determines, as a result of an evaluation under subsection (f), that such levels are reasonable, valid, and informative to the public." The proposed Plan aligns to those priorities; the criteria "Reasonable," "Valid," and "Informative to the Public"

have been indicated in the proposed responses to the NAS Committee recommendations described below.

Input from NCES suggests that the criterion of “informative to the public” as particularly important, where “the public” is interpreted to be groups who are responsible for using NAEP results directly and/or communicating information about NAEP achievement levels to others, including, but not limited to, state and federal legislators, education administrators at all levels, researchers and policy makers who use NAEP data, and media who cover education).

The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014; hereafter the *Standards*) comprise a collection of professional best practices for all aspects of assessment, including achievement level setting. The following Plan was informed by the guidance provided in the *Standards*.

Responding to Recommendations #1, 2, and 3 (Valid)

Recommendation #1: Alignment among the frameworks, the item pools, the achievement-level descriptors, and the cut scores is fundamental to the validity of inferences about student achievement. In 2009, alignment was evaluated for all grades in reading and for grade 12 in mathematics, and changes were made to the achievement-level descriptors, as needed. Similar research is needed to evaluate alignment for the grade 4 and grade 8 mathematics assessments and to revise them as needed to ensure that they represent the knowledge and skills of students at each achievement level. Moreover, additional work to verify alignment for grade 4 reading and grade 12 mathematics is needed.

Recommendation #2: Once satisfactory alignment among the frameworks, the item pools, the achievement-level descriptors, and the cut scores in NAEP mathematics and reading has been demonstrated, their designation as trial should be discontinued. This work should be completed and the results evaluated as stipulated by law: (20 U.S. Code 9622: National Assessment of Educational Progress: <https://www.law.cornell.edu/uscode/text/20/9622> [September 2016]).

Recommendation #3: To maintain the validity and usefulness of achievement levels, there should be regular recurring reviews of the achievement-level descriptors, with updates as needed, to ensure they reflect both the frameworks and the incorporation of those frameworks in NAEP assessments.

The first three recommendations of the evaluation are inter-related. Recommendation #1 is focused on reading and math and covers all of the ALDs throughout the process, whereas Recommendation #3 is more general and primarily focused on monitoring the reporting ALDs. To some extent, Recommendation #3 has already been substantially addressed by the recently updated and approved Governing Board policy on NAEP achievement level setting. One remaining element related to Recommendation #3 is the development of a timeline and process for reviewing ALDs, along with prioritization for content areas beyond reading and math—a task that the Governing Board is now pursuing. An Achievement Levels Procedures Manual to

address the implementation of the policy will include details about the process for conducting these studies.

The Governing Board does not have direct responsibility for Recommendation #2. The NCES Commissioner makes the decision about the trial status and is not required to adhere to this NAS recommendation.

Regarding Recommendation #1, there are general policy definitions that apply to all NAEP assessments. These policy ALDs are elaborated into several different types of content ALDs under the revised Board policy. Additional clarity on the labels and uses of different types of ALDs should be described in the Achievement Levels Procedures Manual, including:

- Content ALDs developed with an assessment framework (generally by content area) are used to inform item development.
- Content ALDs that apply to a framework overall (across content areas) are used to conduct standard setting. These ALDs may be created as part of the framework development process or by re-convening framework panels (or similar individuals) after the assessment has been administered, prior to standard setting.
- Reporting ALDs, as described in the Board’s revised policy statement, will be created following administration of an assessment to communicate about what performance at each NAEP achievement level indicates about what students do know and can do.

Addressing Recommendation #1 should focus on the current reporting ALDs for mathematics and reading at grades 4, 8, and 12. The methodology will be similar to what was done to evaluate the alignment and revise the 2009 NAEP Reading ALDs for grades 4, 8, and 12 (Donohue, Pitoniak, & Beaulieu, 2010) and the 2009 NAEP Mathematics ALDs for grade 12 (Pitoniak, Dion, & Garber, 2010). This process will generate new reporting ALDs that comply with the revised Board policy statement. A potential additional step is to examine and/or document the alignment between the item pools and the NAEP frameworks, including information about the extent to which each NAEP administration faithfully represents the NAEP frameworks. Finally, alignment of cut scores can be evaluated using item maps, as part of the work to review and revise the reporting ALDs. Frameworks should be taken as a given; validation of the frameworks is beyond the scope of this work and evidence for their validity results from the Board’s framework development process.

The following table provides a draft of the activities, timeline and responsibilities for responding to Recommendations 1-3. Work will begin with reading and mathematics ALDs (based on 2019 data, to be used in reporting 2021 results). Reporting ALDs for other subjects will be reviewed and revised according to when they next appear on the NAEP Assessment Schedule. In accordance with Principle 4 of the Board policy on NAEP achievement level setting, reporting ALDs will be reviewed and revised on a regular basis (at least every 3 administrations or every 10 years, or when there is a major framework update). For example, the NAEP Mathematics and

Reading ALDs will need to be revisited following the 2025 administrations under the revised assessment frameworks.

Proposed Activity	Responsibility	Timeline
COSDAM approval of Achievement Levels Procedures Manual (described in policy statement)	COSDAM	May 2020
Conduct studies to examine and/or document alignment between NAEP Math and Reading Frameworks and item pools for grades 4, 8, 12	NCES	December 2020
Conduct studies to review and revise Math and Reading ALDs at grades 4, 8, and 12 Conduct studies to review and revise U.S. History, Civics, and Science ALDs at grade 8	NAGB	Contract awarded summer/fall 2020; complete by summer 2021 (reading/math); for other subjects the timeline will be determined by Assessment Schedule (ALDs updated in time for reporting of next administration after 2020)
Full Board action on revised Reporting ALDs	NAGB	Math/Reading at grades 4, 8, 12 (August 2021); for other subjects the timeline will be determined by Assessment Schedule (Board action will take place prior to release of results)
Conduct studies to examine and/or document alignment between NAEP U.S. History, Civics, Science, and TEL Frameworks and item pools for grade 8	NCES	December 2021

Responding to Recommendation #4 (Informative to the Public)

Recommendation #4: Research is needed on the relationships between the NAEP achievement levels and current or future performance on measures external to NAEP. Like the research that led to setting scale scores that represent academic preparedness for college, new research should focus on other measures of future performance, such as being on track for a college-

ready high school diploma for 8th-grade students and readiness for middle school for 4th-grade students.

Recommendation #4 is interpreted as articulating the need to provide context and relevance for NAEP results to show where NAEP fits in the constellation of other major assessments and external indicators of student achievement that are familiar to the general public, such as international assessments and indicators of postsecondary preparedness. Because NAEP and external indicators typically have different purposes, administration conditions, target populations, and other distinguishing characteristics, the purpose of this particular recommendation is not to make judgments about which results are “right” or “wrong” but to make the reporting of NAEP results more meaningful, useful, AND informative to the public.

This recommendation refers to both linking studies of NAEP and other measures of student achievement, as well as efforts to use NAEP to predict future performance. There are many different existing measures of student achievement, and we are aware of several efforts to link NAEP to various other measures, particularly in math, reading, and science. In order to consider what new studies might be pursued, it is important to better understand the resources that already exist, in addition to discussing how new efforts fit into the Board’s ongoing work and Strategic Vision.

The Governing Board’s work on reporting and dissemination includes the production of infographics and other descriptive reporting that describe student achievement in terms of several contextual variables. This work has typically been done using scale scores but could be expanded to include achievement level information, possibly including efforts to provide descriptive information about contextual factors associated with performance at the *NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced* achievement levels.

To address the issue of how best to synthesize and report information about how NAEP relates to other assessments and indicators, the Governing Board has commissioned a technical memo on recommendations for synthesizing relevant findings from multiple studies in ways that are informative to a general audience. The purpose of this effort is to explore how to place NAEP in a meaningful context of other familiar assessments and indicators, and to generate additional ideas. The following table provides a draft of the activities, timeline and responsibilities for responding to Recommendation 4.

Proposed Activity	Responsibility	Timeline
Review of technical memo on various ideas (including pros/cons) for synthesizing and representing findings about how other assessments and external indicators of student performance relate to NAEP (including a summary of existing linking studies) and what the findings mean for NAEP.	NAGB	Spring 2020
As the Governing Board works to develop its next Strategic Vision, deliberations will take place as part of that effort to determine how to approach the goal of making NAEP more relevant by connecting NAEP results to important real world indicators of student achievement.	NAGB	August 2020
Decision on additional studies that should be pursued to connect NAEP to other assessments and external indicators of student performance	NAGB/NCES	November 2020

Responding to Recommendations #5 & #6 (Reasonable, Valid, Informative)

Recommendation #5: Research is needed to articulate the intended interpretations and uses of the achievement levels and collect validity evidence to support these interpretations and uses. In addition, research to identify the actual interpretations and uses commonly made by NAEP's various audiences and evaluate the validity of each of them. This information should be communicated to users with clear guidance on substantiated and unsubstantiated interpretations.

Recommendation #6: Guidance is needed to help users determine inferences that are best made with achievement levels and those best made with scale score statistics. Such guidance should be incorporated in every report that includes achievement levels.

The *Standards* clearly indicate that any validation plan should begin with specifying the intended interpretations and uses of test scores. It is important to recognize that NAEP ALDs do not make claims about the achievement levels predicting performance on other current or future criteria (e.g., college readiness); however, strong claims *are* asserted about mastery of the content covered by relevant NAEP frameworks. Therefore, Recommendations #5 and #6 are related and should be considered together. The Governing Board is currently working on developing a statement of intended purpose and meaning for NAEP, which includes intended interpretations and uses for scale scores and achievement levels at a general level. The full Board discussed this document at the November 2019 Board meeting and is expected to take action during the upcoming March 2020 Board meeting (Appendix B). The Reporting and Dissemination Committee and COSDAM have provided initial guidance on an interpretative guide for the NAEP achievement levels.

After the Board reaches consensus about the intended interpretations and uses, the next step is to gather and document the evidence that exists related to those interpretations and to identify areas where additional evidence may be needed. This would take the form of building validity arguments to document the evidence that exists to support intended interpretations and uses; separate activities would be appropriate for supporting NAEP *scale scores* and NAEP *achievement levels*.

Gathering and summarizing validity evidence regarding interpretations of NAEP scale scores should primarily be a responsibility of NCES. This may be a matter of gathering and synthesizing documentation of existing NCES procedures that provided validity evidence for NAEP interpretations (e.g., qualifications of item writers, procedures for reviewing items, pilot testing, cognitive labs, etc.). This activity would also help to uncover areas where more research and evidence is needed.

Gathering and summarizing validity evidence regarding interpretations of NAEP achievement levels is a responsibility of the Governing Board (via COSDAM). Research undertaken to address Recommendation #1 should also provide evidence to address part of Recommendation #5, because the ALDs represent the intended meaning of NAEP achievement level categories.

In contrast to the established traditions for validating score meaning (e.g., the *Standards*), broadly endorsed procedures or criteria for gathering and evaluating evidence regarding score (or achievement category) use do not yet exist. Nonetheless, the interpretative guide contemplated by COSDAM and R&D would be one source of evidence to address Recommendations #5 and #6.

The Board recognizes that some stakeholders may hold misconceptions of the achievement levels. For example, legislators or education writers have sometimes confused performance at the *NAEP Proficient* level with grade-level performance. To respond to these misconceptions, we propose to work to create and provide materials and to conduct new outreach activities. The first step to addressing the misconceptions is to better understand how various stakeholder groups are interpreting and using the NAEP achievement levels. The NAS evaluation included reviews of existing materials and conversations with multiple audiences to begin to understand and articulate how various stakeholder groups are interpreting and using the NAEP achievement levels. The NAS evaluation did uncover several existing misconceptions about the NAEP achievement levels, and the Board will need to conduct additional work to more fully understand actual interpretations and uses of the NAEP achievement levels. We will need to develop and refine additional materials in formats most relevant to targeted audiences, (e.g., print, video, workshops) to address existing misconceptions and promote appropriate use. It would also seem desirable to engage in a companion evaluation effort to assess the effectiveness of these new materials and outreach activities.

Recommendation #6 (need for explicit guidance about when to use scale scores versus achievement levels) appears to be very narrow in scope, referring specifically to the

inappropriateness of using the percentage above a cut score to describe changes over time and across groups. To best address Recommendation #6, the interpretative guide should explicitly include information about which inferences are best made with scale scores versus achievement levels.

Effective communication of the NAEP achievement levels is an important aspect of Recommendations #5 and #6. There is a need to better understand how users interpret the policy definitions and ALDs for *NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced*. For example, what does “solid academic performance” mean, and is it possible to describe this educational goal more effectively?

Further development of these ideas (and others) will be needed to address these recommendations, and the staff plans to convene an ongoing advisory group on communication of NAEP achievement levels. The following table provides a draft of the activities, timeline and responsibilities for responding to Recommendations #5 and #6.

Proposed Activity	Responsibility	Timeline
Convene ongoing advisory group to discuss and provide feedback on the development of materials for communicating NAEP achievement levels	NAGB/NCES	Spring 2020 – Spring 2021
Collect information about current uses of NAEP achievement levels via focus groups and evaluate appropriateness of interpretations and uses that are not directly intended	NAGB	Spring/summer 2020
Adopt statement of intended purpose and meaning of NAEP (Appendix B)	NAGB	March 2020
Improve communications of what NAEP frameworks and achievement levels represent	NAGB/NCES	Ongoing
Develop and finalize interpretative guide for NAEP achievement levels; iterative drafts will be discussed by COSDAM and R&D	NAGB	Spring 2020-Spring 2021
Collect and document validity evidence to support intended interpretations and uses of NAEP achievement levels <ul style="list-style-type: none"> Collect and summarize validity evidence to support intended interpretations and uses of NAEP scale scores 	NAGB NCES	Spring 2021 Spring 2021

Responding to Recommendation #7 (Valid)

Recommendation #7: NAEP should implement a regular cycle for considering the desirability of conducting a new standard setting. Factors to consider include, but are not limited to: substantive changes in the constructs, item types, or frameworks; innovations in the modality for administering assessments; advances in standard setting methodologies; and changes in the

policy environment for using NAEP results. These factors should be weighed against the downsides of interrupting the trend data and information.

Recommendation #7 has been addressed by inclusion in the revised policy statement (Principle 4). It will be necessary to develop a process for carrying out a cut score review, but this should occur under COSDAM's purview as part of the development of the Achievement Levels Procedures Manual.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Donahue, P., Pitoniak, M., & Beaulieu, N. (2010). *Final report on the study to draft achievement-level descriptions for reporting results of the 2009 National Assessment of Educational Progress in reading for grades 4, 8, and 12*. Princeton, NJ: Educational Testing Service.
- National Academies of Science, Engineering, and Medicine. (2017). *Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress*. Washington, DC: The National Academies Press.
- Pitoniak, M., Dion, G., & Garber, D. (2010). *Final report on the study to draft achievement level descriptions for reporting results of the 2009 National Assessment of Educational Progress in mathematics for grade 12*. Princeton, NJ: Educational Testing Service.

National Assessment Governing Board’s Response to the National Academies of Sciences, Engineering, and Medicine 2016 Evaluation of NAEP Achievement Levels

Legislative Authority

Pursuant to the National Assessment of Educational Progress (NAEP) legislation (Public Law 107-279), the National Assessment Governing Board (hereafter the Governing Board) is pleased to have this opportunity to apprise the Secretary of Education and the Congress of the Governing Board response to the recommendations of the National Academies of Sciences, Engineering, and Medicine evaluation of the NAEP achievement levels for mathematics and reading (Edley & Koenig, 2016).

The cited legislation charges the Governing Board with the authority and responsibility to “develop appropriate student achievement levels for each grade or age in each subject area to be tested.” The legislation also states that “such levels shall be determined by... a national consensus approach; used on a trial basis until the Commissioner for Education Statistics determines, as a result of an evaluation under subsection (f), that such levels are reasonable, valid, and informative to the public; ... [and] shall be updated as appropriate by the National Assessment Governing Board in consultation with the Commissioner for Education Statistics” (Public Law 107-279).

Background

NAEP is the largest nationally representative and continuing assessment of what our nation’s elementary and secondary students know and can do. Since 1969, NAEP has been the country’s foremost resource for measuring student progress and identifying differences in student achievement across student subgroups. In a time of changing state standards and assessments, NAEP serves as a trusted resource for parents, teachers, principals, policymakers, and researchers to compare student achievement across states and select large urban districts. NAEP results allow the nation to understand where more work must be done to improve learning among all students.

For 25 years, the NAEP achievement levels (*Basic*, *Proficient*, and *Advanced*) have been a signature feature of NAEP results. While scale scores provide information about student achievement over time and across student groups, achievement levels reflect the extent to which student performance is “good enough,” in each subject and grade, relative to aspirational goals.

Since the Governing Board began setting standards in the early 1990s, achievement levels have become a standard part of score reporting for many other assessment programs in the US and abroad.

Governing Board Response

Overview

The Governing Board appreciates the thorough, deliberative process undertaken over the past two years by the National Academies of Science, Engineering, and Medicine and the expert members of the Committee on the Evaluation of NAEP Achievement Levels for Mathematics and Reading. The Governing Board is pleased that the report concludes that the achievement levels are a meaningful and important part of NAEP reporting. The report states that, “during their 24 years [the achievement levels] have acquired meaning for NAEP’s various audiences and stakeholders; they serve as stable benchmarks for monitoring achievement trends, and they are widely used to inform public discourse and policy decisions. Users regard them as a regular, permanent feature of the NAEP reports” (Edley & Koenig, 2016; page Sum-8). The Governing Board has reviewed the seven recommendations presented in the report and finds them reasonable and thoughtful. The report will inform the Board’s future efforts to set achievement levels and communicate the meaning of NAEP *Basic*, *Proficient*, and *Advanced*. The recommendations intersect with two Governing Board documents, the Strategic Vision and the achievement levels policy, described here.

On November 18, 2016, the Governing Board adopted a Strategic Vision (<https://www.nagb.org/content/nagb/assets/documents/newsroom/press-releases/2016/nagb-strategic-vision.pdf>) to guide the work of the Board through 2020, with an emphasis on innovating to enhance NAEP’s form and content and expanding NAEP’s dissemination and use. The Strategic Vision answers the question, “How can NAEP provide information about how our students are doing in the most innovative, informative, and impactful ways?” The Governing Board is pleased that several of the report recommendations are consistent with the Board’s own vision. The Governing Board is committed to measuring the progress of our nation’s students toward their acquisition of academic knowledge, skills, and abilities relevant to this contemporary era.

The Governing Board’s approach to setting achievement levels is articulated in a policy statement, “Developing Student Performance Levels for the National Assessment of Educational Progress” (<https://www.nagb.org/content/nagb/assets/documents/policies/developing-student-performance.pdf>). The policy was first adopted in 1990 and was subsequently revised in 1995,

with minor wording changes made in 2007. The report motivates the revision of this policy, to add clarity and intentionality to the setting and communication of NAEP achievement levels.

The seven recommendations and the Governing Board response comprise a significant research and outreach trajectory that the Governing Board can pursue over several years in conjunction with key partners. The Governing Board will implement these responses within resource constraints and in conjunction with the priorities of the Strategic Vision.

Evaluating the Alignment of NAEP Achievement Level Descriptors

Recommendation #1: Alignment among the frameworks, the item pools, the achievement-level descriptors, and the cut scores is fundamental to the validity of inferences about student achievement. In 2009, alignment was evaluated for all grades in reading and for grade 12 in mathematics, and changes were made to the achievement-level descriptors, as needed. Similar research is needed to evaluate alignment for the grade 4 and grade 8 mathematics assessments and to revise them as needed to ensure that they represent the knowledge and skills of students at each achievement level. Moreover, additional work to verify alignment for grade 4 reading and grade 12 mathematics is needed.

The report's primary recommendation is to evaluate the alignment, and revise if needed, the achievement level descriptors for NAEP mathematics and reading assessments in grades 4, 8, and 12. The Governing Board intends to issue a procurement for conducting studies to achieve this goal. The Governing Board has periodically conducted studies to evaluate whether the achievement level descriptors in a given subject should be revised, based on their alignment with the NAEP framework, item pool, and cut scores. The Governing Board agrees that this is a good time to ensure that current NAEP mathematics and reading achievement level descriptors align with the knowledge and skills of students in each achievement level category. In conjunction with the response to Recommendation #3, the updated Board policy on NAEP achievement levels will address the larger issue of specifying a process and timeline for conducting regular recurring reviews of the achievement level descriptions in all subjects and grades.

The Governing Board agrees strongly with the recommendation that, while evaluating alignment of achievement level descriptors is timely, it is not necessary to consider changing the cut scores or beginning a new trend line at this time. The NAEP assessments are transitioning from paper-based to digital assessments in 2017, and current efforts are focused on ensuring comparability between 2015 and 2017 scores. The Governing Board articulated this in the 2015 Resolution on Maintaining NAEP Trends with the Transition to Digital-Based Assessments (<https://www.nagb.org/content/nagb/assets/documents/policies/resolution-on-trend-and-dba.pdf>).

Recommendation #2: Once satisfactory alignment among the frameworks, the item pools, the achievement-level descriptors, and the cut scores in NAEP mathematics and reading has been

demonstrated, their designation as trial should be discontinued. This work should be completed and the results evaluated as stipulated by law: (20 U.S. Code 9622: National Assessment of Educational Progress: <https://www.law.cornell.edu/uscode/text/20/9622> [September 2016]).

Ultimately, the Commissioner of Education Statistics is responsible for determining whether the “trial” designation is removed. The Governing Board is committed to providing the Commissioner with the information needed to make this determination in an expedient manner.

Regular Recurring Reviews of the Achievement Level Descriptors

Recommendation #3: To maintain the validity and usefulness of achievement levels, there should be regular recurring reviews of the achievement-level descriptors, with updates as needed, to ensure they reflect both the frameworks and the incorporation of those frameworks in NAEP assessments.

The Board’s current policy on NAEP achievement levels contains several principles and guidelines for *setting* achievement levels but does not address issues related to the continued use or reporting of achievement levels many years after they were established. The revised policy will seek to address this gap by including a statement of periodicity for conducting regular recurring reviews of the achievement level descriptors, with updates as needed, as called for in this recommendation. The Governing Board agrees that it is important to articulate a process and timeline for conducting regular reviews of the achievement level descriptors rather than performing such reviews on an ad hoc basis.

Relationships Between NAEP Achievement Levels and External Measures

Recommendation #4: Research is needed on the relationships between the NAEP achievement levels and concurrent or future performance on measures external to NAEP. Like the research that led to setting scale scores that represent academic preparedness for college, new research should focus on other measures of future performance, such as being on track for a college-ready high school diploma for 8th-grade students and readiness for middle school for 4th-grade students.

In addition to the extensive work that the Governing Board has conducted at grade 12 to relate NAEP mathematics and reading results to academic preparedness for college, the Governing Board has begun research at grade 8 with statistical linking studies of NAEP mathematics and reading and the ACT Explore assessments in those subjects. This work was published while the evaluation was in process and was not included in the Committee’s deliberations. Additional studies in NAEP mathematics and reading at grades 4 and 8 are beginning under contract to the National Center for Education Statistics (NCES). The Governing Board’s Strategic Vision includes an explicit goal to increase opportunities for connecting NAEP to other national and

international assessments and data. Just as the Board's previous research related grade 12 NAEP results in mathematics and reading to students' academic preparedness for college, the Governing Board anticipates that additional linkages with external measures will help connect the NAEP achievement levels and scale scores to other meaningful real-world indicators of current and future performance.

Interpretations and Uses of NAEP Achievement Levels

Recommendation #5: Research is needed to articulate the intended interpretations and uses of the achievement levels and collect validity evidence to support these interpretations and uses. In addition, research to identify the actual interpretations and uses commonly made by NAEP's various audiences and evaluate the validity of each of them. This information should be communicated to users with clear guidance on substantiated and unsubstantiated interpretations.

The Governing Board's Strategic Vision emphasizes improving the use and dissemination of NAEP results, and the Board's work in this area will include achievement levels. The Governing Board recognizes that clarity and meaning of NAEP achievement levels (and scale scores) are of utmost importance. The Governing Board will issue a procurement to conduct research to better understand how various audiences have used and interpreted NAEP results (including achievement levels). The Governing Board will work collaboratively with NCES to provide further guidance and outreach about appropriate and inappropriate uses of NAEP achievement levels.

Guidance for Inferences Made with Achievement Levels versus Scale Scores

Recommendation #6: Guidance is needed to help users determine inferences that are best made with achievement levels and those best made with scale score statistics. Such guidance should be incorporated in every report that includes achievement levels.

The Governing Board understands that improper uses of achievement level statistics are widespread in the public domain and extend far beyond the use of NAEP data. Reports by the Governing Board and NCES have modeled appropriate use of NAEP data and will continue to do so. This recommendation is also consistent with the goal of the Strategic Vision to improve the dissemination and use of NAEP results. The Governing Board will continue to work with NCES and follow current research to provide guidance about inferences that are best made with achievement levels and those best made with scale score statistics.

Regular Cycle for Considering Desirability of Conducting a New Standard Setting

Recommendation #7: NAEP should implement a regular cycle for considering the desirability of conducting a new standard setting. Factors to consider include, but are not limited to: substantive changes in the constructs, item types, or frameworks; innovations in the modality for administering assessments; advances in standard setting methodologies; and changes in the policy environment for using NAEP results. These factors should be weighed against the downsides of interrupting the trend data and information.

When the Board’s achievement levels policy was first created and revised in the 1990s, the Board was setting standards in each subject and grade for the first time and had not yet considered the need or timeline for re-setting standards. To address this recommendation, the Governing Board will update the policy to be more explicit about conditions that require a new standard setting.

Board’s Commitment

The Governing Board remains committed to its congressional mandate to set “appropriate student achievement levels” for the National Assessment of Educational Progress. The Board appreciates the report’s affirmation that NAEP achievement levels have been set thoughtfully and carefully, consistent with professional guidelines for standard setting, and based on extensive technical advice from respected psychometricians and measurement specialists. The Board also takes seriously the charge to develop the current achievement levels through a national consensus approach, involving large numbers of knowledgeable teachers, curriculum specialists, business leaders, and members of the general public throughout the process. This is only fitting given the Governing Board’s own congressionally mandated membership that explicitly includes representatives from these stakeholder groups.

The Governing Board remains committed to improving the process of setting and communicating achievement levels. The Governing Board is grateful for the report recommendations that will advance these aims.

Reference

Edley, C. & Koenig, J. A. (Ed.). (2016). *Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress*. Washington, DC: The National Academies Press.

The Intended Meaning of NAEP Results

The primary purpose of the National Assessment of Educational Progress (NAEP), also known as the Nation's Report Card, is to measure the educational achievement and progress of the nation's students at established grades and ages in relation to the content of NAEP frameworks. NAEP results also enable comparisons of what representative students know and can do among states and jurisdictions, among various demographic groups, and over time.

The authorizing legislation for NAEP and the National Assessment Governing Board states that the purpose of the NAEP program is broadly to, "conduct a national assessment and collect and report assessment data, including achievement data trends, in a valid and reliable manner on student academic achievement in public and private elementary schools and secondary schools..." (Public Law 107-279, Section 303(b)(2)(B)). That legislation also prohibits NAEP from maintaining any system of personally identifiable information. Thus, NAEP assesses the educational progress of groups of representative students, not individuals.

The National Center for Education Statistics (NCES) develops numerical score scales for each NAEP subject. NAEP scale scores convey the degree to which students have mastered the content described in the NAEP assessment frameworks, with higher scores indicating greater levels of mastery.

The Governing Board is charged with setting NAEP achievement levels and has established [general policy definitions](#) for *NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced*. Percentages at or above achievement level cut scores indicate the percentage of students in a group who *meet or exceed* the knowledge and skills represented by specific content achievement level descriptions. These specific descriptions are found in the NAEP [assessment frameworks](#) and [reports](#). Additional information about the NAEP achievement level descriptions can be found in the Governing Board policy statement on [Developing Student Achievement Levels for NAEP](#).

NAEP results describe educational achievement for groups of students at a single point in time, progress in educational achievement for groups of students over time, and differential educational achievement and progress among jurisdictions and subpopulations.

There are several features of NAEP that distinguish it from many other assessment programs. For example:

- 1) NAEP produces results for the nation and participating states and districts. NAEP does not produce results for individual students or schools.
- 2) NAEP measures progress based on successive cohorts of students. NAEP does not produce results about the growth of individual students or groups of students over time.
- 3) NAEP results measure achievement and progress; however, NAEP results alone cannot indicate either why or how progress has occurred. Educational policies and practices that concur with NAEP progress may have caused this progress or been coincidental.
- 4) NAEP assessments are based on independent assessment frameworks developed through a national consensus approach described [here](#). NAEP frameworks do not represent any single state or local curricula.

Plans for Design of 2021 NAEP Reading and Mathematics Assessments

Currently, NAEP assesses students with a *single-subject, two-block design*, where each student receives two 30-minute cognitive blocks of items of a single subject (e.g., reading or mathematics) in addition to survey questions. In the 2021 reading and mathematics assessments, about 78% of the total sample will continue to take a single subject two-block assessment. The remaining 22% of the sample will be taking a longer assessment consisting of three 30-minute cognitive blocks of items. This is in preparation to transition to a three-block assessment design in the future for the purpose of increasing efficiency in our assessments by assessing most students in two subjects (e.g., reading and mathematics). NAEP's design will continue to feature its hallmark sampling matrix structure where students are given only a portion of content, but the new design will enable us to gather more psychometric information from each student. While the three-block design will result in longer testing for any given student sampled to participate in NAEP, the overall testing footprint in American schools will be significantly reduced. Specifically, the new design will have several notable advantages once we fully transition to a three-block design:

- About a third fewer students and as many as 2,500 fewer schools will be sampled for each assessment;
- We will spend less time in each school on any given day of testing; and,
- The number of sessions needed per day in schools will be reduced from two to one.

In addition, the three-block design will afford the possibility to report on the relationships between student performance on different subjects such as reading and mathematics.

The transition to a three-block design was discussed with the Board in the spring and summer of 2019 as part of the deliberations on the NAEP Assessment Schedule. More recently, the specifics of the 2021 design were discussed with COSDAM at the November 2019 Governing Board meeting. In this session, we will update COSDAM in terms of two changes that have taken place since then: the increase in sample size for the entire 2021 sample and the change in the configuration of the blocks for the 22% of the sample that will receive three blocks of items. We will provide final sample sizes at both national and state levels and discuss the implications of the final design and sample sizes on precision.

Draft Achievement Levels Procedures Manual

In November 2018, the Governing Board unanimously adopted a revised policy on [Developing Student Achievement Levels for NAEP](#), replacing a previous policy that had been in place since 1995. One of the major changes from the previous version of the policy was to focus on general principles for best practice and to remove procedural details from the policy document itself. The introduction to the policy states, “In conjunction with this policy the Board shall maintain a procedures manual to establish and document additional details about how this policy is to be implemented” (page 3).

The primary audience for the Achievement Levels Procedures Manual is Board staff and contractors. This document is not intended to be of interest to the general public although it does provide transparency for anyone who is interested in the implementation details of the Board policy. It also provides documentation of institutional knowledge of NAEP achievement level setting processes.

The Board has not previously maintained a procedures manual for achievement level setting. Material for this initial draft has been drawn from a variety of sources, including: existing literature on best practices in standard setting; past practice from previous NAEP achievement level setting activities; expert panels, literature reviews, and technical memos that were commissioned over the last few years to inform the revision of the Board policy; and procedural details that were removed from the previous policy. Much of this version was drafted by expert consultant Susan Loomis under subcontract to HumRRO, as part of their Technical Support contract to the Governing Board.

The Achievement Levels Procedures Manual will be finalized with approval from COSDAM (planned for May 2020) but is intended to be a living document; substantive updates will be made only with COSDAM approval. Because this work relates to implementation of an existing policy rather than creation of new policy, this document is not intended to be a full Board agenda item (unless an issue arises that COSDAM feels is worthy of full Board consideration).

The purpose of the March 2020 COSDAM discussion is to gather initial feedback and surface any substantive concerns. Suggestions for line edits can be sent to Assistant Director for Psychometrics Sharyn Rosenberg by email. Following the COSDAM discussion at the March quarterly meeting, an optional call will take place in late March or early April to discuss how the concerns have been addressed. A revised document will be included in the May 2020 Board materials for COSDAM approval.



ACHIEVEMENT LEVELS PROCEDURES MANUAL

**DRAFT for COSDAM Review
February 2020**

Introduction

Under provisions of the National Assessment of Educational Progress Authorization Act of 2002 (P.L. 107-279), Congress authorized the Governing Board to, develop, “achievement levels that are consistent with relevant widely accepted professional assessment standards and based on the appropriate level of subject matter knowledge” (Section 303(e)(2)(A)(i)(II)). To carry out this statutory responsibility, the Governing Board has had a policy statement on NAEP achievement level setting beginning in 1990.

In November 2018, the Governing Board unanimously adopted a revised policy on [Developing Student Achievement Levels for NAEP](#), replacing a previous policy that had been in place since 1995. The current policy establishes the following policy definitions for the NAEP achievement levels, as expectations of what students should know and be able to do:

NAEP Basic

This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for performance at the NAEP Proficient level.

NAEP Proficient

This level represents solid academic performance for each NAEP assessment. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter.

NAEP Advanced

This level signifies superior performance beyond NAEP Proficient.

The policy contains the following six principles for developing NAEP achievement levels:

Principle 1: Elements of Achievement Levels

Principle 2: Development of Achievement Level Recommendations

Principle 3: Validation and Reporting of Achievement Level Results

Principle 4: Periodic Review of Achievement Levels

Principle 5: Stakeholder Input

Principle 6: Role of the Governing Board

One of the major changes from the previous version of the policy was to focus on general principles for best practice and to remove procedural details from the policy document itself. The introduction to the policy states, “In conjunction with this policy the Board shall maintain a procedures manual to establish and document additional details about how this policy is to be implemented. As professional standards evolve and new consensus documents are

released, this policy and the procedures manual shall be updated to the extent that new professional standards require” (page 3).

This Achievement Levels Procedures Manual has been developed to describe procedural details of each policy principle, when necessary. Some principles (or subprinciples) do not seem to require additional procedural detail, such as Principle 6: Role of the Governing Board, which recaps information provided in other sections. The full text of the principles from the policy statement itself has been included in this document (indicated by grey highlighting). The elaboration of procedures appears below the relevant text from the policy statement.

As indicated in the introduction of the policy statement, the achievement level setting process is carried out by contractors selected through a competitive bidding process. The Governing Board’s Assistant Director for Psychometrics typically oversees this process and serves as the Contracting Officer’s Representative (COR). The statement of work (SOW) and contractor proposals need to be consistent with the information contained in the Achievement Levels Procedures Manual and in the policy document itself.

The National Center for Education Statistics (NCES) designates a liaison to work with the Governing Board COR. The NCES liaison works closely with the COR to provide data, materials, sample assessments, and other operational information needed to carry out the achievement level setting process. The NCES liaison coordinates necessary communication with NCES contractors and attends all meetings of the achievement levels panels and the Technical Advisory Committee on Standard Setting (TACSS).

Principle 1: Elements of Achievement Levels

The Governing Board is responsible for developing student achievement levels for each NAEP assessment. Achievement levels for each NAEP assessment consist of content achievement level descriptions (ALDs), cut scores that demarcate adjacent levels, and exemplar items or tasks that illustrate performance at each level.

- a) Content achievement level descriptions (ALDs) translate the policy definitions into specific expectations about student knowledge and skills in a particular content area, at each achievement level, for each subject and grade. Content ALDs provide descriptions of specific expected knowledge, skills, or abilities of students performing at each achievement level. Content ALDs reflect the range of performance that items and tasks should measure. During the achievement level setting process, the purpose of content ALDs is to provide consistency and specificity for panelist interpretations of policy definitions for a given assessment. During reporting, content ALDs communicate the specific knowledge and skills represented by *NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced* for a given assessment.

The policy statement uses the term “content ALDs” to distinguish the content-specific statements for a given assessment from the general policy definitions that apply to all NAEP assessments. “Content ALDs” is not a common term in achievement level setting more broadly. It is an umbrella term for several different types of ALDs, described below.

Policy definitions		The policy defines three NAEP achievement levels: <i>NAEP Basic</i> , <i>NAEP Proficient</i> , and <i>NAEP Advanced</i> . These policy definitions apply to all main NAEP assessments.
Content ALDs	ALDs in Framework (for item development and achievement level setting)	Under the revised policy and procedures for framework development, the framework committee may develop content ALDs both by content area (to inform item development) and overall (for use in the achievement level setting activities). The Framework Development Panel may determine that one set of ALDs can serve both of these purposes. These ALDs will continue to be written in terms of what students should know and be able to do. If there is a specific need to revise the overall ALDs in advance of an achievement level setting, then a separate activity will be undertaken to do so, but this is not intended to be necessary in most cases.
	Threshold/ Borderline ALDs (if applicable)	If descriptions of performance right at the cut scores are needed for the standard setting methodology (e.g., Bookmark), then threshold (or borderline) ALDs will be developed by ALS panelists. Panelists are typically told that the threshold ALDs are for their own use only and will not be reported with the NAEP results. The rationale for having the ALS panelists create threshold ALDs rather than providing them at the beginning of the process is that it is an important task to help ALS panelists fully internalize the ALDs. The transient nature of the threshold ALDs is intended to prevent panelists from spending undue amounts of time on minor edits and wordsmithing.
	Reporting ALDs	The policy calls for conducting a study following the first operational administration of an assessment (and again every 3 administrations or 10 years, whichever comes later) to revise the content ALDs for the purpose of reporting, using empirical data of student performance. The reporting ALDs will be written in terms of what students do know and can do.

- b) Cut scores mark the minimum threshold score, the lower bound, for each achievement level. Performance within a given achievement level begins at the cut score for that level and ends just below the cut score for the successive achievement level.
- c) Exemplar items or tasks, including student responses, illustrate student performance within each of the achievement levels. They provide specific examples to help the public better understand what students in each achievement level know and can do.

Principle 2: Development of Achievement Level Recommendations

The Governing Board shall develop student achievement levels for NAEP, consistent with relevant widely accepted professional assessment standards, based on the appropriate level of subject matter knowledge.

- a) A Design Document shall be developed at the beginning of the achievement level setting process, to describe in detail the scope of the achievement level setting project being undertaken, including but not limited to all planned materials, procedures, and analyses needed for the project. The Design Document shall be posted for public review with sufficient time to allow for a response from those who wish to provide one.

Purpose of the Design Document

The purpose of the Design Document is to provide a detailed description of each aspect of the proposed achievement level-setting process. The Design Document serves as the guide for the project, and it is the document against which the implementation of procedures is compared and evaluated. The Design Document must be submitted for approval early in the project (typically within 30-60 days of contract award) in order to guide developments throughout the process. Modifications to procedures require modifications to the Design Document.

Content of the Design Document

The Design Document elaborates on the proposed procedures and must clearly describe the key aspects to be implemented for the entire project. Each aspect of the process required in the statement of work issued for the procurement must be addressed in the Design Document, as well as any additional features proposed for the project.

Each component of Principle 2: Developing Achievement Level Recommendations must be described in the Design Document. The purpose of each step in the achievement level setting (ALS) process, the personnel engaged in each step, materials and resources required, and timelines for each must be described in detail sufficient to clearly convey an understanding of the process to be implemented. A draft agenda must be provided for each component of the project for which a panel is to be convened.

Required panel studies to be detailed in the Design Document include the pilot study and the operational ALS panel study. If additional research is required prior to finalizing the design of the ALS process, these studies should be conducted as field trials. Field trials must be conducted prior to the pilot study, and the field trials must be fully described in the Design Document. The research must be completed prior to the pilot study to help assure that the pilot study can be conducted according to the design for the operational ALS.

Validity research studies must be detailed in the Design Document, and a clear rationale must be provided for each study, along with procedures for collecting data and implementing studies.

A complete project schedule presented both by type of study and chronologically must be included in the Design Document.

Process of Design Document Review

The COR for the ALS process will review drafts of the Design Document and coordinate the process of review, modification, and finalization of the document. In coordination with the COR's reviews, the Design Document will be shared for review and evaluation by the Technical Advisory Committee for Standard Setting (TACSS). After modifications to meet the recommendations of the TACSS have been incorporated, the COR will share the Design Document with the Committee on Standards, Design and Methodology (COSDAM) for their evaluation.

When the Design Document has met the general approval of COSDAM, it will be distributed to key individual and organizational stakeholders and users of NAEP data for review and comment. The review process must have ample publicity to produce broad-based reviews and comments, and the review process must provide ample time for that purpose (typically at least 30 days). A variety of methods of providing review comments should be made available. Recommendations collected through this review will be evaluated by the ALS contractor, the COR, and TACSS to determine additional modifications to finalize the Design Document.

The final version of the Design Document will be presented to COSDAM for formal approval. Modifications to the design require revisions to the Design Document.

- b) The development of content achievement level descriptions (ALDs) shall be completed initially through the process that develops the assessment frameworks. (See the Governing Board Policy on Framework Development for additional details). The Board may then review and revise content ALDs to advance the purposes they serve, whether that is guiding an achievement level setting or informing the public about the meaning of achievement levels. Whether revised or not, the ALDs that guide achievement level setting shall be articulated in terms of what students *should know and be able to do*. There shall be no content ALDs developed for performance below the *NAEP Basic level*.

If New Content ALDs Must be Developed For Achievement Level Setting

Typically the content ALDs for use in achievement level setting will be developed as part of the framework development process, as outlined in the Board policy on [Framework Development](#) and its accompanying procedures manual. In some rare cases, it may be necessary to revise the content ALDs that were created during the framework development process for use in achievement level setting. For example, the content ALDs would need to be revised if one aspect of the framework could not be operationalized and the ALDs refer to knowledge and skills that are not represented by the item pool. Sometimes the use of content ALDs in a field trial or pilot study surfaces concerns that

were not anticipated in advance. There may be other situations that arise to threaten the utility of the content ALDs for standard setting, as identified by the Assistant Director for Psychometrics, COSDAM, and/or the achievement levels contractor.

If it is necessary to revise the content ALDs for use in achievement level setting, it is desirable to include some members of the Framework Development Panel to conduct this work. The number of persons involved in development of the content ALDs will depend upon the number of grade levels involved, but a minimum of three content experts per grade is advised.

The content ALDs must follow best practices for developing performance level descriptions:

- The ALDs must describe measurable attributes and not attitudes or behaviors of students.
- Calibration of the ALDs to distinguish performance at one level from that at another should not include ambiguous terms that are subject to individual interpretation such as few, some, often, seldom, and rarely.
- ALDs should be succinct descriptions of the key knowledge, skills, and abilities that describe the performance of students at each level of achievement, relative to the framework features and in alignment with the policy definitions.
- To the extent feasible, the ALDs should describe the same performance attributes across the three levels of achievement. Performance is assumed to be cumulative across levels such that higher levels subsume performance described at lower levels. If the level of performance does not change for a higher level, there is no need to repeat the description.

Key factors for the evaluation of ALDs

- Alignment of ALDs to key aspects of the assessment framework
- Alignment of ALDs to policy definitions of each achievement level: NAEP Basic, NAEP Proficient, and NAEP Advanced
- Alignment of ALDs within each achievement level across grades
- Alignment of ALDs across achievement levels within each grade

Public Comment Collection

If new content ALDs are developed for achievement level setting, they should be shared for a broad-based review by key stakeholders in the content area and users of NAEP achievement levels and data. The review should focus attention on the alignment factors listed above and invite additional comments regarding the clarity and usefulness of the statements. Review comments will be evaluated by the ALD development panel to determine additional changes deemed necessary.

Review and Approval by COSDAM for Use in the ALS Process

If new content ALDs are developed for achievement level setting, they will need to be approved by COSDAM for use in the ALS process. After the ALD development panel has incorporated feedback from public comment, the ALDs will be presented to COSDAM for review and approval. Additional revisions may be recommended by COSDAM before approval for use in the ALS process. Initial approval by COSDAM is provisional, for use in

all panel studies of the ALS process. In accordance with the policy (Principle 3g), final and official approval by the Governing Board for reporting purposes is determined after the assessment has been administered and reporting ALDs are created based on empirical data.

Additional revisions may be needed as a result of panel studies conducted in preparation for the operational ALS process. If modifications impacting the calibration of the ALDs become necessary, the revised ALDs will again be presented to COSDAM for provisional approval, based on recommendations of content experts.

c) An achievement-level setting panel of subject matter experts shall be convened to recommend achievement level cut scores and exemplars.

- i. Each panel shall reflect diversity in terms of gender, race/ethnicity, region of the country, urbanicity, and experience with students with disabilities and English language learners. To ensure that they are qualified to make the judgments required by the achievement level setting process, individual panel members shall have expertise and experience in the specific content area in which the levels are being developed, expertise and experience in the education of students at the grade under consideration, and a general knowledge of assessment, curriculum, and student performance.
- ii. Each panel shall include teachers, non-teacher educators, and other interested members of the general public with relevant educational background and experience. Teachers shall comprise the majority of the panel, with non-teacher educators (e.g., curriculum directors, academic coaches, principals) accounting for no more than half the number of teachers. The remaining panelists shall be non-educators who represent the perspectives of additional stakeholders representing the general public, including parents, researchers, and employers.
- iii. The size of the panels shall reflect best practice in standard setting and be operationally feasible while being large enough to allow for split panels. Most NAEP achievement level settings have historically included approximately 20-30 panelists per grade, divided into two comparable groups.

Selection of Panelists

The selection of panelists is of critical importance to the success and validity of the ALS process. The process must be systematic, replicable, and transparent. A nomination process should be used to identify well-qualified individuals who are broadly representative of a variety of demographic characteristics and professional credentials to serve as achievement level setting panelists. First, it is necessary to identify individuals who are in a position to know persons having the qualifications required for ALS panelists.

While not a requirement, it may be useful to draw a nationally representative sample

using principles of representative sampling to identify geographical units, such as states, cities, or school districts to serve as the basis of representation. Private schools should be included in the recruitment and identification of educators. From these sampled units, individuals holding specific positions within the content area may be identified to serve as nominators of panelists.

Nominators should be provided with guidelines regarding the requirements of panelists and the credentials needed to serve as panelists. Nominators that meet the qualifications for panelists may self-nominate.

Representativeness of Candidates Recruited

The demographic characteristics to be represented on the achievement level setting panels are specified by the policy, which requires diversity on each panel. Appropriate goals for diversity may be proportional to national population distributions. For example, the number of private school panelists may be based on the proportion of accredited private school teachers in the nation in the subject area. The representativeness of these characteristics should be met for each grade level panel in the ALS process—not simply across all grade-level panels.

Extra panelists should be recruited to avoid a shortfall in meeting the distributional targets for the panels, as well as the targeted number of panelists. In recognition of the uneven distribution of some demographic characteristics across different content areas and grade levels, however, the representativeness of some aspects may vary by grade and subject.

Appropriate Credentials of Candidates Recruited

From the pool of nominees, those with the most outstanding content and education credentials should be given highest priority for selection as panelists. All panelists must have educational training and experience in the content of the subject assessed, direct experience with students at the grade level for which they are to serve as panelists, and a general knowledge of assessment, curriculum, and student performance. The specific credentials required will vary by type of panelist, and they must meet the requirements of the policy.

Teachers: Teacher panelists must currently teach in the grade and subject for which they are to serve as an ALS panelist. A minimum of five years of teaching with two years of teaching in the grade and subject is required for NAEP ALS teacher panelists. Teachers who have won teaching awards or other professional recognition should be given priority consideration for selection as panelists.

Non-teacher educators: Non-teacher panelists are educators who are not teaching in the K-12 education system, although candidates need experience or training in the subject area and grade level range for which they are to serve as a panelist. Curriculum directors in the subject area at the state, regional, district, or school level and other such educators typically have credentials appropriate to serve as panelists in this category. In addition, post-secondary educators who train teachers in the content and grade level are also candidates for this role.

General public: Members of the general public who have an educational background and/or training and work experience in the subject area and who have direct experience with children in the grade level are eligible to serve as panelists. Retired or former educators who spent a majority of their working career as educators are not eligible to serve as representatives of non-educators.

Requisite Composition of Panels

The policy provides general guidance for the representation of each panelist type in the ALS process: the majority of panelists are to be teachers; non-teacher panelists are to be no more than half the number of teachers; and, while the general public is to be represented, there is no set requirement for the number or proportion of non-educators. The distribution of panelists by type applies to each grade level in the ALS process.

Demographic characteristics should be distributed approximately equally within each grade level in the ALS process. For some subjects, the distribution by gender varies by grade level, and that must be acknowledged in the distribution across grade levels. For some characteristics, such as geographic region, representation across all the grades may be sufficient.

Drawing Panels and Assigning Panelists to Groups

A simple coding scheme (with 3-5 categories or levels) of candidates' credentials may facilitate the process of selecting outstanding candidates and assuring representation on each panel with respect to demographic characteristics. A computerized algorithm to maximize selection of outstanding panelists within panelist type while meeting specified proportional constraints on panelists' demographic characteristics can ease the process of selecting panelists. Additionally, to assure the sense within panels that they are "broadly representative," it is advised that no more than one panelist from the same school or district serve on a grade-level panel.

Selection of panelists for each type of ALS panel

NAEP ALS panels are generally larger than those for state standard setting. The larger number of panelists is related to the requirements for broad-based and national representation of panels, different types of panelists, requirements for statistical precision, and the large size of NAEP item pools.

- Thirty panelists have typically been recruited for each grade-level panel in the operational ALS process.
- Twenty panelists have typically been recruited for each grade-level pilot study panel.
- The number of panelists for each grade level in a field trial depends upon the purpose of the study. NAEP panel studies typically require at least ten panelists. The study design, along with advice of the TACSS, will determine the exact number required for field trials.

The composition of the field trial panels is typically more flexible, but the requirements should be determined by the purpose of the study. A nationally representative panel is generally not required, although representation by panelist type and demographic characteristics is typically advised.

Split panels

In addition to the grade level panels, the policy calls for split panels. The purpose of split panels is to lessen the burden of the judgment task for panelists given the large item pools for NAEP. The split panel design reduces the amount of time required for the judgment process, and it lessens the potential for panelist fatigue. In addition, the split panel design provides the opportunity for comparisons of results, albeit limited, between the split panel groups.

Both the panels and the assessment items to be judged in the ALS process are split. Two subpanels should be sufficient for each grade level in most ALS procedures. The combination of ALS methodology for collection of judgments and the number of item judgments to be made may require more subpanels. Both panelists and item pools should be divided so that each group of panelists and each set of items is as equivalent as possible. Equivalence of panelist groups is numerical, by panelist type and by demographic characteristics. Equivalence of the item pools is numerical, by item format/type, item difficulty, and content framework designation in the assessment, for example. A subsample of items should be included in the pool assigned to each panel group in order to have common items for which judgment comparisons can be made. The results of the first round of judgments can be evaluated as resulting from roughly replicate panels.

Table groups

ALS procedures should be implemented with 4-6 panelists assigned to each table group in each grade panel to provide the opportunity to have small discussion groups. The goal is to assign panelists to table groups so that each group is equivalent with respect to panelist type and as equivalent as possible with respect to demographic characteristics. There should be at least one representative of each panelist type in each table group.

- d) Panelists shall receive training on all aspects of the achievement levels setting process to ensure that panelists are well-prepared to perform the achievement level setting tasks required of them. Panelists shall be instructed that their role is to make achievement level recommendations to the Governing Board. Training shall include but not be limited to: the purpose and significance of setting achievement levels for NAEP; the NAEP assessment framework for the given subject area; and administration of a sample assessment under NAEP-like conditions that students experience. It is important for panelists to arrive at a common conceptualization of *NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced* based on the content ALDs. Panelists shall be trained on each element of the judgmental task they perform, including the selection of exemplar items. They should be led by capable *content facilitators* (who are content experts and have previous experience with achievement level setting) and *process facilitators* (who have background in standard setting and experience leading panelists through the achievement level setting process). Facilitators shall take a neutral stance and not attempt to influence panelist judgments.

While standard setting is a process based on judgments, the judgments must be well informed. Panelists must be clear that their informed judgment, not their opinion, is

required. Training in all aspects of the process is necessary to assure that panelists have a clear understanding of each part and are trained with the knowledge and skills needed to make informed judgments. Training should be provided as an iterative process and with a mix of plenary (whole group) sessions and grade-level sessions. The training modules should build on one another so that there is appropriate repetition and reinforcement throughout the process. The timing of information and the amount of information shared at a given time are both important considerations in training panelists for the NAEP ALS process.

Advance Materials

Advance materials should be provided to panelists to begin the training process by informing panelists about what they will be doing when the panels are convened. Once notified of their selection to serve on an ALS panel, panelists need to have communications that provide assurance that the process is well organized, in addition to information that starts their training for the ALS process. Advance materials should be designed to provide the following key types of information. Information communicated in different formats and through different modes is advised.

Purpose of standard setting: A clear statement of the purpose of standard setting.

Framework document: Describe the role of the framework for the development of the assessment and provide instructions regarding the focus of their attention to prepare for the ALS process.

Policy definitions: Describe the central role that the policy definitions play in the NAEP ALS process.

Achievement level descriptions (ALDs): Describe the relationship of the ALDs to the policy definitions and the central role that the ALDs play both in the ALS process and for reporting student performance on NAEP.

Overview of process and description of each step: A user-friendly video is recommended as an engaging way of introducing the process to panelists. A briefing booklet may also be used to provide the overview and detailed information about the process in advance of the panel meeting.

Draft agenda for the panel meeting: A draft agenda must provide enough specific information to convey the activities to be accomplished each day and the relative emphasis on each aspect of the process. Panelists need to know in advance that they will have long days filled with training and standard setting activities and that attendance at and participation in each session is mandatory.

Details of travel and lodging arrangements: Panelists travel from throughout the U.S. to participate in the NAEP ALS panel meetings. Panelists must have information early enough and in sufficient detail to feel confident about travelling alone to a place that they have perhaps never visited and for a procedure that they have never experienced.

Format of Meetings

The format of the panel meetings should be designed to facilitate thorough training as well as to provide some variety in activity and setting.

Whole group/plenary sessions: These sessions include panelists from all three

grades assessed by NAEP: 4th, 8th, and 12th. The whole group sessions are designed to provide the initial training in each key step of the process. The purpose of whole group sessions is to increase standardization across the grade groups by assuring that everyone hears the same information. If only one grade level is involved in the standard setting, there is no need for a whole group session.

The overview of the process is presented in the opening whole group session, and this provides information about the NAEP program, the National Assessment Governing Board, the National Center for Education Statistics (NCES), and the various contractors that are involved in the NAEP program.

If the achievement level-setting process is being conducted to develop new achievement levels to replace those developed for a previous framework and assessment of the same subject, the Governing Board COR should provide the previous cut scores and data in the opening session and explain how and why they are not to influence judgments in this ALS process.

Training in the key steps of the process should be provided first in the whole group sessions. The presentations in the whole group sessions should be aimed at describing the purposes and uses of each aspect of the ALS process, and they should be designed to help panelists know how activities fit into the overall process of developing NAEP achievement levels.

An overview of the NAEP assessment framework should be provided in a whole group session on the first day of the process. The presentation helps panelists understand clearly how the framework is organized and why specific aspects are and are not included. The content ALDs should also be presented and the process for their development should be described so that panelists are assured that these have been carefully crafted by content experts. The relationship between policy definitions and content ALDs should be made clear in this session scheduled in the early part of the process.

In order to avoid the influence of cut scores and performance data from previous ALS procedures, as well as to avoid the influence of cut scores and performance data across grade panels in an on-going ALS procedure, the use of different scales for reporting feedback in each grade and for each panel study in the ALS process has been effective.

In addition to the whole group sessions scheduled to introduce new procedures in the process, a final, closing session should be scheduled to thank the panelists for their service and provide responses to any last minute questions or concerns. Panelists must leave with a clear notion that their contributions are valued.

Grade group sessions: Instruction and training in the details of the procedures should be provided in grade group sessions. These are the working sessions. Panelists should be assigned to table groups where computers and materials for implementing the process are provided to each panelist. While panelists work in table

groups and have discussions in table groups, training and instruction should always involve discussion by the grade group as a whole. It is vital that a common understanding and general agreement be reached by all grade level panel members.

Facilitators: In addition to the lead facilitator for whole group sessions, grade group facilitators should include both process facilitators and content facilitators. The pair should work together as a team in each grade group, but the process facilitator takes the lead in training and instruction. All facilitators must have training and experience in standard setting.

Process facilitators should be well trained in the ALS methodology to be implemented and experienced in leading standard setting panels. They must be both skilled at working with people and skilled in statistics and, preferably, psychometrics. The NAEP program is complex, and a strong background in quantitative methods and analysis is necessary for the process facilitator.

A facilitator guide should be developed to include all instructions and information to be presented to panelists. This helps to assure that the process is implemented in a standardized manner across the different panel meetings and grade levels. The facilitator guide should be the basis for training facilitators for the specifics of the ALS process, and the facilitators should be made aware that they are to follow the guide. The guide must emphasize that facilitators are to present a neutral position and that their role is not to persuade the panelists.

Content facilitators ideally should be selected from among the members who participated in the Framework Development Panel. The subset of Framework Development Panelists who worked on developing the ALDs represent those most appropriate to serve in this role for the ALS process, if possible. It is helpful for content facilitators to have extensive experience with the NAEP framework and ALDs to add to their authority in leading the panelists in content matters.

Content facilitators should provide training in the framework and ALDs, and they should lead the work with the panelists to develop the borderline ALDs and other aspects of the process that involve assessment content.

Observers: Observing the ALS process should not be open to the public due to the need to maintain the security of the assessment material and the requirement that NAEP achievement levels be released only by the NCES Commissioner. Approved observers should be strongly urged to attend all sessions throughout the process in order to understand how the process works and to have that understanding be reliable and accurate. Observers should be seated at specific tables reserved for them; they should not sit with panelists in the meeting room. Observers must take care not to cause any distractions or disturbances to the process. Observers should be encouraged to engage in social conversation with panelists, but they should be instructed not to discuss the process with panelists.

Observers typically include key staff of the Governing Board and NCES, and members

of the Governing Board and the TACSS. The inclusion of other observers should be at the discretion of the COR in order to assure that those with an interest and need to observe the ALS process are included while also assuring that the number of observers is not so large as to be distracting for panelists.

- e) The achievement level setting method that generates cut score recommendations shall have a solid research base and be appropriate for the content area, item types, number of items, scoring rubrics, and mode of administration, as applicable.

Criteria Regarding the Choice of Methodology for Achievement Level Setting

Solid research base: A solid research base has been a requirement throughout the history of NAEP achievement level setting. Research studies should be conducted to try out new methodologies or modifications of existing methodologies.

Appropriate for item types: Holistic methodologies are not appropriate for assessments with mixed item formats and they would not be practical for an entire assessment of dichotomous items because it would be difficult to form a holistic judgment over a large number of discrete items. Procedures requiring item-by-item judgments seem most appropriate for assessments with many discrete items.

Number of items: Assessments with a large number of items require a large number of item judgments which can lead to fatigue and perhaps judgment error. The methodology used with a large number of items, such as is typical for NAEP, must be easy to implement and use. Even complex data presented in a graphic format can be successfully incorporated into methodologies, such as the Mapmark method, as feedback to inform judgments of panelists.

Scoring rubric: Panelists must understand the scoring rubrics and how they are applied. There is no direct, one-to-one relationship between scoring rubrics and ALDS, however. Some NAEP assessments use clusters of items, and alternative combinations of responses for scoring. This requires an ALS procedure that can accommodate such judgments.

Mode of administration: NAEP has transitioned from paper-and-pencil administration to digital administration. The ALS procedures implemented with the digital assessments should also be computerized. The panelists must be able to experience the assessment as students experienced it, and the methodology for collecting their judgments of the knowledge, skills, and abilities required for a correct response must be consistent with the administration mode.

Lessons Learned About the Choice of NAEP ALS Methodology

In addition to the criteria specified in Principle 2e, some additional considerations should be taken into account:

- The ALS methodology must be consistent with the NAEP scaling methodology. An ALS procedure that allows for conjunctive judgments is not consistent with the

NAEP compensatory scaling model.

- The methodology must be easy for panelists to understand and use. The relationship between judgments and cut scores should be clear and easily understood. Similarly, the way to modify judgments and subsequent cut scores must be easy to understand and implement by panelists.
- Computerized methodologies generally require much less time because computation of results and feedback is much faster.
- Implementation of the ALS methodology must be efficient. The methodology with the least impact on resources—time, labor, materials—should be selected, other considerations being equal.

Quality Control Procedures

It is critical that quality control measures be in place given the large number of points where mistakes may occur throughout the process. For data entry, it is preferable that panelists enter their judgments directly into a computerized system to reduce manual errors of entry; however, if any data are entered manually, they should be 100% verified using a double-entry, cross-checking procedure. Computerized entries still need to be verified to confirm that there are no out-of-range or out-of-sequence data.

Software programs designed to complete analyses on the judgment data must be run with simulated data in advance of the panel meetings to de-bug and provide quality control. The software programs should detect logical errors and other kinds of problems that could result in incorrect results being generated. During the panel meetings, two data analysts should independently run all analyses on-site and verify that they produce the same results before feedback is shared with panelists.

Following the conclusion of the panel meetings, the NAEP operations contractor should confirm that the final cut scores have been mapped onto properly weighted and equated scales, before achievement level setting results are communicated to the Board.

- f) Evaluations shall be administered to panelists throughout the achievement level setting process, in accordance with current best practices. Evaluations shall be part of every major component of the process, and panelists shall be asked to confirm their readiness for performing their tasks. Evaluation data may be used for formative purposes (to improve training and procedures in future meetings); summative purposes (to evaluate how well the process was conducted and provide procedural validity evidence); and to inform the Governing Board of any relevant information that could be useful when considering cut score recommendations. The panelists shall have an opportunity to indicate to the Board whether they believe the recommended cut scores are reasonable.

Purposes and Uses

Evaluations should be administered for formative purposes to collect information that can be used to improve upon the timing, content, and format of information and training provided to panelists. Evaluations should be administered for summative purposes to ascertain whether the design and implementation of the procedures were effectively implemented and successful in accomplishing their purpose. Information collected

through evaluations is essential for establishing procedural evidence. Evaluations provide information to the Governing Board to inform their deliberations for setting the achievement levels.

Schedule of Evaluation Administrations

Evaluations should be administered throughout the achievement level setting process for each type of panel study in order to capture the opinions and attitudes of panelists at key points. Evaluations should be reviewed by the process facilitator each day to ascertain if any panelist is struggling with any aspect of the process. An identification code is recommended for evaluations to maintain confidentiality while allowing the facilitator to identify panelists in need of one-on-one help.

To the extent feasible and appropriate, a common set of questions should be asked without modification for each panel meeting. Similarly, when feasible and appropriate, the same questions included in previous ALS procedures should be asked without modification in order to make comparisons and to have a base for judging the relative success of a specific ALS panel meeting.

Panelists should be asked to confirm their readiness to perform key judgment tasks in the process:

- Prior to judgments regarding student performance at each level of achievement
- Prior to selection of exemplar items
- Prior to recommendations regarding the final cut scores and performance data

Panelists should be administered an evaluation at key points throughout the process to focus on a specific step and preparation for that step. The agenda and timing of specific steps will determine whether separate evaluations are necessary for steps. For example, it may be sufficient to collect information about the judgment round and feedback information in a single evaluation. Evaluations are recommended for the following steps in the process:

- At the end of the first day of the process to evaluate training and instruction
- After completing training in ALDs and development of borderline descriptions
- Following the first round of judgments
- Following the first round of feedback
- Following each round of judgments and feedback
- Following selection of exemplar items to recommend for reporting

In addition, panelists should be asked to evaluate the final cut scores and student performance data and make recommendations to the Governing Board regarding these data—including suggested changes or modifications recommended. A “consequences data questionnaire” has typically been used for collecting this information.

A final evaluation of the entire process should be administered at the completion of the process. Panelists should be asked clear and straightforward questions about their cut score recommendations.

Types of information to be collected in evaluations will vary to some extent according to the specific methodology used for setting achievement level cut scores. Further, the specific information collected will likely be more in-depth with regard to procedures that have not previously been used for a NAEP ALS process. The following should serve as general guidelines:

- Keep evaluations as brief as possible
- Maintain comparability of evaluation data from previous ALS procedures when feasible
- Collect key information consistently across the rounds of judgments
- Use a variety of ways to collect information
- Avoid statements and questions for which the response is highly predictable or likely to show little disagreement
- Avoid ambiguity
- Avoid questions that require self-evaluations of confidence or competence

g) In accordance with current best practices, feedback shall be provided to panelists, including “impact data” (i.e., the implications of their selected cut scores on the reported percentages of students at or above each achievement level).

Feedback is a key component of a standard setting process. Feedback is generally based on a combination of panelists’ judgments and student performance. The understanding of the relationship between panelists’ judgments and student performance must help the panelists evaluate how well performance on the assessment, relative to their judgments, represents the performance required in the policy definitions and achievement level descriptions.

NAEP ALS procedures should include a variety of feedback designed to better inform the panelists’ judgments during the process. A variety of feedback is helpful for providing a clear understanding of performance and for informing the judgments of performance relative to the ALDs. Providing an additional type or format of feedback at each round of judgments helps to provide variety as well as to manage the burden of new information. The particular types or formats of feedback provided will differ to some extent with different standard setting methodologies.

Purposes of Types of Feedback

Group-level cut scores and variability data should be provided as feedback for each round. The primary outcome of the ALS process is recommended cut scores, and the cut scores resulting from the ALS process are to be recommended to the Governing Board for use in reporting NAEP results. Panelists should be given information about where their cut scores fall at each round of judgments. The cut scores, per se, do not provide great insights into the relationship between their judgments of student performance and the statements of what students should know and are able to do; but the remainder of the feedback and discussion in preparation for subsequent rounds of judgment should be based on the cut scores. The median typically should be used as the cut score unless there is a compelling rationale to use a different statistic, since the median is not sensitive to

outliers. During the ALS process, the cut score feedback should be based on different score scales for each grade to avoid any attempts to adjust cut scores to match across grades. When multiple grade levels are involved in a NAEP ALS process, the cut score feedback should be shared across grade groups.

Inter-rater data should be provided as bar graphs to show panelists the distribution of cut scores at each achievement level for each panelist in the group. Panelists should be able to evaluate the location of their cut score at each achievement level with that of other panelists in the group. The distribution should also show any overlap in cut scores at adjacent achievement levels. These data are intended to help panelists understand that variability in the distribution of their cut scores represents a lack of general agreement regarding the minimal performance required to reach each level of achievement. Panelists should be able to trace the pattern of inter-rater consistency data across rounds of feedback to see how the consistency of their judgments and the level of agreement regarding required performance change across rounds.

An inter-rater consistency exercise should be implemented as a second type of inter-rater feedback after the first round of judgments, particularly for holistic procedures having few items. A list of items for which judgments showed least agreement should be provided to panelists for discussion. Several ways of showing a lack of agreement should be used to select approximately 10 items for discussion, including:

- Items for which judgments are closest to a 50-50 split between two achievement levels
- Items for which judgments are generally spread across all three achievement levels
- Items for which judgments are largely split between two non-adjacent levels

Panelists should be given the list and data for the items and asked to discuss their own judgments for the items. The discussion should clarify their understanding of the performance required for the item in relation to the ALDs. Through this discussion, panelists should enrich and strengthen their common understanding of performance at each level of achievement.

Impact data should be provided to panelists as a reality check to help them evaluate whether their judgments seem realistic in light of both the ALDs and student performance on the assessment. Although their initial discussion of the data may be challenging, panelists should focus on the comparison of their judgments, based on their understanding of the ALDs, relative to student performance. They should then evaluate whether any modification either to their understanding of the ALDs or to their cut score is in order. Panelists who have been well trained in the ALDs are generally committed to giving priority to their common understanding of the meaning and interpretation of the ALDs relative to student performance.

Presentation of Feedback

Each type of feedback should be distributed separately—not all at once. It is especially important that panelists not have access to a new type of data before they have been

instructed in its use. Panelists must have sufficient time to understand the feedback and discuss it with others. This is especially important for the initial presentation and discussion of impact data.

Guidelines in Provision of Feedback

- The goal of providing feedback is to inform the judgments of panelists regarding student performance relative to the content achievement level descriptions.
- The quantity of feedback should be sufficient to assure that panelists feel confident about making judgments; it should not be overwhelming.
- The data should be clear and concise. Panelists must understand how to use feedback data in order to use it. They should be instructed in how to incorporate the feedback information into their judgments to modify their cut scores.

The following recommendations are based on previous research conducted during NAEP achievement level setting activities:

- Cut score data should be distributed after each round of judgments.
- Inter-rater consistency graphs should be distributed after each round of judgments.
- Impact data should be first presented after round 2 judgments in preparation for round 3 judgments. The format for that presentation is numerical data and graphs. A pie chart shows the percentage data for performance within achievement levels and a cumulative bar chart shows the percentage at or above each achievement level.
- Following round 3 judgments, an interactive tool should be added to the review of impact data. Panelists should be able to determine the cut score associated with impact data that they judge to be both consistent with the ALDs and more reasonable in light of all the information they have received throughout the process. They can evaluate numerous cut score and impact data combinations and discuss them with other panelists. Their decision should be their own. Panelists should then be asked to respond to a questionnaire that is designed to capture their judgments regarding the cut scores and associated impact data to recommend to the Governing Board for reporting the NAEP results.

An inter-rater consistency exercise can be implemented as a second type of inter-rater feedback after the first round of judgments when using a holistic ALS procedure such as the Body of Work methodology. A list of student responses (Bodies of Work) for which judgments showed least agreement may be provided to panelists for discussion. Several ways of showing a lack of agreement can be used to select approximately 10 examples of student performance (Bodies of Work) for discussion:

- Bodies of Work for which judgments are closest to a 50-50 split between two achievement levels
- Bodies of Work for which judgments are spread across the most achievement levels
- Bodies of Work for which judgments are largely split between two non-adjacent levels

- Bodies of Work for which the classification is a reversal (a majority of booklets are classified at a level inconsistent with the modal classification of booklets around it)

Panelists may be given the list and data for the booklets and asked to discuss their own judgments for each. The discussion is to clarify their understanding of the performance required in order to align with the ALDs. Through this discussion, panelists enrich and strengthen their common understanding of performance at each level of achievement.

h) The process shall consist of at least two achievement level setting meetings with distinct groups of panelists, a pilot study, and an operational meeting. The purpose of the pilot study is to conduct a full “dress rehearsal” of the operational meeting, including but not limited to: an opportunity to try out materials, training procedures, collection of panelist judgments, feedback given to panelists through the process, software used to conduct analyses, meeting logistics, and other essential elements of the process. The pilot study may result in minor changes to the procedures, as well as major changes that would need additional study before being implemented in an operational meeting. The pilot study provides an opportunity for procedural validity evidence and to improve the operational meeting. At the discretion of the Governing Board, other smaller-scale studies may be conducted prior to the pilot study or in response to issues raised by the pilot study. The criteria in Principle 2a apply to panelists of both meetings.

Two types of panel meetings are required for each ALS procedure: a pilot study and an operational ALS panel meeting. The policy specifies that the pilot study will be implemented to carry out the exact procedures designed for the operational ALS. The design must be reviewed and approved prior to implementation. In addition, if research is needed prior to the pilot study to examine new methods and procedures with participation of panelists, this type of study is designated as a field trial. If necessary, a field trial may also be called for after the pilot study and before the operational ALS. Whether the need is for a field trial or a second pilot study will be determined through the technical advice and expertise of COSDAM, TACSS, and the COR.

Certain features of the panel studies should be standardized across all studies:

Panelists: The same procedures must be used for recruitment and selection of panelists for the pilot study and operational ALS. The criteria for the pilot study panel may be less stringent, if necessary, in order to meet fully the requirements for the operational ALS panel. Whether panelists for a field trial need to meet the requirements for pilot and ALS panels depends upon the purpose of the study.

Facilitators: The same facilitators must be used for the pilot study and operational ALS. If one or more process facilitators are needed for the field trial, they should be the same as facilitator(s) that serve for the pilot and operational ALS. The same holds for content facilitation.

Materials: All materials planned for the operational ALS must be provided for the pilot study using the content and format planned for use in the operational ALS. This includes

the advance materials, feedback, and evaluations. If the need for modifications to materials is revealed in the pilot study, however, those changes must be made for the operational ALS. Technical advice will determine the need for an additional research study prior to implementation with the modified materials. Any materials required for the field trial that are planned for use in the operational ALS should, as nearly as practicable, use the content and format planned for use in the operational ALS.

Meeting Logistics: The pilot study and operational ALS must be conducted in the same facility and using the same meeting room layout.

Security: Many individuals will have access to various parts of the secure NAEP item pool during the achievement level setting process and will have information on NAEP results prior to the official release of the data. It is imperative for the COR to ensure that the contractor has processes in place to ensure that secure materials and data are securely controlled as well as confidentiality maintained at all times. Security is a serious concern; it is a felony to disclose confidential NAEP data or materials.

Each achievement level setting project must include effective data security plans that demonstrate how security procedures will be employed and monitored at all times for the duration of the contract. This includes security procedures for (1) item distribution, (2) item review, (3) data review, (4) storage of computers/tablets containing secure materials, (5) server security and avoidance of distributed denial of service (DDoS), if applicable, and (6) hotel and other staff security maintenance. Data security plans should be incorporated into the Design Document.

NCES requires that any person(s) who will be reviewing or using secure data or materials sign nondisclosure agreements. Throughout an achievement level setting project, there may be a need to share secure data and materials across several individuals and groups with have signed nondisclosure agreements; this should occur via a secure project workspace rather than by email.

Procedures and Methodology

All procedures for the ALS process must be implemented in the same way for both the pilot study and operational ALS. This includes, but is not limited to, the agenda, software, instructions and training, standard setting methodology, feedback, and evaluations. If the need for changes to the procedures or methodology is revealed at any point prior to the operational ALS, these changes must be evaluated by the TACSS. The recommendations of TACSS should be considered by COSDAM when determining the need for additional panel studies prior to the operational ALS.

- i) The Governing Board shall ensure that a Technical Advisory Committee on Standard Setting (TACSS) is convened to provide technical advice on all achievement level setting activities. Technical advice provided by standard setting experts throughout the project is intended to ensure that all procedures, materials, and reports are carried out in accordance with current best practices, providing additional validity evidence for the process and results. The Board or its contractor may also seek technical advice from other groups as appropriate, including NCES and the larger measurement

community (e.g., the National Council on Measurement in Education).

Purpose and Role

A TACSS will be convened to provide technical advice on all achievement level setting activities. The members of TACSS should be appointed by the achievement level setting contractor according to the policy and with the approval of the COR. Contractually, advice and recommendations of the TACSS are not under direct supervision of the Governing Board, and all advice and recommendations to the Governing Board are the responsibility of the contractor.

Qualifications and Composition

The number of TACSS members may vary, depending upon the particular requirements of the ALS process and contract, but a minimum of six members is required. In consultation with NCES, TACSS membership should include a representative of the Design, Analysis, and Reporting (DAR) contractor to NCES who is involved with all operational procedures for NAEP data scaling and analysis. The TACSS must include individuals with expertise in NAEP scaling and analysis procedures and in achievement level setting procedures. It is not necessary for TACSS members to have content expertise in the subject for which achievement levels are being set. At least one TACSS member must have been involved in a previous NAEP achievement level setting process for the Board.

Key members of the contractor's staff, the COR, and the NCES liaison to the Governing Board for ALS procedures regularly attend TACSS meetings. Additional staff from each may be invited to attend at the discretion of the COR.

Meetings

The number of TACSS meetings may vary, depending upon the particular requirements of the ALS process and contract. In-person TACSS meetings, as well as webinars, should be scheduled throughout the entire contract to coordinate with key points in the planning, implementation, and reporting periods of the ALS process. At a minimum, the TACSS should review the following:

- Design Document
- Plans for panelist recruitment
- Composition of panels relative to the design for recruitment
- Instructional materials, such as the orientation video in advance materials
- Materials to be used in the panel meetings
- Software to be used for collection of panelist data
- Software to be used for analysis of panelist data
- Feedback and results from the panel meetings
- Evaluations of the process and analyses of data
- Reports to be presented to the Governing Board
- Validity evidence

In addition to participation in scheduled meetings of TACSS, two members will be invited to observe ALS panel meetings and special studies. In addition, one or two TACSS

members may be asked to attend a specific meeting of the Governing Board or meetings of other organizations when topics of importance to the NAEP ALS process are presented or discussed.

Other Sources of Technical Advice

Throughout the ALS process, teleconference meetings (on a regular basis or as needed) should be scheduled with the COR, the ALS contractor, and NCES contractors for the NAEP program. These meetings will be for the exchange of information regarding the ALS process and schedule of activities requiring data and other inputs from NCES contractors. The NAEP contractors are involved in technical aspects of the assessment administration and analysis of results that impact the ALS process.

Technical advice may also be requested through outreach to organizations such as the National Council on Measurement Education. Finally, content and other experts may be invited to inform TACSS about specific issues, as needed.

- j) All aspects of the procedures shall have documentation as evidence of the appropriateness of the procedures and results. This evidence shall be made available to the Board by the time of deliberations about the achievement levels. A summary of the evidence shall be available to the public when the achievement level results are reported.

Evidence to Evaluate the Procedures

Adherence to the approved design for the process and procedures from start to finish is necessary to show that the procedures implemented were not changed arbitrarily or for convenience during the implementation of the ALS process. This includes adherence to the design for the recruitment and selection of panelists and success in meeting targets for the composition of the panels; the qualifications of facilitation staff; the materials and information provided to train panelists relative to those called for in the design of the process; materials and information provided as feedback to inform panelists' judgments relative to the design; and, the evaluation of the process and outcomes by panelists.

The evidence documented for these procedures must be consistent with the approved design of the process and consistent with best practices. The procedures must be implemented satisfactorily and evaluated positively. Panelists' understanding of the process, sense of confidence when applying procedures and making judgments, and statistical agreement should increase with each round of judgments. This signals that the procedures have functioned appropriately and successfully.

Evidence to Evaluate the Results

Evidence to evaluate the appropriateness of results may come both from the procedures implemented in the ALS process and from sources external to the process. Panelists' evaluations of the results should lead to their recommendation that the results be adopted by the Governing Board. If panelists judge the procedures to be appropriate and their implementation to be appropriate, they are likely to judge the results as appropriate. But, if the results appear to be inconsistent with panelists' judgments regarding the relationship between ALDs and cut scores, they will likely judge the results to be

inappropriate.

Panelists' judgments should demonstrate common agreement on the level of achievement required to reach each level of performance. The variance of their individual judgments for cut scores should decrease across each round of judgments. Differences in cut scores derived from judgments of panelists grouped by panelist type, table group, demographic characteristics, and so forth should also be statistically equivalent, and they should exhibit patterns predicted in advance.

Statistical comparisons of results from the operational study should be made to results from similar standard settings from similar assessments. Identification of assessments for valid comparisons with NAEP is challenging. There are differences between NAEP and other assessments that must be acknowledged and accounted for when comparing results. If state samples are a part of the NAEP assessment, it may be possible to identify results of state standard setting to compare.

Summary Evidence to Present to the Public

It is important to be transparent about the ALS procedures and results. The policy requires that the evidence of appropriateness of procedures and results be summarized and made available for public review at the time the ALS results are reported.

All achievement level setting projects should result in a final report that contains information about the final recommendations and describes the full process and evidence for arriving at those recommendations. It is essential that the report contains clearly stated and well-organized documentation of the logistical, methodological, and technical aspects of the achievement levels process. The report should also be of a quality and style that will yield information accessible to the broad audience of NAEP achievement levels, including the education community, policymakers, and the interested public.

The report should consist of three sections: an executive summary, the full text and discussion, and appendices containing all relevant tabularized materials. The executive summary and full text and discussion components should be written in a way that allows each to be presented as a standalone document suitable for public distribution separate from the appendices.

The report should include the following sections: description of achievement level setting process; technical advice and decisions reached during the project; data analysis procedures; materials, procedures, and analysis; recommended achievement level results; achievement level descriptions and exemplars; validation study activities; procedures and results related to obtaining public comment; and recommendations for future achievement level setting activities.

Two versions of the final report should be prepared: one that contains secure data and materials (for internal use only), and one version in which the secure data and confidential materials have been redacted. The redacted version of the report should be posted on the Governing Board website on the day of the Report Card release in which the given achievement level results are incorporated.

Customized summaries may also be produced and distributed to address interests and concerns of specific audiences, but a common set of information must be included in each summary. The information must be clear, concise, and engaging. Examples of the evidence may be helpful for demonstrating the appropriateness of procedures and results.

- k) Sample items and student responses known as exemplars shall be chosen from the pool of released items for the current NAEP assessment to reflect performance in the *NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced* regions of the scale. The use of exemplars is intended to help the public better understand what performance in each achievement level represents for each subject and grade. When possible, exemplars may also be chosen that reflect performance at threshold scores. The collection of exemplars shall reflect the content found in the achievement level descriptions and the range of item formats on the assessment.

Purposes and Uses of Exemplar Performances

Exemplar items are intended to increase public understanding of the knowledge and skills required in the ALDs. Items should be selected by panelists to represent the knowledge and skills that demonstrate the performance required to match the achievement level description for each level. The procedure for selection of exemplar items should be implemented after the final round of feedback has been presented and discussed. At this point, panelists should be intimately familiar with the ALDs and have a clear understanding of how item performance relates to the ALDs. Exemplar items and the accompanying performance data are used in reporting NAEP ALS results to make the ALDs more concrete. The exemplars are specific examples of knowledge that a student performing within an achievement level knows or a task that a student performing within a level of an achievement level can accomplish.

Exemplar items must be selected from among the items that will be made public and no longer be used in future NAEP assessments. The number of items so designated varies according to the format of the assessment. NCES, in collaboration with the Governing Board, determines the items to be released to the public for reporting results for each NAEP assessment.

Criteria for Selection of Exemplar Performances

Criteria for exemplar selection must include both empirical, statistical evidence and judgments regarding requirements for performance according to the ALDs. Panelists should make their judgments regarding the relationship between the level of performance required by the items and the ALDs for the items that have been selected for their consideration according to statistical criteria. Empirical evidence should confirm that students scoring within the cut score range of an achievement level are likely to answer the item correctly. A minimum requirement is that students have at least an average .50 probability of answering the assessment item correctly or of scoring at a specific rubric score level for a constructed response item. If an item mapping standard setting methodology is used, the probability of correct response used for identification of exemplar items should correspond to the response probability for item mapping. The same criteria should apply for exemplar items selected to represent performance at the

cut score of the achievement level, but the criterion should be applied at the cut score, rather than across the range.

Items should be classified at the lowest achievement level for which the statistical criteria are met. Panelists should be instructed to determine whether the performance required by students is well aligned to the ALD. Panelists also should be instructed to consider whether the probability of correct response for the item seems appropriate: not too high, indicating the item was very easy for students in the level, and not too low, indicating that the item was very difficult for students in the level.

There is no limit to the number of items panelists can select for recommendation to the Governing Board to use in reporting. In general, panelists should be encouraged to select all items that they judge to be appropriate for representing performance at the achievement levels. Items recommended to the Governing Board for use should be approved by a majority of panelists and the level of disapproval should be minimized.

- 1) The outcomes from the achievement level setting panel meetings (recommended cut scores, exemplars, and ALDs for use in reporting) shall be forwarded to the Board for their consideration.

The outcomes from the achievement level setting panel meetings (recommended cut scores, exemplars, and reporting ALDs) shall be presented to the Board for their consideration. The Governing Board by-laws assign the responsibility of monitoring and overseeing the achievement level setting process to COSDAM. In order to provide technical guidance and to be prepared to reach agreement on a recommendation for the full Board, COSDAM must be updated regularly and kept fully informed regarding the ALS procedures and progress.

Presentations to COSDAM

The major outcomes of the ALS process should be reported to COSDAM for both the pilot study and the ALS. COSDAM should be briefed through written reports and/or in person at quarterly meetings regarding each key aspect of the ALS process, and interim briefings will be scheduled as necessary if there are time-sensitive aspects of the process that cannot wait for the next quarterly meeting. It is important that COSDAM be informed throughout the process in order to anticipate the outcomes and the decisions to be made regarding the cut scores.

Panelists' final recommendations should be made available to COSDAM in a timely manner to allow ample time for their discussion and consideration of the data. COSDAM should have information for discussion during at least one meeting prior to the quarterly meeting at which achievement levels are to be formally adopted by the Governing Board.

Presentations to the Governing Board

The final decision for setting achievement levels is made by the full membership of the Governing Board, based on the recommendations of COSDAM. The presentation of ALS information to the Board is generally made by one or more members of COSDAM, the COR, and the ALS contractor's project director. Members of the TACSS and others directly

involved in the ALS process may be invited to participate in the presentation of information and findings to the Governing Board. The determination of presenters will generally be decided by the COSDAM Chair in coordination with the COR.

The final decision by the Governing Board for setting achievement levels requires ample time for reaching an understanding of the results and the real and potential impacts of the results. The Governing Board must be given preliminary data regarding the cut scores and performance relative to the cut scores, as well as the ALDs and exemplar items in a timely manner to enable Board members to reach the necessary understanding of the recommendations being made and to reach agreement on the final levels to be set. This generally requires a briefing prior to the quarterly Board meeting at which the final decision is to be made.

Principle 3: Validation and Reporting of Achievement Level Results

The achievement level setting process shall produce results that have validity evidence for the intended uses and interpretations and are informative to policy makers, educators, and the public.

- a) Professional testing standards require evidence to support the intended interpretations and uses of test scores. Among the sources of evidence supporting the validity of test scores is evidence bearing on the standard setting process and results. Standard setting is necessarily judgmental, and the Board shall examine and consider available evidence about the procedural integrity of the achievement level setting process, the reasonableness of results, and other evidence in order to support intended uses and interpretations.
- b) The Board shall examine and consider all evidence related to validity of the achievement level setting activities. These data shall include, but not be limited to: procedural evidence such as training, materials and panelist evaluation data; reliability evidence such as consistency across panelist type, subpanels, rounds, and meetings, if appropriate; and external comparisons to other similar assessments, if appropriate, with necessary caveats. The results from validation efforts shall be made available to the Board in a timely manner so that the Board has access to as much validation data as possible as it considers the recommendations regarding the final levels.

Throughout the process, the COR and TACSS should monitor the process and interim results. Information should be shared with COSDAM in regularly-scheduled quarterly meeting, and as needed, to help assure that the process is functioning as designed and to avoid future issues.

Alignment and Fidelity of Implementation with Process Design

A comparison of the procedures implemented with the procedures detailed in the Design Document should serve as a basis for evaluating the procedural validity of the achievement level setting process. Procedural validity is a necessary, but not sufficient, condition for establishing the validity of a standard setting process. As noted in Principle

2a, the Design Document should provide a detailed description of each step in the process, and be vetted for confirmation that the important details are included and sufficiently described to serve as a guide for implementation of the ALS procedures. The procedures must be implemented according to the design, and panelists must evaluate the implementation positively. Minor variances may be acceptable, and all variances must be documented and explained.

Reasonableness of Results

Panelists' evaluations (Principle 2f) of each key step in the process should serve as a basis for judging the reasonableness of results. Reasonableness, however, is a judgment. By the last round of performance judgments, panelists are expected to have a thorough understanding of the ALDs and to be well prepared to evaluate the reasonableness of the performance data relative to the ALDs. There must be clear evidence that this is the case. Positive evaluations by panelists of the process and the results of the process should provide evidence of the reasonableness of the results. Asking panelists directly if they would be willing to sign a statement supporting the reporting and use of results has typically served to provide confirmation of their judgment of the results as reasonable.

In addition to the evaluation by panelists of the reasonableness, results may be presented for evaluation by content experts and measurement experts with knowledge of NAEP and standard setting. The TACSS members, in particular, are well versed in the process and should judge the integrity of the process and reasonableness of results. Their judgment and judgments of other experts regarding the reasonableness of results relative to the ALDs and their knowledge of student achievement can confirm evidence of the reasonableness of the results. Ultimately, COSDAM and the Governing Board must judge the reasonableness of the results. This judgment should take into account the evidence presented from other reviewers and data presented.

Criteria for Evaluating Evidence of the Validity of the Achievement Level Setting Process

Training: Criteria for validity are positive evaluations by panelists of the amount of time allocated for training, positive comparisons of the amount of time for training relative to previous achievement level settings, and positive evaluations of training for each key task.

Panelist evaluation data: Positive evaluations of the process and the outcomes of the process serve as evidence of validity. Comparisons of the evaluation responses throughout the process, increasingly positive evaluations across iterations of procedures, and comparisons of evaluation responses to previous ALS processes must be evaluated. Positive results of these evaluations serve as the criteria for establishing the validity of the process and outcomes.

Materials used in the process: Materials used to instruct panelists must be accurate, clearly stated, and easily understood. The volume of material must be sufficient to thoroughly train panelists without creating undue burden. Materials should inform panelists for the tasks using multiple approaches and media. The timing and combination of materials provided must be evaluated as appropriate to the purpose

and effective for achieving the purpose. Panelists must evaluate the materials as meeting these criteria.

Statistical analyses: In addition to the qualitative analyses of procedures and results, quantitative analyses of procedures are required to provide support for the validity of the process and results. Statistical analyses of evaluations of the process and results of the process must provide confirmation that panelists are well trained and able to carry out procedures successfully to achieve the purpose and that the results of the process are statistically sound.

Results of statistical analyses of evaluation data, cut score data, and other data outputs from the ALS process should yield evidence of reliability. The results should show no statistically significant differences based on key attributes of panelists and organizational features of the process. These data should be analyzed across rounds, as appropriate, in addition to the following breakdowns:

- Demographic characteristics of panelists: e.g. sex, geographic region
- Panelist type: teacher, non-teacher educator, general public
- Table groups
- Split panel groups
- Pilot study results and operational ALS results

Additional data may be available for analysis of procedural evidence and results of the ALS procedure with other NAEP ALS procedures and with standard setting in the content area for other assessments judged to be appropriate for comparison.

- c) NAEP achievement levels are intended to estimate the percentage of students (overall and for selected student groups) in each achievement level category, for the nation, and for states and trial urban districts (TUDAs) for some assessments. NAEP is prohibited by law from reporting any results for individual students or schools.
- d) In describing student performance using the achievement levels, terms such as “students performing at the *NAEP Basic* level” or “students performing at the *NAEP Proficient* level” are preferred over “*Basic* students” or “*Proficient* students”. The former implies that students have mastery of particular content represented by the achievement levels, while the latter implies an inherent characteristic of individual students.
- e) In reporting the results of NAEP, the three achievement levels of *NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced* refer to the three regions of the NAEP scale at and above each respective cut score. The remaining region that falls below the *NAEP Basic* cut score shall be identified as “below *NAEP Basic*” when a descriptor is necessary.
- f) In describing the *NAEP Proficient* level, reports shall emphasize that the policy definition is not intended to reflect “grade level” performance expectations, which are typically defined normatively and can vary widely by state and over time. *NAEP*

Proficient may convey a different meaning from other uses of the term “proficient” in common terminology or in reference to other assessments.

- g) To facilitate valid uses of ALDs for the purpose of reporting, the Board shall ensure that the descriptions of performance for the achievement levels reflect what the empirical data reveal about the knowledge and skills demonstrated by students in that score range. To develop ALDs for reporting, following the achievement level setting the Board shall revisit and may revise content ALDs to ensure that they are consistent with empirical evidence of student performance. In particular, these “Reporting ALDs” chosen to illustrate the knowledge and skills demonstrated at different achievement levels shall be written to incorporate empirical data from student performance. Reporting ALDs shall describe what students at each level *do* know and *can* do rather than what they *should* know and *should* be able to do.

Following Board action to adopt new achievement levels, anchoring studies may be used to evaluate the ALDs from the ALS process to determine what modifications are needed for reporting results. Anchoring studies (also known as item mapping studies) use empirical data from student performance to evaluate items that “anchor” or “map” within each achievement level range of the score scale. The goal of these studies is to assure that the reporting ALDs for each achievement level describe performances that reflect empirical evidence of the knowledge, skills, and abilities demonstrated within each achievement level. The Reporting ALDs should describe what students performing at each level of achievement actually do know and can do. Modifications to the ALDs used in the achievement level setting process for reporting purposes may include both the addition and deletion of statements. At a minimum, if no additions or deletions are needed, statements about what students *should* do will be revised to what students *can* do.

The statement of work for an achievement level setting project should also include the development of reporting ALDs as a final task in the contract. The Design Document described in Principle 2a should include procedures for developing the reporting ALDs, and the TACSS should also oversee this work. The full item pool (also used in the recently conducted achievement level setting activities) should be used for developing reporting ALDs.

Panelists for Anchoring Studies

A strong and high level of content expertise is required for this task, and previous experience with the relevant NAEP framework is highly desirable for the majority of participants, if feasible. This includes persons who served on the NAEP Framework Development Panel, NAEP Standing Committee for item development, development of ALDs for achievement level setting (if performed in a separate step from the framework development process), or achievement level setting panels. This process is typically limited to teachers and non-teacher educators with relevant content expertise.

Replicate panels should be used. For each grade and replicate panel, there should be at least 3-5 panelists, and each grade should have the same number of panelists. Panelists should reflect diversity in terms of gender, race/ethnicity, region of the country, and urbanicity.

Materials and Procedures for Anchoring Studies

Many of the procedural details described under Principle 2 (e.g., advance materials, format of meetings, training, evaluations) are also applicable to anchoring studies to develop reporting ALDs, but modifications will be needed given the difference in purpose of the meetings.

Noteworthy procedural details that are unique to developing reporting ALDs are as follows:

Statistical Criteria: Anchoring studies evaluate items that “map” or “anchor” within the score range of achievement levels using statistical criteria—typically a response probability (RP) and often a discrimination criterion. The statistical criteria should be chosen to consider both consistency with the ALS process and how results and exemplar items will be reported in the Nation’s Report Card. In the most recent anchoring studies performed with the 2009 reading assessment and the 2009 math assessment at grade 12, a response probability of .67 was used. A discrimination criterion is often used to assure that there is a *reasonable* difference between the probability of correct response at two adjacent levels. A discrimination criterion at the 40th percentile of differences in RP at adjacent levels has typically been used in previous studies.

Decision Rule: The ALDs used in achievement level setting may be modified to develop the Reporting ALDs either because too few items map within the achievement level range to justify specific mention of the knowledge, skill or ability in the ALD or because several items map within the achievement level range for which there is no descriptor of the knowledge, skill, or ability in the ALD. An appropriate decision rule must be adopted for making the modifications—either to add descriptors of performance or delete descriptors. The item pools for NAEP vary somewhat from one assessment cycle to the next. Both items and ALDs are written to represent the framework, but specific assessments may not have the same number of items measuring a particular aspect of the framework or ALDs. The decision to modify the ALDs for reporting purposes should be based on more than one or two discrepant items. Approximately 10% of the items in the item pool for the grade level is recommended for consideration as the criterion, but the judgment of the panelists regarding the importance or significance of the performance may override the statistical criterion. A convention for NAEP achievement levels is that a descriptor does not need to be repeated for a higher achievement level if the performance requirement does not change at that next higher level. But, if the judgment is that some mention should be made for clarification of the performance requirement for a knowledge, skill, or ability, then that judgment may override the statistical/quantitative decision rule.

Item Difficulty: Previous anchoring studies have typically excluded items that did not anchor because the items were too difficult. The criteria for determining that has been RP.50: items that were so difficult that the probability of correct response, even at the NAEP Advanced level, did not reach .5. Anchoring study panelists should be made

aware of the RP associated with all items, and they should evaluate the items with RP.50 or lower in order to understand the performance requirements of those items as part of the evaluation of the empirical data relative to the ALDs.

Alignment to Exemplar Items: Finally, the draft Reporting ALDs should be evaluated relative to the exemplar items to represent each achievement level. It is important that the exemplar items serve to illustrate the performance described in the reporting ALDs.

Results from Anchoring Studies

The reporting ALDs drafted by the replicate panels must be adjudicated to produce a single reporting ALD for each subject and grade. Differences must be discussed and the rationale for each understood by both groups. Reporting ALDs must clearly represent the policy definitions and the same calibration of achievement as the ALDs used in the ALS process.

The reporting ALDs must be submitted for review and evaluation relative to the ALDs used for the ALS process. Key stakeholder groups, including content, policy, educator, and parent groups should be contacted and encouraged to participate in the review, particularly if there are substantial changes to the ALDs. NAEP content experts should also be sought out as judges for this comparison. Reviewers should be provided with the assessment framework, the policy definitions, and the reporting ALDs. The evaluation should focus on the alignment of the reporting ALDs with respect to the policy definitions, the alignment of the reporting ALDs across the three achievement levels within each grade, and the alignment of the reporting ALDs for each level across the three grades.

Recommendations should be evaluated by the panelists who created the reporting ALDs to determine whether any additional modifications are needed. The results of these reviews, as well as a complete description of the process, should be reviewed by TACSS.

The composite of information and recommendations will be presented to the Governing Board (via COSDAM) for consideration and approval.

- h) An interpretative guide shall accompany NAEP reports, including specific examples of appropriate and inappropriate interpretations and uses of the results.

The recent evaluation of NAEP achievement levels (National Academies of Sciences, Engineering, and Medicine, 2017) includes a recommendation to provide guidance “to help users determine inferences that are best made with achievement levels and those best made with scale score statistics” (p. 13). Users need a solid understanding of the achievement levels to interpret and understand what the percent of students at or above *NAEP Proficient* means. To assist users, an interpretative guide should include illustrative uses of NAEP data, but it cannot be considered an exhaustive list. Guidance should include information about why certain uses and interpretations are inappropriate rather than merely a listing of what is appropriate or inappropriate.

The interpretative guide should be easily accessible and should have a link directly from

the Nation's Report Card. Some information can be the same for each assessment, while other information will likely need to be customized. Many misuses of NAEP data occur when people make inappropriate causal conclusions, interpret NAEP Proficient as representing grade level performance, or construe gap trends using achievement level results. Common misuses should be included with a rationale for why certain uses are inappropriate.

The Governing Board plans to begin including interpretative guides with the release of the 2021 results, as described in the Achievement Levels Work Plan.

Principle 4: Periodic Review of Achievement Levels

Periodic reviews of existing achievement levels shall determine whether new achievement level descriptions and/or cut scores are needed to continue valid and reliable measurement of current student performance and trends over time.

- a) At least once every 10 years or 3 administrations of an assessment, whichever comes later, the Governing Board, through its Committee on Standards, Design and Methodology (COSDAM), shall review the alignment between the content ALDs and items, based on empirical data from recent administrations of NAEP assessments. In its review, COSDAM (in consultation with the Assessment Development Committee) shall solicit input from technical and subject matter experts to determine whether changes to the content ALDs are warranted or whether a new standard setting shall be conducted, making clear the potential risk of changing cut scores to trends and assessment of educational progress. Relevant factors may include but not be limited to: substantive changes in the item types or in the balance of item types; changes in the mode of administering assessments; advances in standard setting methodologies; and changes in the policy environment for using NAEP results.

The purpose of these reviews is to determine whether changes to the content ALDs are warranted or whether a new ALS process must be implemented. An anchoring study methodology can be used for this review. The methodology described in Principle 3g for developing Reporting ALDs can be used for these analyses, but some modifications will be needed.

Evaluation of Existing ALDs

The initial goal should be to evaluate the existing ALDs rather than to modify them. Results of the evaluation may indicate the need for modification, however. Anchor descriptions must be developed from the items that are mapped into each cut score range. The entire item pool for each grade level in the assessment should be used for these studies, and the study should include items from two recent administrations of the assessment. If modifications have been made to the assessment that are of concern—change in mode of administration, the balance of item types, and so forth, then it may be most appropriate to use only the most recent assessment that incorporates the changes.

Over the period of 10 years or 3 administrations (whichever comes later), several types of changes may have taken place that could impact the alignment of items to the ALDs. Examples of these factors are included in Principle 4a). If the Anchor descriptions do not align with the policy definitions, that would signal the potential need for new cut scores because the knowledge, skills, and abilities required for performance on items within the cut score ranges would not match the performance described in the policy definitions. The lack of match or alignment can be indicated by finding that the anchor descriptions for an achievement level include performances that are higher/more difficult or lower/less difficult than that required in the policy definition for the level. Changes in the balance of item types or mode of assessment administration could lead to this finding.

If the performance within the range of each achievement level does align with the performance requirements of the policy definitions, the anchor descriptions should be compared against the existing ALDs to evaluate differences in the types of performances represented in each and to compare the calibration of the performance for the achievement described in each. Note that although the calibration of a specific aspect of performance in one or more achievement levels may be judged to be unaligned with the policy definition, this would not generally signal the need for a new set of cut scores. Rather, this would signal the need for revised reporting ALDs to address the misalignment of this particular aspect in the ALDs describing what students know and can do. The reviewers should look for evidence about whether the types of knowledge, skills, and abilities are the same, as well as whether the level of performance is consistently the same.

The anchor descriptions should also be compared against the existing ALDs to determine whether the two are approximately equivalent in terms of the level of performance required and the specific performance type of knowledge, skills, or abilities required. The reporting ALDs are more specific than the policy definitions, and differences may become apparent in this comparison that were not evident in the comparison to policy definitions.

If the calibration of the two sets of descriptions appears to be approximately the same, this would support maintaining the current cut scores. If, however, the performances represented in the anchor descriptions are included in the ALDs for a different achievement level, this would signal a misalignment of cut scores and ALDs. A change in administration mode or in the item format used for assessing the same performance could lead to this finding. Further research would be needed to determine appropriate adjustments, and the magnitude of differences would be a key indicator of whether adjustments to the ALDs would be sufficient or whether new cut scores will be required.

If the calibration of performance requirements of knowledge, skills, and abilities of the anchor descriptions and ALDs are judged to be approximately equivalent, it is possible that the content of the performance in the two sets of descriptions is found to be at variance. In this case, some knowledge, skills, or abilities may be included/excluded in the anchor description that are/are not in the ALDs. Changes to the item types or proportion of item types included in the assessment may lead to this sort of result. This finding of

consistency in calibration, along with inconsistency in specific performance described, would call for modifications to the ALDs without modifications to the cut scores.

Consequences of Judgments for Change

Any judgments for change require more research and evaluation:

- A judgment for the need to make changes to reporting ALDs requires that the source of the need is fully understood.
- The decision to set new cut scores means that trend data on achievement levels is lost. That is an important decision.
- If the results of the anchoring study indicate that performance relative to the cut scores is not well aligned to the policy definitions and/or ALDs, a decision must be made for whether a change in cut scores or ALDs should be made. In some cases, the ability to maintain the trends for reporting relative to the cut scores is most important, and an anchoring study design can be used to develop new ALDs.
- If the judgment is that new cut scores are needed to correspond more closely with the current ALDs, then a new ALS process will have to be implemented.

The Design Document for this work will specify the decision steps and the role of the TACSS and content experts providing evidence for COSDAM's deliberation.

- b) Within the period for a review of achievement level descriptions and cut scores, changes may occur to a NAEP framework. If a framework is replaced or revised for a major update, a new achievement level setting process may be implemented, except in circumstances where scale score trends are maintained. In this latter instance, COSDAM shall determine how to revise the ALDs and review the cut scores to ensure that they remain reasonable and meaningful.

If a framework is replaced or revised for a major update and scale score trends are not maintained, then a new achievement level setting will be necessary. If a framework is replaced or revised for a major update and scale score trends are maintained, then a new achievement level setting may or may not be necessary. The procedures described under Principle 4a can be used to help determine whether a new achievement level setting should be performed, or whether the ALDs should be revised instead. It is possible that a framework update could result in a decision to maintain scale score trends but establish new cut scores; the decision of whether to maintain scale score trends and achievement levels are related but distinct.

- c) If there are major updates to a NAEP framework, the ALDs shall be updated by the Framework Visioning and Development Panel. (See the Governing Board Policy on Framework Development for additional details). Following an assessment administration under the revised framework, COSDAM shall use empirical data to revise content ALDs to align with the revised framework.

If the procedures described in Principle 4b result in a decision to maintain the cut scores and revise the ALDs created by the Framework Development Panel for reporting purposes, then reporting ALDs should be developed as described in the procedures for Principle 3g.

- d) As additional validation evidence becomes available, the Board shall review it and make a determination about whether the achievement levels should be reviewed and potentially revised.

It is very challenging to obtain readily available data from external assessments with nationally representative samples that represent similar constructs, ALDs, and achievement levels. Additional research undertaken by the Governing Board or other groups after achievement levels have been set may provide relevant validity evidence.

To address some validity questions, it may be necessary to conduct original research after the conclusion of an achievement level setting project, such as a study of original data collection of teacher judgments. An ad hoc technical advisory group consisting of content and technical experts should be convened to review such evidence and make recommendations to the Board about whether the achievement levels should be reviewed and potentially revised.

Principle 5: Stakeholder Input

The process of developing student achievement levels is a widely inclusive activity. The Governing Board shall provide opportunities to engage multiple stakeholders throughout the achievement level setting process and shall strive to maximize transparency of the process.

- a) The process of seeking nominations for the achievement level setting panels shall include outreach to relevant constituencies, such as: state and local educators; curriculum specialists; business representatives; and professional associations in a given content area.

The process of seeking nominations for achievement level setting panelists is outlined in the section under Principle 2 on *Procedures for the Recruitment and Selection of Achievement Level-Setting Panelists*. A nomination process helps to assure that outstanding individuals are identified to serve as panelists and that the representativeness of the achievement level setting panels is diverse. Further, the nomination process increases vastly the number of persons included in the standard setting process and helps to focus nationwide attention on the activity.

- b) The Design Document (describing in detail all planned procedures for the project) shall be distributed for review by a broad constituency and shall be disseminated in sufficient time to allow for a thoughtful response from those who wish to provide one. All interested stakeholders shall have an opportunity to provide public comment.

The procedures for development and review of the Design Document are presented in Principle 2. Stakeholders for the review of the Design Document should include content individuals, groups, and organizations as well as members of the technical and policy

communities. All users and potential users of NAEP achievement levels results should be encouraged to participate in the review of the design for the ALS process. Activities for obtaining public comment may include, but will not be limited to, meetings, canvassing of various groups and individuals, hearings, and written and oral communication to engage a broadly representative group who has a vested interest in the process and results of the NAEP achievement levels.

- c) Achievement level setting panelists shall include teachers, non-teacher educators, and other interested members of the general public with relevant educational background and experience, including parents, researchers, and employers. Each panel shall reflect diversity in terms of gender, race/ethnicity, region of the country, urbanicity, and experience with students with disabilities and English language learners.

The procedures for selecting ALS panelists are described in the section under Principle 2 on *Procedures for the Recruitment and Selection of Achievement Level-Setting Panelists*.

- d) All achievement level setting activities shall be informed by technical advice throughout the process. The Technical Advisory Committee on Standard Setting shall provide ongoing technical input from standard setting and assessment experts, and other groups with relevant technical expertise may be consulted periodically as needed.

Standard setting is a judgmental process that is encased within a strong base of technical support and advice. The procedures associated with that advice are described in the section under Principle 2 on *Sources of Technical Advice*.

COSDAM

The Governing Board structure includes the Committee on Standards, Design and Methodology to be the Board's technical oversight and source of advice regarding achievement levels. COSDAM meets quarterly and is regularly briefed on the achievement level setting process from preparation of the procurement through approval and reporting of results. In addition to the regularly scheduled quarterly meetings, COSDAM has interim meetings scheduled whenever information from the Committee is needed to assure efficient progress in the process and information to brief members is needed for the Committee to take action. COSDAM must approve all key components of the ALS process and results, and COSDAM makes recommendations to the Governing Board regarding the setting, reporting and maintenance of achievement levels.

TACSS

The Technical Advisory Committee on Standard Setting is a required source of technical advice called for in procurements issued for setting achievement levels. The number of members and frequency of meetings is somewhat flexible. A minimum of six members is required, and TACSS meetings are to be scheduled in order to secure advice prior to each key step in the process. Although the TACSS is a requirement of the Governing Board, TACSS members officially report to the ALS contractor. This provides more independence to the TACSS in relation to the Governing Board and increases the probability that

recommendations by TACSS are objective and free of any conflict of interest or the appearance of a conflict of interest with the Governing Board.

Process and Technical Reports from Previous ALS Procedures

Final reports of ALS procedures provide complete documentation of the process and analyses of results. The reports are detailed and complete, including minutes of TACSS meetings. These documents provide guidance regarding the choice of procedures and their successful implementation. Some aspects of these reports are secure but redacted versions are prepared for public posting on the Governing Board website.

Other sources of Technical Advice within the NAEP Program

The Assistant Director for Psychometrics typically is in charge of the Governing Board's achievement level setting work. This staff person usually serves as the COR for all of the Governing Board's ALS procurements. Regular meetings with NCES staff help to assure that Governing Board staff are informed of developments in the NAEP program that may impact achievement levels and that NCES staff are informed of developments in the ALS process that may impact their work with the NAEP program.

In collaboration with NCES staff, the team of primary contractors (currently termed the NAEP Alliance) meet via teleconference with key members of the Governing Board's ALS contractor staff and ALS COR to assure that technical and logistical planning are coordinated to maximize efficiency and effectiveness of the ALS procedures.

The Design and Analysis Committee (DAC) is appointed by the Design, Analysis and Reporting (DAR) contractor to NCES to provide technical advice to the NAEP program. The DAC meets periodically to review and discuss technical issues regarding the NAEP program and to make recommendations to the DAR contractor regarding how to address the issues. The Assistant Director for Psychometrics typically is invited to attend these meetings and to provide updates regarding the Governing Board's work. Recommendations by the DAC may impact the ALS process, and understanding their deliberations and the rationale for their recommendations is important to the process.

From time to time, the Governing Board convenes panels of technical advisors to address special issues that may arise, and additional technical experts may be called upon to advise on special issues related to the ALS process.

Sources of Technical Advice External to the NAEP Program

Technical advice may also be requested through organizations such as the National Council on Measurement in Education (NCME). Presentations at the annual meetings of NCME help to communicate information about the NAEP ALS process and to collect feedback from participants who attend the presentation sessions. Members of the organization may be asked to provide technical information on specific topics and issues, and they may participate in panels convened for that purpose.

Governing Board staff regularly attend other meetings of NAEP stakeholders and may provide information regarding NAEP achievement levels to collect feedback. External stakeholders may be especially helpful in identifying potential sources of data for

research to evaluate the validity of NAEP achievement levels.

- e) Ongoing input and coordination with staff and contractors from the National Center for Education Statistics (NCES) is necessary to ensure that all achievement level setting activities are carried out in a manner that is consistent with the design, analysis, and reporting of NAEP assessments.

As noted in the Introduction, NCES designates a liaison to work with the Governing Board COR. The NCES liaison works closely with the COR to provide data, materials, and other operational information needed to carry out the achievement level setting process.

Principle 6: Role of the Governing Board

The Governing Board, through its Committee on Standards, Design and Methodology (COSDAM), shall monitor the development and review of student achievement levels to ensure that the final achievement level descriptions, cut scores, and exemplars recommended to the Governing Board for adoption comply with this policy.

- a) The Committee on Standards, Design and Methodology (COSDAM) shall be responsible for monitoring the development and review of achievement levels that result in recommendations to the Governing Board for any NAEP assessment under consideration. COSDAM shall provide direction to the achievement level setting contractor, via Governing Board staff. This guidance shall ensure compliance with the NAEP legislation, Governing Board policies, Department of Education and government-wide regulations, and requirements of the contract(s) used to implement the achievement level setting project.
- b) If there is a need to revise the initial achievement level descriptions (ALDs) created at the time of framework development for use in achievement level setting and/or reporting, the Governing Board shall take final action on revised ALDs based on recommendations from COSDAM.
- c) COSDAM shall receive regular reports on the progress of achievement level setting projects.
- d) COSDAM shall review and formally approve the Design Document that describes all planned procedures for an achievement level setting project.
- e) At the conclusion of the achievement level setting project, the Governing Board shall take final action on the recommended cut scores, exemplars, and ALDs for use in reporting. The Governing Board shall make the final determination on the NAEP achievement levels. In addition to the panel recommendations, the Board may consider other pertinent information to assess reasonableness of the results, such as comparisons to other relevant assessments.

- f) Following adoption by the Governing Board, the final ALDs, cut scores, and exemplars shall be provided to the National Center for Education Statistics (NCES) for reporting the results of the NAEP assessment(s) under consideration.
- g) Consistent with Principle 4 above, COSDAM shall periodically review existing achievement levels to determine whether it is necessary to revise achievement level descriptions or conduct a new standard setting.

At the Conclusion of the Achievement Level Setting Process

At the end of each panel meeting, the COR and project director should thank panelists for their important contributions and provide a reminder of important information and next steps. Panelists must be reminded that it is permissible to talk about the achievement level setting *process*, but they cannot reveal information about secure data and results. The COR should provide contact information if the panelists have questions about what information can and cannot be shared prior and after the official release of the Nation's Report Card.

The COR and achievement levels project director should both sign letters and certificates of appreciation to send to the panelists shortly after the meeting in which they participated concludes. In addition, the COR should notify panelists of the release date for the Nation's Report Card when it is determined. If the Board modifies the panel's recommendations, the COR should notify the panelists of that decision and the rationale for the decision.

After the Board takes action on achievement level recommendations, the COR should send official written notification of the achievement level cut scores, exemplars, and achievement level descriptions to the NCES Associate Commissioner for Assessment for inclusion in the Nation's Report Card.

NAEP Linking Studies: Overview and Frequently Asked Questions (SV #2)

One of the goals of the Strategic Vision is to “Increase opportunities to connect NAEP to administrative data and state, national, and international student assessments.” Over the past four years, COSDAM has had several discussions about linking studies that have been completed, are currently underway, or could be undertaken with future administrations of NAEP. NAEP linking studies have been conducted by NCES and their contractors, although some studies have been led and funded by the Governing Board (primarily in support of the Board’s previous research on academic preparedness for college).

General information about NAEP linking studies is provided below in the form of frequently asked questions and answers. The purpose of this short information session during the March 2019 COSDAM meeting is to solicit questions and feedback from COSDAM members on what additional information would be useful to inform future discussions.

What is a NAEP linking study?

NAEP linking studies generally involve connecting data from a particular NAEP assessment to data from another assessment, providing information about where a NAEP score would fall on the scale of another assessment, and/or where a score from another assessment would fall on the NAEP scale. Results from a NAEP assessment can be connected to data from another assessment under the following conditions: 1) some items from another assessment are included as part of the administration of NAEP; 2) a common group of students takes both NAEP and another assessment (typically at different points in time); or 3) randomly equivalent groups of students take NAEP and another assessment. Linkages to other assessments are either concurrent (i.e., relating NAEP to another outcome that takes place within the same time frame) or predictive (i.e., relating NAEP to a future outcome).

In addition to referring to assessment data, the term “linking study” has also been used to describe efforts to connect information from NAEP to data from other NCES surveys. For example, the Early Childhood Longitudinal Study – Kindergarten (ECLS-K) and the High School Longitudinal Study (HSLs) include a parent questionnaire but NAEP does not; parent-reported data about income and occupation have been used in ongoing efforts to develop and validate socio-economic status (SES) measures for NAEP¹.

Previous NAEP linking studies have connected NAEP to other NCES surveys and longitudinal studies (via intra-agency agreements within NCES); data from state longitudinal databases (via agreements with state agencies); and external assessments (via agreements with other testing programs).

¹ For several years, the NAEP program has been engaged in efforts to establish an improved measure of SES but no changes have been made to NAEP reporting of SES at this time; currently the program is still using eligibility for the National School Lunch Program and highest level of parental educational attainment.

What is the purpose of NAEP linking studies?

NAEP linking studies have been performed for a variety of purposes, such as:

- To estimate state-level performance on international assessments (e.g., linking NAEP and TIMSS was used to estimate TIMSS scores for all 50 states)
- To compare NAEP achievement levels with external benchmarks (and to understand the stringency of those performance standards) (e.g., a linking study of NAEP grade 4 reading and the Progress in International Reading Literacy Study [PIRLS]) found that the NAEP achievement levels are more stringent than the PIRLS benchmarks, explaining why fewer students reach the NAEP Proficient level in comparison to the PIRLS High benchmark)
- To compare state performance standards on a common scale (e.g., the state mapping studies use NAEP as a common metric for comparing the stringency of performance standards on state assessments)
- To estimate student performance on an external indicator of achievement, such as the percentage of grade 12 students academically prepared for college (e.g., several studies including a national NAEP-SAT linking study were used to determine the point of the NAEP grade 12 reading and mathematics scales that corresponded to likelihood of placement in non-remedial college coursework)
- To inform the collection of non-academic outcomes, such as the development of a new measure of socio-economic status (e.g., a link between NAEP and parent questionnaires from ECLS-K and HSLs were used to explore new measures of SES for NAEP)

What NAEP linking studies have been performed over the past 10 years?

Several studies have been performed or are currently underway. A brief summary of each follows, with a link to completed reports (where available) for additional information.

1. *NAEP-SAT Linking Study*: The purpose of this study was to identify a reference point or range on the NAEP 12th-grade reading and mathematics scales that might be associated with the College Board's SAT preparedness benchmarks. The NAEP and SAT scores for 12th-grade students who had taken both assessments in 2009 were the basis for this linking. The report based on the results of this study can be found at: https://www.nagb.gov/focus-areas/reports/preparedness-research/docs/statistical-relationships/landing/SAT-NAEP_Linking_Study.html.
2. *Longitudinal Analyses of Performance on NAEP Related to Performance in College and Other Outcomes of Florida Students*: The purpose of this study was to relate NAEP scores to ACT and SAT scores, college performance and other outcomes. Working with Florida state officials and their longitudinal database, scores for students who had participated in the 2009 NAEP 12th-grade assessments and were subsequently enrolled in Florida's public colleges in 2010 were linked to a variety of outcome indicators. The

report can be found at: https://www.nagb.gov/focus-areas/reports/preparedness-research/docs/statistical-relationships/landing/Florida_Statistical_Study.html.

3. *High School Transcript Study*: NCES periodically surveys the curricula of our nation's high schools and the course-taking patterns of high school students through its High School Transcript Study (HSTS). In conjunction with the administration of 12th-grade NAEP assessments, the HSTS also offers information on the relationship of student course-taking patterns to student achievement at grade 12. Transcripts were collected from seniors who graduated in 1987, 1990, 1994, 1998, 2000, 2005, 2009, and 2019. The most recent report that is available is for 2009 (the report for the 2019 data collection has not yet been released). Results from the 2009 study can be found at: http://www.nationsreportcard.gov/hsts_2009/.
4. *NAEP-TIMSS Linking Study*: NCES initiated this study to link the NAEP scale to the Trends in International Mathematics and Science Study (TIMSS) scale so that states could compare the performance of their students with that of students in other countries. The study was conducted in 2011 with eighth-grade students in all 52 states/jurisdictions that participated in the NAEP mathematics and science assessments. The report based on the results of this study can be found at: https://nces.ed.gov/nationsreportcard/studies/naep_timss/.
5. *NAEP-PIRLS Linking Study*: The purpose of this study was to obtain a statistical comparison between NAEP and the Progress in International Reading Literacy Study (PIRLS). The results of the 2011 NAEP grade 4 reading assessment were expressed in terms of the metric of the 2011 PIRLS assessment thereby providing international benchmarks for the NAEP grade 4 reading achievement levels. The report based on the results of this study can be found at: <http://files.eric.ed.gov/fulltext/ED545246.pdf>.
6. *NAEP-HSLS Linking Study*: Data for students who participated in both the 2013 NAEP grade 12 assessments and the 2009 High School Longitudinal Study (HSLS) were linked so that information from the HSLS student and parent questionnaires could provide a broader context for understanding NAEP results. In addition, the study explored using the relationship between the HSLS questionnaire variables and NAEP scores to predict NAEP mathematics scale scores for the full HSLS sample. The NAEP/HSLS overlap sample is being used to conduct studies on: mathematics motivation and its relationship with mathematics performance; predicting college admissions immediately out of high school from grade 12 NAEP mathematics scores; college enrollment benchmarks on the grade 12 NAEP mathematics assessment; noncognitive factors as predictors of college outcomes; STEM course-taking in high school in the prediction of grade 12 NAEP mathematics scores; and studying students' entrance into STEM fields. This research has not yet been published.

7. *NAEP-PISA Linking Study*: NCES conducted a pilot study to investigate the feasibility of creating a statistical link between the NAEP mathematics scale and the Program for International Student Assessment (PISA) mathematics literacy scale. Two states that participated in the 2013 NAEP state-level 12th-grade pilot and had participated in the 2012 PISA were included in this study. In each state, additional samples of students in grades 9, 10, and 11 were administered a version of the NAEP mathematics assessment. Although it was determined that establishing a statistical link between NAEP and PISA may be feasible, the validity of the predicted PISA results requires further evaluation. Some questions remain to be addressed in terms of the validity of the linking results, such as the constructs measured by NAEP and PISA, the definition of the target populations between NAEP and PISA given differences in the timing window (different assessment years), exclusion policies, etc. Also, the NAEP-PISA link was established based on two states only, with no additional states for cross-validation. No report has been published on this work.
8. *NAEP-Lexile® Study*: The Lexile® framework and measures (owned by MetaMetrics®) include a vertical reading scale that spans grades 1 to 12, in addition to benchmarks for college and career readiness. The purpose of the study was to identify scores on the NAEP scale that correspond to preparedness benchmarks on the Lexile scale. To accomplish this link, a subsample of students in the 2013 NAEP assessment were administered Lexile items. The NAEP-Lexile study was successful in demonstrating a strong relationship between NAEP reading and the Lexile measure of comprehension ($r = .89$). At the recommendation of the NAEP Design and Analysis Committee, ETS attempted to further evaluate the validity of the linking by comparing the estimated proportions of readiness based on the state-level linking results with the proportions estimated by this linking study. Data from only two states were available and several more would be needed to evaluate consistencies among the results. So, at this stage, the validity or generalizability of the study results are still inconclusive. No report has been published on this work.
9. *NAEP-ACT Linking Study*: The purpose of this study was to identify a reference point or range on the NAEP 12th-grade reading and mathematics scales that might be associated with the ACT preparedness benchmarks, along with the point on the ACT scales that are associated with the NAEP Proficient level. The NAEP and ACT scores for grade 12 students who had taken both assessments in 2013 were the basis for this linking. A draft of this report was discussed with COSDAM at the March 2019 meeting; the final report has not yet been published but should be available soon.
10. *Linking of Grade 8 and 12 NAEP to State Longitudinal Data Systems in Select States*: As part of the Governing Board's preparedness research agenda, a variety of statistical linking studies were performed with the 2013 NAEP data. They include 1) linking of

NAEP and ACT with a group of select states, 2) linking NAEP and SAT scores within one state, and 3) linking grade 8 NAEP and ACT EXPLORE[®] with a group of select states. Pending availability of data and re-negotiated data sharing agreements, additional analyses may be performed with data from state longitudinal databases. The reports that are currently available have been shared with COSDAM during previous Board meetings and are linked below:

- NAEP-ACT EXPLORE Linking Studies: [TN](#), [NC](#), [KY](#)
- NAEP-SAT Linking Study: [MA](#)
- NAEP-ACT Linking Study: [TN](#), [MI](#)

11. NAEP-ECLS-K:2011: NCES conducted this study to link results from the Early Childhood Longitudinal Study-Kindergarten Class 2010-2011 (ECLS-K:2011) and the 2015 NAEP grade 4 reading and mathematics assessments. Students in the ECLS-K:2011 study who were also sampled for NAEP were asked to complete a supplemental SES-related questionnaire at the conclusion of the NAEP administration. These student responses were compared to responses provided by parents to similar SES-related questions. In addition, this study made it possible to explore predictors of NAEP reading performance based on data collected from kindergarten to third grade as part of ECLS-K:2011; and a similar study is underway for mathematics. The NAEP/ECLS-K:2011 overlap sample has been used to analyze reading and mathematics growth patterns from kindergarten to grade 4. This work has not yet been published.

12. Linking of NAEP to State Assessments for Grade 4 and 8 Reading and Mathematics: The NAEP Validity Studies (NVS) Panel, under contract to NCES, has linking studies currently underway to compare 2017 NAEP grade 4 and 8 reading and mathematics assessments with a sample of 2-4 state assessments. The NVS work includes both an item comparison component and a statistical component to better understand the similarities and differences between the NAEP mathematics, reading and writing assessments and the current generation of states' mathematics and English language arts learning goals, as reflected in states' accountability assessments. The statistical component is intended to inform the item comparison analyses with regard to the relative difficulty (location on the NAEP scale) of the cognitively complex items on the college and career readiness aligned assessments. Performance data from students who participated in both their state assessment and NAEP will be used to link and then jointly scale each separate state assessment with the corresponding NAEP assessment.

What has been done with the reports and results from NAEP linking studies?

Results from NAEP linking studies primarily have been released as research and technical reports, in addition to conference presentations. As part of the Achievement Levels Work Plan, the Board intends to explore ideas for synthesizing results from multiple studies and making the findings more accessible to non-technical audiences. A technical memo providing suggestions for approaching this has been commissioned as part of the Board's Technical Support contract with the Human Resources Research Organization (HumRRO). During the May 2020 Board meeting, COSDAM and the Reporting and Dissemination (R&D) Committee may meet in a joint session to discuss ideas for maximizing the utility of NAEP linking studies by providing context for how NAEP relates to other assessments and important indicators of student achievement.

Can we perform additional NAEP linking studies?

Periodically, the Board provides input to NCES on the desirability of additional NAEP linking studies. There are several factors that affect the feasibility of undertaking new linking studies, including:

Alignment of the NAEP Assessment Schedule with other assessments of interest: Most linking studies are based on the same sample of students (or randomly equivalent groups of students) taking NAEP and another assessment. The Governing Board has taken this into account when making some decisions about the assessment schedule, such as ensuring that the NAEP mathematics and science assessments are administered in the same years as the TIMSS mathematics and science assessments. In some cases, the administration years do not always align due to variation in periodicity (e.g., PIRLS is administered every 3 years and NAEP reading is administered every two years, so they only overlap once every 6 years).

Agreements with other testing and/or survey programs: In order to connect additional data to students in the NAEP sample, special permission is needed. This is the case whether the study involves another NCES data collection or an agreement with a state agency or external testing program. The data sharing agreements with state agencies and external testing programs typically have taken multiple years to negotiate with legal and contracts departments of multiple parties and have been very labor intensive. Many data sharing agreements are very specific in terms of what research questions can be addressed by the scope of the study; additional questions cannot be added later unless the agreements are re-negotiated.

Funding: Undertaking new linking studies is also a function of available funding. Some linking studies are much more expensive than others (e.g., when additional data collection is required compared to a naturally occurring overlap of samples).

Content similarity: In order for results from a linking study to be useful, a precursor step is to evaluate whether the constructs measured by NAEP and the other assessment are similar enough to allow for meaningful comparisons. Content alignment studies are generally performed in advance of conducting statistical linking studies.