National Assessment Governing Board
Committee on Standards, Design and Methodology
Friday, November 16, 2018
10:30 am – 12:30 pm
AGENDA

10:30 – 10:35 am	Welcome and Review of Agenda *Andrew Ho, COSDAM Chair*	
10:35 – 10:40 am	**ACTION:** Policy Statement on Developing Student Achievement Levels for NAEP *Andrew Ho*	See Achievement Levels Policy Tab
10:40 – 11:20 am	Update on Implementing the Board's Response to the Evaluation of NAEP Achievement Levels *Sharyn Rosenberg, Assistant Director for Psychometrics* *Andrew Ho*	Attachment A
11:20 – 11:25 am	Questions on Information Items (see below)	
11:30 am – 12:30 pm	Joint Session with Reporting & Dissemination Committee: Communication and Interpretation of NAEP Achievement Levels (SV #3) *Rebecca Gagnon, R&D Chair* *Andrew Ho*	Attachment B
	Information Item Update on Implementing the Strategic Vision (SV #2-10)	Attachment C

Implementing the Governing Board's Response to the 2016 Evaluation of NAEP Achievement Levels

The final report of the most recent evaluation of NAEP achievement levels was released on November 17, 2016; a free PDF of the full report can be downloaded at: https://www.nap.edu/catalog/23409/evaluation-of-the-achievement-levels-for-mathematics-and-reading-on-the-national-assessment-of-educational-progress. The Governing Board received a briefing from staff at the National Academies of Sciences, Engineering, and Medicine and members of the interdisciplinary review committee during the November 2016 Board meeting. As required by law, the Governing Board adopted a formal response to the evaluation (see attached) that was sent to the Secretary of Education, the Committee on Education and the Workforce of the House of Representatives, and the Committee on Health, Education, Labor, and Pensions of the Senate on December 20, 2016.

Over the past couple of years, COSDAM has had several conversations about how to implement various aspects of the Board's response to the evaluation. The table below summarizes the current status, recent work, and planned next steps for each of the recommendations. During the upcoming November Board meeting, COSDAM members will engage in discussions about how to proceed with various activities, including potential collaborations with NCES and the Reporting and Dissemination Committee.

Recommendations and Board Response	**Current Status and Recent Work**	**Planned Next Steps**
Recommendation #1: Evaluating the alignment of NAEP achievement level descriptors (ALDs) *The Governing Board intends to issue a procurement for conducting studies to achieve this goal* *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 ALDs in all subjects and grades (in conjunction with Recommendation #3)*	The proposed revised policy that COSDAM has been working on since March 2017 addresses ALDs, process for reporting ALDs, periodic review of all ALDs In July 2018, HumRRO convened an expert panel to discuss the approach for review and revision of ALDs and developing reporting ALDs under the proposed revised policy (for more detail, see the attached minutes) In August 2018, COSDAM had a preliminary discussion about this work and agreed that reporting would begin with the 2021 math and reading results	After the proposed revised policy is adopted, the procurement work can begin, including any additional COSDAM discussion of the study design There are funds in the Governing Board's Fiscal Year 2019 budget to issue a procurement for this work

Recommendations and Board Response	Current Status and Recent Work	Planned Next Steps
Recommendation #2: Determination of the trial status of the NAEP achievement levels *The NCES Commissioner is responsible for determining whether the trial designation is removed*	The evaluation stated that only the first recommendation (followed by an additional evaluation) was necessary for removing the trial designation, but this decision is at the discretion of the NCES Commissioner	Implement the Board's response to this evaluation NCES will determine the appropriate time for conducting an additional evaluation At the completion of a new evaluation, the then-current NCES Commissioner will make a determination about whether the trial designation should be removed
Recommendation #3: Regular recurring reviews of the ALDs *The revised policy will include a statement of periodicity for conducting regular recurring reviews of the ALDs, with updates as needed*	This is addressed by Principle 4 (Periodic Review of Achievement Levels) in the proposed revised policy	Adopt the proposed revised policy Review and revision of math and reading ALDs will occur first, but other subjects will follow shortly thereafter after the policy is adopted
Recommendation #4: Relationships between NAEP achievement levels and external measures *The Governing Board and NCES have some additional linkages planned and underway* *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*	Governing Board and NCES staff presented an initial idea of how to synthesize existing linking study findings at the May 2017 COSDAM meeting Additional information about current and possible future linking studies was shared with COSDAM in August 2017 and March 2018	Additional work is needed to figure out a comprehensive plan for conducting and reporting this research in collaboration with NCES

Recommendations and Board Response	Current Status and Recent Work	Planned Next Steps
Recommendation #5: Interpretations and uses of NAEP achievement levels *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*	To inform the statement of intended uses of NAEP, HumRRO has been working on a synthesis of existing uses of NAEP, which can also inform future data collection efforts There have been several COSDAM discussions about developing a statement of intended uses for NAEP, including for the achievement levels (most recently at the August 2018 COSDAM meeting; see attached) There was a discussion about considerations for developing a validity argument for the NAEP achievement levels at the March 2018 COSDAM meeting (see attached technical memo) The proposed revised policy refers to interpretative guide to accompany NAEP releases; there will be a discussion of how to approach this guide (and defining uses of NAEP achievement levels) at the upcoming November 2018 joint session with R&D	Develop statement of intended uses and interpretations of the NAEP achievement levels Perform additional research to better understand how various audiences are interpreting achievement levels and how communications can be improved Develop interpretative guides to be linked to the Nations Report Card Develop a validity argument for the NAEP achievement levels based on the intended uses and interpretations Perform additional outreach Develop a comprehensive plan to achieve the steps above, in conjunction with the R&D Committee and NCES
Recommendation #6: Guidance for inferences made with achievement levels versus scale scores *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*	This work is contingent upon having a statement of intended uses of NAEP and should be incorporated into the interpretative guide, in addition to other communication materials	Develop a plan to implement this recommendation, in conjunction with the R&D Committee and NCES

Recommendations and Board Response	**Current Status and Recent Work**	**Planned Next Steps**
Recommendation #7: Regular cycle for considering desirability of conducting a new standard setting *The Governing Board will update its policy on setting achievement levels for NAEP to be more explicit about conditions that require a new standard setting*	This is addressed by Principle 4 (Periodic Review of Achievement Levels) in the proposed revised policy	Adopt revised policy statement Consider whether any current standards should be reviewed

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.

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

Expert Panel Meeting on NAEP Achievement Level Descriptions

Meeting Minutes

Prepared for:	National Assessment Governing Board 800 North Capitol St., NW, Suite 825 Washington, DC 20002-4233 Attn: Sharyn Rosenberg, Asst Director for Psychometrics
Prepared under:	National Assessment Governing Board 800 North Capitol St., NW, Suite 825 Washington, DC 20002-4233 Attn: Munira Mwalimu, Contracting Officer Contract # ED-NAG-17-C-0002
Authors:	D. E. (Sunny) Becker Arthur Thacker Monica Gribben Hillary Michaels Sheila Schultz
Date:	September 14, 2018

Headquarters: 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314 | Phone: 703.549.3611 | humrro.org

Notes of the Expert Panel Meeting on NAEP Achievement Level Descriptions July 12-13, 2018 National Assessment Governing Board

Contents

List of Figures

Notes of the Expert Panel Meeting on Achievement Level Descriptions

July 12–13, 2018
National Assessment Governing Board

At the request of the National Assessment Governing Board, HumRRO organized and facilitated a meeting with a select group of leading experts in assessment and achievement level setting. The purpose of this meeting was to elicit input about proposed changes to the development and use of NAEP achievement level descriptions (ALDs) in general and about how to approach the first recommendation from the recent evaluation of NAEP achievement levels (National Academies of Sciences, Engineering, and Medicine, 2017), which is focused on reviewing and revising the NAEP ALDs for mathematics and reading. The meeting was timed to share high level outcomes with the Committee on Standards, Design and Methodology (COSDAM) as they consider policy changes and next steps for addressing the first recommendation from the evaluation.

The purpose of this document is to summarize the themes and comments from the rich discussions at the meeting.

Background

We were fortunate to assemble an exceptional panel of experts (hereafter referred to as "the Experts"): **Dr. Susan Davis-Becker**, ACS Ventures; **Dr. Karla Egan**, EdMetric; **Dr. Steve Ferrara**, Measured Progress; **Dr. Ed Haertel**, Stanford University; **Dr. Andrew Kolstad**, P20 Strategies, **Dr. Susan Loomis**, Consultant; **Dr. Barbara Plake**, University of Nebraska-Lincoln, and **Dr. Lauress Wise**, HumRRO. The full list of meeting attendees is included in Appendix A.

Four papers on aspects of ALDs were commissioned and provided as read-ahead materials in advance of the panel meeting:

- *Reporting Achievement Level Descriptors for the National Assessment of Educational* Progress; Dr. Hillary Michaels, HumRRO; Dr. Karla Egan, EdMetric; Dr. Art Thacker, HumRRO; and Dr. Sheila Schultz, HumRRO
- *Validating Achievement Level Descriptors*; Dr. Marianne Perie, University of Kansas
- *Anchor Studies for Analysis of NAEP Achievement Levels*; Dr. Susan Loomis, Consultant
- *The Basis of Scale Anchoring in Item Mapping: Some Issues of Concern*; Dr. Andrew Kolstad, P20 Strategies

The meeting was held on July 12–13, 2018 in Alexandria, Virginia. In advance of the meeting, the Experts received an agenda and the following read-ahead materials: (a) the National Academies of Sciences, Engineering, and Medicine's *Evaluation of the Achievement Levels for Mathematics and Reading on the National Assessment of Educational Progress*, (b) the Governing Board's formal response to the evaluation of NAEP Achievement Levels, (c) the four papers listed in the above bulleted list, and (d) an example of a previous NAEP anchor study. Appendix A contains the agenda and list of read-ahead materials.

Dr. Sunny Becker, HumRRO, welcomed the Experts and led the attendees through introductions. She also reviewed the agenda and stated the goals for the meeting.

The morning of the first day was devoted to a series of presentations. Copies of all presentation slides are in Appendix B. The remainder of the meeting comprised group discussions of several topics, including (a) use and range of reporting ALDs, (b) methodology, (c) panelists and procedures, (d) special considerations for mathematics and reading ALDs, (e) special considerations for reporting ALDs when setting new achievement levels, (f) validation/vetting of reporting ALDs, and (g) recommendations for special studies. Because this meeting was held so close in time to the COSDAM meeting it was meant to inform, there was insufficient time to prepare and vet these meeting notes to submit to COSDAM. In lieu of these notes, Dr. Sharyn Rosenberg offered a summary of important themes for the Governing Board and COSDAM to consider regarding the development and revision of reporting ALDs; the Experts agreed to this summary.

Presentations

To set the context for the Experts, **Dr. Sharyn Rosenberg**, Assistant Director for Psychometrics on the Governing Board staff, provided an overview of NAEP achievement level setting, including current and proposed roles of the ALDs throughout the process. Figure 1 depicts the current process,

Figure 1. Depiction of ALD development: Current Governing Board policy and procedures

Table 1 describes the types of ALDs, and uses of each, according to the current policy and procedures.

Table 1. Types and Uses of ALDs: Current Governing Board Policy and Procedures

Policy definitions	The current policy on *Developing Student Performance Levels for NAEP* defines three NAEP achievement levels: *Basic*, *Proficient*, and *Advanced*. These policy definitions apply to all subjects and grade levels for which NAEP achievement levels are set.
Preliminary ALDs	The current policy refers to preliminary ALDs, which are developed by the framework committee. These preliminary ALDs typically have separate statements for each content area and grade level, and are intended to inform item development (as described by the Board policy on *Item Development and Review*). The statements are written in terms of what students should know and be able to do.
Final ALDs	The current policy refers to final ALDs but is ambiguous about when the ALDs are revised and "locked down" from further changes. Since 1998, the preliminary ALDs have been reviewed and revised by a panel of content experts prior to beginning the achievement level setting activities. The rationale for finalizing the ALDs in advance of (rather than during) the achievement level setting meetings is to allow for thoughtful review and vetting (including public comment) on the final ALDs aimed to assure appropriate alignment to the policy definitions across achievement levels within each grade and across grades within each achievement level. The final ALDs typically contain summary statements for the

	subject overall and may also contain additional details by content area. The statements are written in terms of what students should know and be able to do.
Threshold 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 are developed by achievement level setting panelists. Panelists are told that the threshold ALDs are for their own use only and will not be reported with the NAEP results.
ALDs in NAEP reports	The ALDs currently included in NAEP reports are generally the same as the final ALDs used in the achievement level setting. An exception to this practice is when there were framework changes that required revisions to the ALDs (e.g., 2009 reading) or changes to cut scores that necessitated development of new ALDs (e.g., 1996 science) In several other cases, anchoring studies were conducted to evaluate the validity of the existing final ALDs but did not result in changes to the ALDs for reporting (see Loomis paper in Appendix A).

Dr. Rosenberg noted that the Governing Board is in the process of revising its policy on developing student achievement levels for NAEP. COSDAM developed a draft revised policy, but this document was not shared in the read-ahead materials because the initial full Board discussion was planned for the August Board meeting, shortly after the panel meeting took place.

In her presentation, Dr. Rosenberg shared excerpts from the proposed revised policy that were directly related to ALDs to inform the discussion:

- There are a few minor edits to the policy definitions for clarity, including: adding “NAEP” in front of *Basic*, *Proficient*, and *Advanced* to better differentiate the NAEP achievement levels from other uses of these terms; and removing the term “grade” to avoid confusion with “grade-level” performance.

- ALDs for a specific grade and subject (e.g., NAEP grade 4 mathematics) are collectively referred to as “content ALDs” to differentiate from the policy definitions (e.g., NAEP Proficient) that apply to performance on all NAEP assessments.

- Content ALDs are developed initially as part of the framework development process and may be revised to serve other purposes such as guiding an achievement level setting. The content ALDs that guide achievement level setting activities shall be written in terms of what students *should know and be able to do*.

- There will be no content ALDs developed for performance below the *NAEP Basic* level (consistent with the current policy).

- When content ALDs are reported with results (also referred to as reporting ALDs), they shall be written to incorporate empirical data from student performance. They shall describe what students *do* know and *can* do rather than what they *should* know and *should* be able to do (this represents a major change from the current policy and practices).

- There is a new principle on periodic review of achievement levels (and ALDs), to address one of the recommendations from the recent evaluation of NAEP achievement levels.

Figure 2 depicts the proposed revised process.

Figure 2. Depiction of ALD development: Proposed revised Governing Board policy and procedures

Table 2 describes the proposed types and uses of ALDs.

Table 2. Types and Uses of ALDs: Proposed Revised Governing Board Policy and Procedures

Policy definitions		The proposed revised policy on *Developing Student Achievement Levels for NAEP* defines three NAEP achievement levels: *NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced*. These policy definitions apply to all main NAEP assessments.
Content ALDs	ALDs in Framework (two types - for item development and achievement level setting)	Under the revised policy and procedures for framework development, the framework committee will develop content ALDs both by content area (to inform item development) and overall (for use in the achievement level setting activities). 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 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 continue to be developed by achievement level setting panelists. Panelists are told that the threshold ALDs are for their own use only and will not be reported with the NAEP results.
	Reporting ALDs	The proposed revised policy calls for conducting a study following an achievement level setting 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.

Dr. Rosenberg explained that she wanted to understand the technical feasibility and any challenges associated with developing reporting ALDs for NAEP based on empirical data.

She shared the first recommendation from the recent evaluation of NAEP achievement levels in mathematics and reading:

> "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" (National Academies of Sciences, Engineering, and Medicine, 2017).

In its formal response to this recommendation, the Governing Board pledged to conduct studies to review and revise the ALDs in math and reading at grades 4, 8, and 12. Dr. Rosenberg noted that one of the primary goals from this meeting is to seek input on considerations for conducting these studies.

Next, one author of each of the four commissioned papers summarized the paper and answered related questions.

Dr. Hillary Michaels shared information from a recent paper that she, Karla Egan of EdMetric, and other HumRRO colleagues prepared on developing reporting ALDs (Michaels et al., 2018). She emphasized the primary purpose of reporting ALDs is to communicate clearly to stakeholders the knowledge, skills, and abilities that students can demonstrate at each achievement level. Although reporting ALDs have been used for years, she noted the major challenge with reporting ALDs is that they typically are not developed specifically for who or how they will be used. She described four types of ALDs—of which reporting ALDs are one type—as well as the intended purpose and primary audience for each. Although ALDs are developed and used by various stakeholders, policy ALDs are often developed for policy makers and politicians; range ALDs are developed for teachers, content and curriculum experts and item writers; and reporting ALDs are developed to inform the general public and media representatives. Dr. Michaels also reviewed an example of policy, range, and reporting ALDs and discussed the type of information presented in each as well as the differences in level of specificity among them. She ended her presentation by having participants consider (a) what is the proper language to use for reporting ALDs (statements that use would vs. should vs. could), (b) whether content ALDs should be revised following standards setting (thereby creating final ALDs), (c) the extent to which information about use should be included in the dissemination of reporting ALDs, and (d) the importance of gathering validity evidence when developing and using reporting ALDs.

Dr. Marianne Perie joined the meeting remotely to provide a summary of approaches used by states to validate ALDs on their assessments (Perie, 2018). She discussed four goals for ALDs and the evidence states collect to demonstrate validity of their achievement levels: (a) ALDs should be fully aligned to the assessment; (b) ALDs should provide an accurate representation of student knowledge and skills; (c) ALDs should follow a clear progression across levels and grades; and (d) ALDs should be clearly written and easily understandable by a larger audience. Dr. Perie described five approaches states use to validate ALDs: (a) alignment, (b) item mapping, (c)item descriptors, (d) student mapping, and (e) survey. *Alignment* methods focus on using an evidence-centered design approach to match the rigor of the content in the ALD to the rigor of the framework and items. In *item mapping*, experts can review an item and match the content of that item to an appropriate ALD or they use item statistics to map the item to an appropriate ALD. The *item descriptors* approach is similar to item mapping, with the difference being the use of groups of items rather than individual items. Dr. Perie explained that a *student mapping* approach is not appropriate for NAEP because there are no individual scores. *Surveys*

are used to look at readability, clarity, appropriateness, and utility of items. Similar to what states often do, Dr. Perie suggested using multiple approaches for NAEP, including item descriptors and surveys of the intended target audience.

Dr. Susan Loomis provided an historical context by describing previous NAEP anchor studies (Loomis, 2018). The NAS Report recommended conducting studies that evaluate consistency among NAEP frameworks, item pools, ALDs, and cut scores. Although each study is designed to answer specific research questions, the general purpose of an anchor study is to determine the extent to which student performance, within achievement level ranges, demonstrates the knowledge, skills, and abilities described in the achievement level descriptors. Anchor studies are part of the validity evidence supporting the NAEP cut scores and ALDs.

Dr. Loomis highlighted several studies included in her paper, all of which used some type of item mapping technique:

- *Science 1996 Grades 4, 8, and 12 when data from standard setting panels were used as the basis for the Governing Board's adjustments to cut scores as recommended by standard setting panelists.* Two panels developed ALDs from the items order at response probability (RP) 0.67. The Governing Board adopted the ALDs developed by one of the panels.
- *Science 2009 Grades 4, 8, and 12 when all Grade 4 cut scores and Advanced cut scores for Grades 8 and 12 were changed.*
- *Reading 2009 Grades 4, 8, and 12 when a new framework was implemented, and the 1992 cut scores and score scale were unchanged*. Panelists used anchor descriptions to develop ALDs for reporting the 2009 Reading results relative to the 1992 Reading cut scores.
- *Geography 2002 Grades 4, 8, and 12 using all student data and conditional probabilities for performance scores.* This two-year study was conducted in phases. The first phase investigated whether student performance was represented by the ALDs developed in 1994. The second phase convened two independent panels of geography experts. One panel had experts familiar with NAEP while the other panel included experts who were not. Both panels wrote descriptions of items from the Grade 8 2001 assessment that had been anchored to the scale for the study. The results indicated a high level of consistency between the groups.
- *Mathematics 2003 Anchor Study Grades 4 and 8.* The study looked at the extent to which the changes in NAEP item pools over the years since the framework was first implemented had impacted the alignment of items with the 1992 ALDs. The studies found the ALDs remained generally consistent over time in describing what students should know and do even though item types had changed, particularly in geometry.

Dr. Loomis noted considerations for panelist recruitment and materials required to conduct anchor studies to review and revise the mathematics and reading ALDs. Based on previous experience, she suggested the panelists and facilitators should be qualified in their content area and be familiar with NAEP. She also recommended there be two panels of about six panelists each. Panelists should review items from more than one assessment cycle and would review all item types. The ALDs should be considered final until the framework on which they are based changes. In her experience, these guidelines have resulted in clear and generalizable results.

The following additional methodological considerations were discussed: (a) applying a consistent RP criterion, such as 0.67, in both standard setting and anchor studies ; (b) including a measure of discrimination; (c) not using a correction for guessing for anchor studies unless used for standard setting; (d)including a focus on items that do not anchor; (e) conducting two-

way comparisons of descriptions and alignment ratings; and (f)including evaluations by panelists at key points in the process.

Dr. Andrew Kolstad discussed concerns with the convention of using a response probability (RP) value of 67 for standards setting (Kolstad, 2018). His concerns stem from the potentially misguided idea of "item mastery." He described a standards setting in which panelists determine a point in an item booklet ordered by item difficulty where students described by an ALD no longer get items correct. An assumption underlying this method is that the student would get all of the prior (easier) items correct, and all the later (harder) items incorrect. However, if the student missed an earlier item, to get the same score, they would need to get one of the later items correct. Dr. Kolstad argued that this method would create a mismatch between the description of the student indicated by the ALD classification and the KSAs the student actually possessed.

One of the problems with using an RP value of greater than 50 is that it increases the number of false negative classifications of students. A greater number of students who can actually do the things indicated in the ALD will be classified into a lower category (false negative) than the number of students who cannot do those things will be classified into a higher category (false positive). Dr. Kolstad illustrated this phenomenon using a graphic of the items a particular student got correct versus those incorrect using an RP value of 80. In the graphic, he labels both the false negative and false positive classifications. It is clear from the illustration that error is skewed toward the negative when RP is greater than 50. The illustration from Dr. Kolstad's paper is reproduced below in Figure 3.

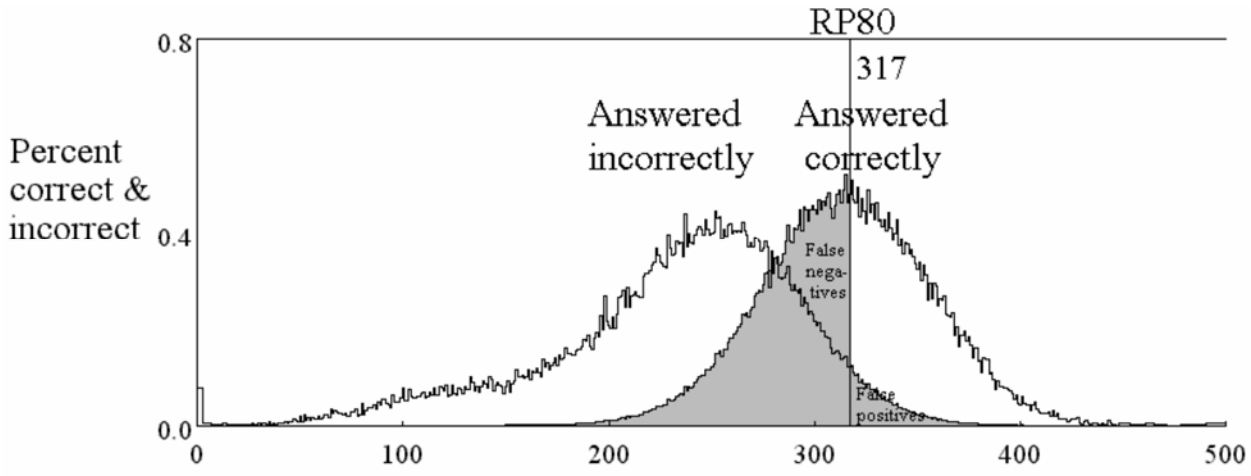

Figure 3. Hypothetical outcomes associated with RP80. (Figure reproduced from Kolstad (2018).)

Dr. Kolstad also provided examples of how basing ALDs on higher RP values can be very misleading, quoting a New York Times article indicating that half of Americans had limited proficiency with English. The article provided an illustrative item from the proficient category of an English assessment and indicated that half of Americans did not get questions similar to the illustrative question correct. In fact, 72% of Americans got the illustrative item correct, but the focus on interpreting "mastery" based on specific items led to the misunderstanding. Dr. Kolstad concludes his paper by describing four options for setting standards that do not rely on RP 67 and that would improve the correspondence between ALDs, test performance, and use of test information. Those options are reproduced in their entirety below.

Option 1: Use a judgmental process to choose an increment to the IRT b parameter that takes into account the principal policy uses of the data. For some policy purposes, the balance between false positives and false negatives may differ from those for other purposes. For example, if special services such as literacy remediation are going to be targeted to low performers, we ought to be very sure that they need the services by setting the response probability convention below 0.50 rather than targeting setting it above 0.50 to ensure that examinees are more than capable of the tasks they are set. Since this is a policy decision that depends on the purpose for which the data are expected to be used, there is no reason to rely on the conventional practice of a 0.65 probability. However, it is difficult for assessment programs with multiple uses to focus its procedures on any one expected use. This approach also uses different criteria to place cognitive items and examinees on an assessment scale and produces a degraded correspondence between the percentage of questions answered correctly and the percentage of items that meet the response probability convention.

Option 2: Map items by matching the distribution of scores in the population and the p-value of the item. By assigning to the test question the scale score corresponding to the point on the latent distribution at which the percentage of the population achieving at least that point matches the percentage of the population that answers the question correctly. This approach matches the population distributions of success on cognitive items and success at points along the proficiency scale. However, this method uses different criteria to place cognitive items and examinees on the assessment scale, places more stringent standards on easy questions and less stringent standards on harder questions (compressing the items together along the scale). When dealing with nontechnical, content-expert panelists, this method might need an explanation to understand how the difficulty of the individual items corresponds to scale scores.

Option 3: Map items using simple response patterns. Under this option, each cognitive question would be assigned the scale score that would be received by an examinee if that question were the most difficult item answered correctly in a simple scalogram pattern of responses. This approach produces a good match between the mapping of items and the percentage of correct answers needed to qualify for the corresponding score. However, this method places less stringent standards on easy questions and more stringent standards on harder questions spreading the items out along the scale). It uses similar, but not identical criteria to place cognitive items and examinees on the assessment scale. In my view, this method has an intuitive explanation that helps nontechnical panelists to understand why the probability of a correct response for an item on the margin is close to 0.50, yet the examinee possesses the ability to answer that or a similarly difficult question correctly.

Option 4: Map items at the IRT threshold parameter (without adjustment for guessing). This approach uses the same criteria to place cognitive items and examinees on the scale, places equally stringent standards on all items, produces a passable correspondence between the percentage of questions answered correctly and the percentage of items that meet the response probability convention, but when dealing with nontechnical, content-expert panelists, will need an explanation to counter the basic intuition about the lack of predictability about success with the individual items that they are responsible for examining closely.

Discussion

The agenda included seven topics to guide the discussion. Dr. Sunny Becker (HumRRO) facilitated a discussion around these topics and related ideas. However, the conversation appropriately meandered among and across the topics. Rather than forcing this summary to the

original discussion topics, ideas are presented here along the following themes that arose organically:

- Proposed revisions to the Board policy on achievement level setting
- Purposes of reporting ALDs
- Audiences for reporting ALDs
- Various item ordering methods for ALD development
- "Top down" and "bottom up" ALD development
- Panelists and procedures for validating ALDs and developing reporting ALDs
- Special considerations for math and reading reporting ALDs
- Steps to validate alignment of current ALDs and develop reporting ALDs
- Validation/vetting of reporting ALDs

We include discussion within each area and some limited repetition for context and continuity.

Proposed Revisions to the Board Policy on Achievement Level Setting

As noted earlier, Dr. Rosenberg shared select excerpts from the proposed revision of the Governing Board policy on developing achievement levels for NAEP. The Experts discussed the language of these revisions. They unanimously endorsed changing the policy labels to *NAEP Basic, NAEP Proficient,* and *NAEP Advanced* to differentiate the NAEP achievement levels from the use of basic, proficient, and advanced in state accountability systems, as well as other assessment programs.

The Experts stated that the proposed wording of Principle 4 – Periodic Review of Achievement Levels was confusing, specifically "past and recent administrations of NAEP assessments". They suggested substituting "current" or "recent" for "past and recent." Awkward wording aside, the Experts generally agreed with the stated frequency of revisiting ALDs "[a]t least once every 10 years or 3 administrations of an assessment, whichever comes later."

The Experts supported the proposal for reporting ALDs to describe what students *do know and can do* and for the ALDs used in achievement level setting to continue to describe what students *should know and be able to do*.

Purposes of Reporting ALDs

The Experts agreed that the purpose of producing reporting ALDs is to support accurate, credible, and defensible statements about NAEP findings reported in the *NAEP Basic*, *NAEP Proficient*, and *NAEP Advanced* achievement levels adopted by the Board. Further, the process to craft these reporting ALDs should inform the evolution of NAEP into the future, maximizing coherence and consistency as frameworks, item pools, and the technology of assessment evolve. This should include defensible standard setting, as well as periodic review of established cut scores.

These purposes imply more specific requirements. Alignment should be integral among frameworks, item pools, ALDs, and cut scores. The process for creation and review of reporting

ALDs must clearly and accurately communicate to multiple stakeholder audiences with overlapping but nonidentical information needs, and should forestall foreseeable misinterpretations, as feasible.

Audiences for Reporting ALDs

An important consideration prior to defining the structure, content, and specificity of reporting ALDs is to determine who the intended audiences are. While the group did not reach consensus on this, the Experts did note that multiple audiences may be interested in reporting ALDs, and they engaged in some discussion regarding whether different audiences might require different information. Potential audiences include state assessment directors, policy makers, and governors' chiefs of staff. Some of the Experts suggested, and others disagreed, that teachers, parents, and students might be considered target audiences.

Once the Governing Board identifies its intended audiences, the list of audiences could inform the best definitional structure and level of specificity for the reporting ALDs. The Experts suggested that a worthwhile step in advance of developing the reporting ALDs would be a study to present various versions of sample reporting ALDs to representatives of the target audiences, and then testing the audiences regarding their understanding of the ALDs.. Further, a step following development of the actual reporting ALDs might be to vet them with a sample of target audience members to ensure they are clear and do not result in misinterpretations. The Experts cautioned against creating a complex system with separate reporting ALDs for specific audiences.

Various Item Mapping Methods for ALD Development

To validate ALDs through anchoring studies, NAEP has typically used an item mapping method with an RP consistent with the one used in standard setting. In doing so, there is uniformity between the two tasks and in how the panelists think about the item. Currently, the RPs are set around 0.67, although reading and mathematics exemplar item selections in the past used an RP value of 0.50. Huynh (2000a & 2000b) suggests that RP 0.67 represents the maximum amount of item information for correct responses. As each NAEP Technical Report has pointed out, the maximum total information, for both correct and incorrect responses, is represented by an RP of 0.50. Other researchers found RP 0.67 to be a useful criterion for mastery because panelists can interpret it (e.g., NRC, 2006), as the response probability where 2/3 of students with a given cognitive score level would answer a question correctly. The item maps and exemplars available to the public are consistent with the definitions and descriptions of items determined by a high RP value, although the RP typically ranges from 0.65 – 0.74 depending on item type (e.g., see https://www.nationsreportcard.gov/itemmaps/?subj=MAT&grade=4 &year=2017).

As described in Dr. Kolstad's paper and presentation, high RPs have disadvantages. Different RPs yield different item locations and different cut scores. He demonstrated that mapping by a high RP value will lead to decreased false positive results (for students who may not have the skill, but still get credit for having knowledge of it) but at the same time will lead to increased false negatives (for students do have the skill, but do not get credit for having that knowledge). Moreover, he provides examples of how lay audiences misinterpret what the RP level of an item means to the underlying scale, item difficulty, total score, and the item. He believes that panelists can interpret RP 0.50 as the response probability of the most difficult item in a score-equivalent pattern of responses in which all items below it are answered correctly and all above it are answered incorrectly.

The group noted that regardless of what method and criteria are used to map items to the score scale, there are likely to be some items that do not map into an achievement level because they do not adequately differentiate students at one level from those at adjacent levels. The Board will need to decide how to handle misaligned items both in terms of the development of reporting ALDs and in terms of future administrations of the assessment. Some Experts suggested removing items from the current pool that do not match the ALDs or do not map to the cut score range for which the content matches the ALD. A large percentage of misaligned items could indicate a larger validity issue that would warrant additional study. The Experts noted that while items that map in the below Basic range would not be used in ALDs, they are essential to measuring a broad range of knowledge and skills.

Developing reporting ALDs should be focused on accurately reflecting what students in each category can do. Knowledge, skills, and abilities (KSAs) may be summarized for students right at the cut score, in the middle of the achievement level range, or at the very top of the achievement level range. In addition, the Experts discussed how reporting ALDs would be most useful if the descriptors included specific information on what *most* students at a particular level can do, what *many* students can do, and what *some* students can do in an achievement level range (e.g., Basic, Proficient).If NAEP results include reporting ALDs, score interpretation guides need to be developed to support appropriate score use.

This prompted the Experts to discuss a variety of methods and their rules for item mapping, since item location and ultimately, the ALD interpretation, will be impacted by the chosen method. The approach could change, based on the use and/or users of the reporting ALDs.

The Experts focused their discussion on these approaches:

- Map items at the IRT threshold parameter (without adjustment for guessing), such as RP 67. This is the current practice. The standard setting panelists are asked to think about where a barely qualified examinee has a 2/3 chance of getting the item correct. This is also described as the response probability of where 2/3 of minimally competent students would answer a question correctly. Once the standards have been set, the items can be sorted in the performance levels based on their parameters. The Experts noted that, once the content and skills of the items are described, the fact that the original order was based on an RP value is often lost.

- Map items using simple response patterns (Dr. Kolstad's preferred method) using a scalogram approach. This method relies on Guttman scaling or cumulative scaling that is often used to measure attitudes. Unlike item response theory, it is not probabilistic. The scale is a unidimensional continuum to help stakeholders predict item responses knowing the total (cumulative) score. However, scales are rarely perfectly cumulative, thus requiring scalogram analyses. In education, the scales are often used to obtain formative information against an expected learning progression. Guttman charts are frequently used in formative assessment because they portray what students know and can do. Guttman scaling has been used in NAEP research to support reporting domains (Schulz & Lee, 2002). To develop the ALDs, patterns of right and wrong responses could be compared. The interpretation becomes easier because it is consistent with stakeholder's instinctive understanding of higher and lower scores.

- Establish conditional p-values based on the students within any achievement level, such as Proficient. This method removes the items from their underlying scale. The p-values could be conditioned at 80% proficient, for example, to describe what *most* students in this achievement level can do, or at 60%, to reflect what *many* of the students in the

achievement level can do. These p-values can be used to identify items that exemplify what students within that category know and can do with different levels of certainty. This would result in a good description of the items that students within a group can do with different levels of probability/certainty, and this can be the basis for the ALDs.

The pros and cons of using RP values for mapping were discussed. The Experts did not come to consensus on whether to continue to approach the creation of ALDs based on a specific RP value (e.g. RP 67). There was concern that electing not to base item sets on RP 67, when that RP value was used during standard setting might create inconsistency in the overall system. Some of the Experts also endorsed the idea that "mastery" meant that students necessarily had more than a 50% chance of answering an item correctly. Others thought "mastery" was a flawed interpretation of a high RP value. Some of the Experts concluded that validation of ALDs based on items selected from an RP value designated for mastery is challenging and may lead to spurious conclusions based on skewed misclassifications.

Since the Governing Board's goal for reporting ALDs is to describe what students within a group know and can do, it may be more straightforward to eliminate RPs from consideration. Therefore, some Experts suggested using conditional p-values. The idea of mastery based on item content led to a discussion about the definition of mastery that should be applied to NAEP. There was some agreement that a small-scale study should be considered to attempt to create ALDs based on conditional p-values of items among students within each NAEP classification. However, not all Experts endorsed this idea.

"Top Down" and "Bottom Up" ALD Development

The Experts noted that the ALDs which are used by standard setting panels and item developers are developed in a top-down way. That is, policy definitions express what each achievement level means in a general, "high-level" sense, without reference to specific content areas or grade levels. These general guidelines drive interpretation of a framework for a specific content area and grade level to describe in greater detail what a student should be able to do. For example, Figure 4 shows the proposed revised policy ALD for *NAEP Proficient* performance and examples of ALDs that could be used to set achievement levels for the NAEP Proficient level at grades 4 and 8. Currently, these ALDs serve as the reporting ALDs for NAEP mathematics.

On the other hand, developing reporting ALDs based on the items that empirically map onto each achievement level (Figure 5) was described by the Experts as a "bottom up" process. The item information provides a smaller grain size than the frameworks, so summaries of what students actually know and are able to do within each achievement level are likely to be at a greater level of detail than summaries developed from the framework objectives. Any given item pool is necessarily a limited subset of the possible items posited by the framework, so the "bottom up" process would inevitably contain some "holes." . Developing reporting ALDs in this way, and then comparing those to the ALDs depicted in Figure 4, offers a rigorous confirmation that the frameworks, items, and cut scores, are all highly aligned. If the reporting ALDs contradict the ALDs used to set achievement levels, or if some of the specific statements are missing from one or the other, reconciliation would be necessary. Any bottom-up description cannot cover all that the framework intends. Small adjustments would not be problematic, but substantial discrepancies would require a deeper investigation and policy decisions.

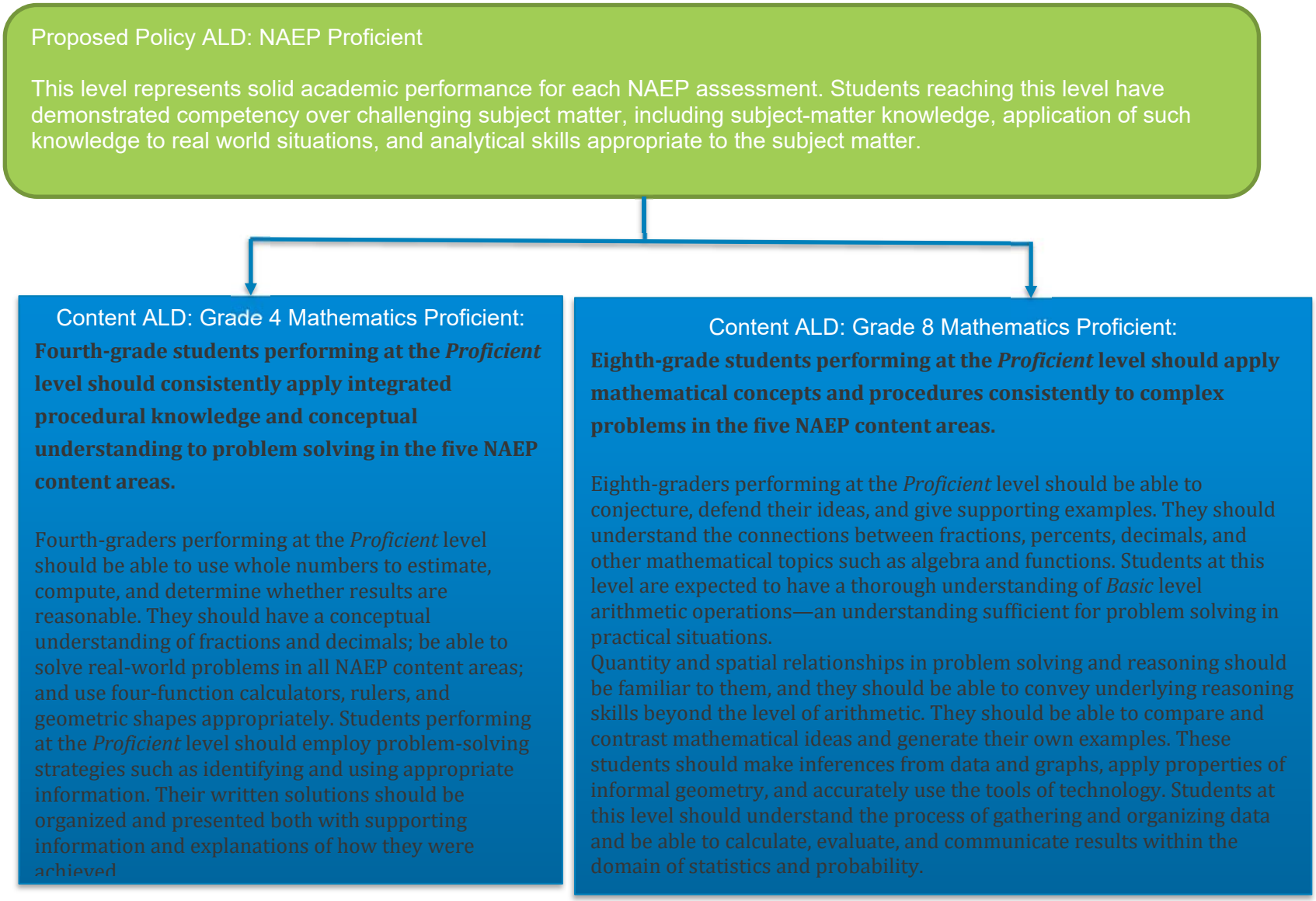

Figure 4. Top-down development of content ALDs from higher level policy ALDs.

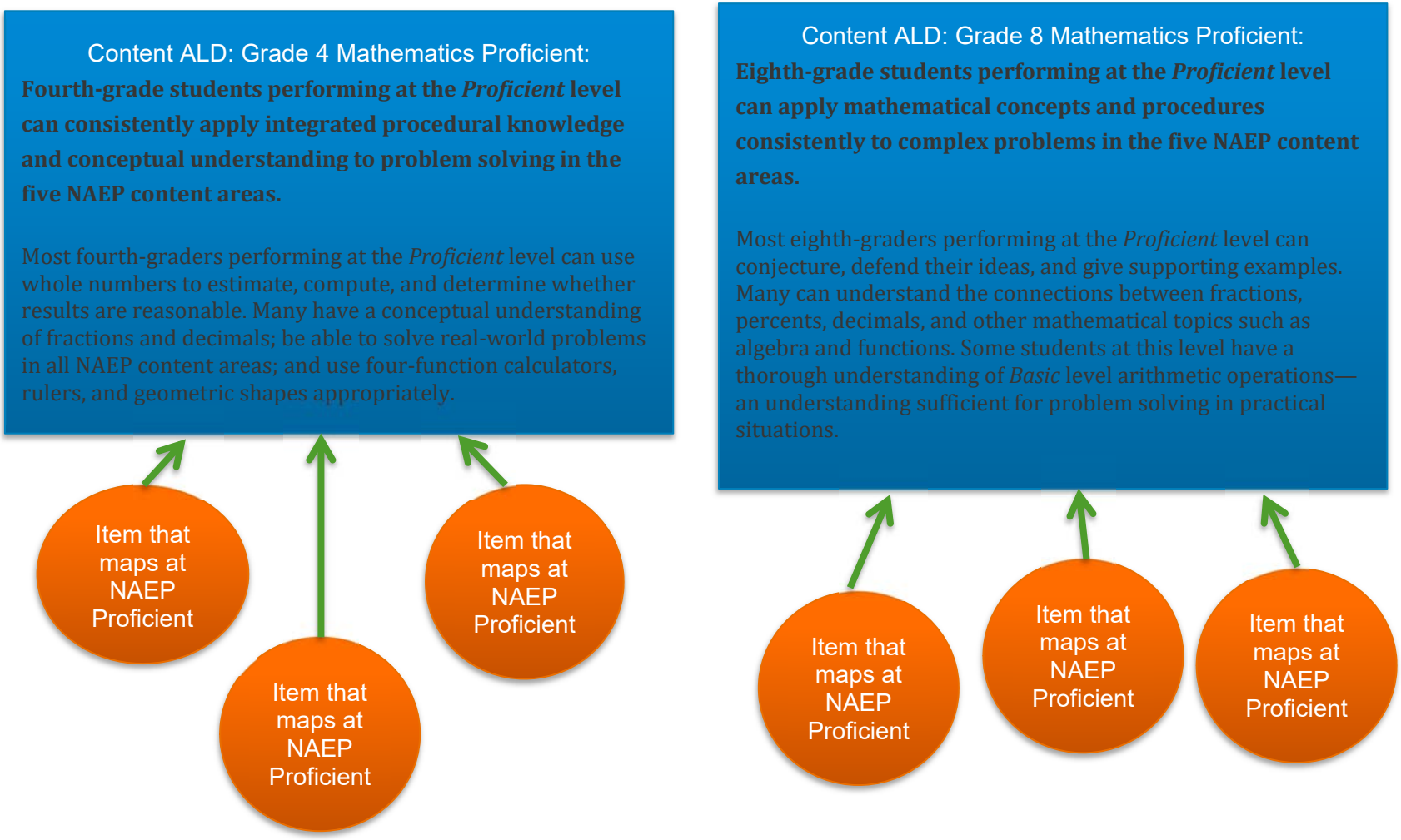

Figure 5. Bottom-up development of hypothetical reporting ALDs from items that map onto a given NAEP achievement level.

Panelists and Procedures for Validating ALDs and Developing Reporting ALDs

The discussion of who should serve on panels to validate existing ALDs and create reporting ALDs began with a review of the panelist qualifications for NAEP achievement level setting activities. The proposed revised policy states that achievement level setting panels shall consist of at least 50% teachers, with non-teacher educators (e.g., curriculum directors, academic coaches, principals) accounting for no more than half the number of teachers. The remaining panelists would be non-educators who represent the perspectives of additional stakeholders representing the general public, including parents, researchers, and employers. Panelists should 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; a general knowledge of assessment, curriculum, and student performance; and 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 Experts agreed that the achievement level setting panelists should not also create reporting ALDs, since the tasks are different, both activities are very time-intensive, and it may be confusing to perform multiple tasks with different instructions and context. There was less consensus on who should serve on panels to validate existing ALDs and develop reporting ALDs. There are benefits to including participants from the framework committee; they come to the task prepared with knowledge of the framework and rationale behind its development. They discussed mixed groups including some members of the framework committee and some new individuals who would have been qualified to serve on the framework committees but did not serve. The Experts felt that it would not be appropriate to include individuals who lacked content expertise in the subject.

In terms of the number of panelists, the Experts suggested including up to five or six panelists per grade and subject. With six panelists, two groups of three can divide the work or replicate the work in each grade/subject. For math, this would mean that there would be approximately 15-18 panelists across grades 4, 8, and 12. The panelists could sit at tables of 3 for some activities and then be grouped as tables of 6 for other activities.

In terms of procedures, the Experts proposed that panelists would need to engage in a validation step to compare results from an item mapping (or similar) procedure to the ALDs currently in use. Most Experts endorsed a strong design where panelists would write new ALDs based on the item mapping procedure. Then the new ALDs based on the empirical data would be compared to the ALDs currently in use (and in most cases, the ALDs that were used to set the achievement levels). Consistency checks between the two sets of ALDs would identify serious mismatches and serve as a validity check; this would effectively be a comparison of the ALDs produced from a “top-down approach” with ALDs produced from a “bottom up” approach. Training and facilitation are very important. The Experts noted that it is challenging to distill ALDs from item sets, and that panelists must be adept at recognizing when items may perform unexpectedly due to idiosyncratic reasons. The panelists would require training on how to take individual items and develop ALD statements, and specific examples would be needed.

Some Experts argued that it might not be necessary to write new ALDs to perform this validation step; instead panelists could be asked to judge the extent to which the current ALDs could be verified using the item mapping data. Other Experts were concerned that this procedure might be more open to bias and that it was important for NAEP to use the strongest design possible to maintain its reputation for being the gold standard, especially when conducting the studies for math and reading. The wording of the first recommendation in the evaluation implies that the

designs used for previous NAEP anchoring studies in reading and math are acceptable, and those studies did use the stronger design of writing new ALDs to compare to the existing ALDs.

Another option that was discussed but largely dismissed was asking panelists to look at items and predict which achievement level they should represent, and then compare those results to the actual results from an item mapping procedure. Several Experts noted that based on their experiences, it is very difficult for panelists to estimate the ALD targets for items.

Following the validation step, the Experts agreed that panelists would need to use the item mapping data to draft statements for reporting ALDs. For example, if the reporting ALDs are written in terms of the things that most/many/some students in each achievement level can do, then panelists would need to examine items and write descriptions based on the percentage of students who correctly answered (or would be expected to correctly answer) various items. Some Experts recommended using statements about what most/many/some students can do rather than what the "typical" student can do. The use of items for creating reporting ALDs should be limited to recent administrations of the assessment. Reporting ALDs should reference specific content and should provide users with as much specificity as they require, within the limits of good measurement practice.

If all of these steps were included, the meetings to validate ALDs and create reporting ALDs would be likely to take at least three or four days.

Special Considerations for Math and Reading Reporting ALDs

Dr. Rosenberg began a discussion of special considerations for developing reading and mathematics ALDs given the first recommendation of the recent evaluation of NAEP achievement levels. She noted because NAEP has transitioned to digital administration, these data currently exist for grades 4 and 8 in 2017; however, the grade 12 reading and mathematics assessments will not be administered digitally until 2019, so those data will not be available until late next year. She explained that NAEP's transition to a digitally-based administration involves "trans-adapting" the items from the paper-pencil to a digital assessment version, resulting in item parameters not being on the same scale. There was consensus among the participants there must be coherence among the reporting ALDs across the three grades; they believed there would be too much risk for inconsistency if the reporting ALDs were developed at different times. The participants suggested developing reporting ALDs at the same time for the three grades but using data from the grade 12 paper-pencil administration for a pilot study to develop draft reporting ALDs and then creating operational reporting ALDs when the grade 12 digital data become available.

Dr. Rosenberg raised concerns about being able to complete all activities so the newly developed reporting ALDs could be used to report the 2019 NAEP results. The Experts noted the tendency for the Governing Board to be deliberate when making changes to the NAEP program. They felt it was not appropriate for activities to be rushed to address the recommendation from the evaluation. The Experts felt strongly the Governing Board should conduct a feasibility study that includes one grade for each content area and a pilot study of the new reporting ALDs at all three grades prior to using them operationally. Additionally, they recommended that the Governing Board plan strategically so there will be sufficient time between the pilot study and operational study to make appropriate changes. The Experts suggested that the Governing Board conduct a special study to examine the efficacy of the new reporting ALDs and to determine their usefulness to the various stakeholders. They also suggested that the Governing Board conduct focus groups to obtain feedback about the use and interpretation of "most/many/some" phrasing in the new reporting ALDs. The Experts

suggested that the Governing Board staff prepare a recommended timeline to develop the new reporting ALDs, along with a rationale that includes the validity evidence that will be collected from each activity, so that Board members can appreciate what is needed to produce reporting ALDs that are useful and widely understood.

Steps to Validate Alignment of Current ALDs and Develop Reporting ALDs

Dr. Becker led the group through development of an explicit list of suggested steps to verify alignment between the item pools, cut scores, and ALDs and to develop new reporting ALDs for mathematics and reading at grades 4, 8, and 12:

1. Governing Board defines users of the NAEP reporting ALDs.
2. Conduct focus groups with users of the NAEP reporting ALDs to determine what they find useful.
3. Issue a Request for Proposal (RFP).
4. Prepare a design document for validating the alignment of items, cut scores, and ALDs and for developing new reporting ALDs; the design will include pilot grade 12 reporting ALDs based on paper-pencil data and operational ALDs based on the 2019 digitally-based administration.
5. Conduct a feasibility study that includes one grade each at reading and mathematics.
6. Conduct a pilot study at each of the three grades for reading and mathematics.
7. Evaluate the results from the pilot study, including by sharing the resulting reporting ALDs with potential users; if major changes are made to the process, then conduct another pilot study.
8. Use this process operationally in 2021 (using 2019 data rather than waiting for 2021 data) to report NAEP results for reading and mathematics.
9. Develop a process to evaluate and vet the reporting ALDs.
10. Develop communications and dissemination plans using the new reporting ALDs.

The Experts discussed the importance of clarity around the different types and sequencing of ALDs, including when ALDs and cut scores are adopted by the Governing Board.. At this point, an ALD alignment study must include decisions and actions regarding any items determined to be misaligned. The subsequent development process for reporting ALDs should include an anchor study/item mapping and an evaluation of alignment between reporting ALDs and content ALDs. Adherence to this formal process would culminate in reporting ALDs representing a coherent and consistent system of frameworks, item pools, ALDs, and cut scores.

Validation/Vetting of Reporting ALDs

The Experts discussed validation of reporting ALDs in terms of their accuracy and utility. They noted that NAEP is often considered to be the “gold standard” of assessments and in some ways, provides a methodological approach that may inform state and other assessments. They noted that the process for producing, monitoring, and managing reporting ALDs affords another opportunity to demonstrate best practices to benefit other testing programs.

First, using common item sets should lead replicate panels to come to essentially the same content for reporting ALDs in each of the categories. The Experts also discussed using both RP values and conditional p-values to create reporting ALDs, and then comparing them. However, there was not agreement on the RP value that would be most appropriate (50 or 67), so this method of validation may be more complex.

The Experts also suggested conducting focus groups or "market research" to determine whether the reporting ALDs were understood by users and met their needs. They identified several potential users (e.g. state education agency officials, governor's office staff, district-level education staff) who might use the reporting ALDs. The vetting of the ALDs would include exploration of potential misuses or misinterpretations of data, as well as gathering feedback on the utility and ease of interpretation of the reporting ALDs.

Summary and Reflections

Dr. Rosenberg expressed her appreciation for the Experts' insights. She highlighted the key take-away points she planned to share with COSDAM at their August meeting. The Experts agreed with this summary.

- Wording of proposed policy changes
  - Endorse new labels of NAEP Basic, NAEP Proficient, NAEP Advanced
  - "Past and recent administrations of NAEP assessments" is not clear.
  - Agree with reviewing reporting ALDs every 3 administrations or 10 years, whichever comes later.
- Articulate the current procedures for verifying that NAEP items are aligned to their frameworks. States are required to have independent reviews to evaluate the alignment between their item pools and frameworks.
- What are the interpretation and use arguments for NAEP ALDs?
  - What sources of evidence are needed?
- Suggestion to phrase reporting ALDs in terms of what most/many/some can do based on actual student performance.
- Use separate panels to set standards and develop reporting ALDs to avoid cognitive shift from one purpose to another.
  - For setting reporting ALDs, use a mix of panelists from framework committees and educators who were qualified to serve on the framework committee but did not.
  - This process may need 3-4 days.
- Be planful. Release of reporting ALDs in 2021 is more feasible than in 2019.
  - Use 2015 grade 12 paper-pencil results and 2017/2019 grades 4 and 8 DBA results. Update grade 12 when 2019 results are available.
  - Pilot test could use 2015/2017 data or 2017 grades 4 and 8 data, depending on how many items are needed.
- Identify validation steps needed for each part of the process (see list of 10 steps noted above to validate alignment of current ALDs and produce reporting ALDs).

References

Huynh, H. (2000a). On item mappings and statistical rules for selecting binary items for criterion- referenced interpretation and Bookmark standard settings. Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans, LA.

Huynh, H. (2000b). On Bayesian rules for selecting 3PL binary items for criterion referenced interpretations and creating booklets for Bookmark standard setting. Paper presented at the annual meeting of American Educational Research Association, New Orleans, LA.

Michaels, H., Egan, K., Thacker, A. and Schultz, S. (2018). *Reporting achievement level descriptors for the National Assessment of Educational Progress* (2018 No. 040). Alexandria, VA: Human Resources Research Organization.

Kolstad, A. (2018). *The basis of scale anchoring in item mapping: Some issues of concern. A white paper developed for the National Assessment Governing Board.*

Loomis, S. (2018). *Anchor studies for analysis of NAEP achievement levels. A white paper developed for the National Assessment Governing Board.*

National Academies of Sciences, 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.

National Assessment Governing Board (2016). *Response to the National Academies of Sciences, Engineering, and Medicine 2016 Evaluation of NAEP Achievement Levels.*

Perie, M. (2018). *Validating achievement level descriptors. A white paper developed for the National Assessment Governing Board.*

Pitoniak, M., Dion, G. and 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.* Prepared under contract to and in conjunction with the National Assessment Governing Board. Princeton, NJ: Educational Testing Service.

Appendix A: Meeting Agenda, Attendees, and List of Read-Ahead Materials

Expert Panel Meeting on NAEP Achievement Level Descriptions
National Assessment Governing Board Technical Support Project
July 12 –13, 2018 | Agenda

DAY 1

9:00 – 9:15	**Welcome, Introductions, and Meeting Goals**	Dr. Sunny Becker
9:15 – 10:00	**Setting the Context**	Dr. Sharyn Rosenberg
10:00 – 10:30	**Considerations for Reporting ALDs for NAEP**	Dr. Hillary Michaels*
10:30 – 10:45	Break	
10:45 – 11:15	**State Approaches to PLD Review and Revision**	Dr. Marianne Perie*
11:15 – 11:45	**History of Anchor Studies for NAEP**	Dr. Susan Loomis*
11:45 – 12:15	**Methodological Considerations**	Dr. Andy Kolstad*
12:15 – 1:00	Break for lunch	
1:00 – 1:15	Review and Revise Discussion Topics	
1:15 – 5:00	**Group Discussion** (break from approximately 3:00 – 3:15)	
	Use and Range of Reporting ALDs Methodology Panelists and Procedures Special Considerations for Math and Reading ALDs	
6:00	**Meet for optional group dinner**	

DAY 2

9:00 – 9:15	**Review of Previous Day and Plan for Today**	Dr. Sunny Becker
9:15 – 12:15	**Group Discussion** (break from approximately 10:30-10:45)	
	Special Considerations for Reporting ALDs When Setting New Achievement Levels Validation/Vetting of Reporting ALDs Recommendations for Special Studies	
12:15 – 1:00	Break for lunch	
1:15 – 2:45	**Group Discussion**	
	Decision Points for the Governing Board Summary of Recommendations and Next Steps	
2:45 – 3:00	**Wrap-up**	Dr. Sunny Becker

* Session will consist of a brief presentation by an author, reminding the Experts about content of a read-ahead document. This will be followed by clarifying questions.

Attendees

Expert Panelists:
Dr. Susan Davis-Becker, ACS Ventures, LLC
Dr. Karla Egan, EdMetric, LLC
Dr. Ed Haertel, Stanford University
Dr. Steve Ferrara, Measured Progress
Dr. Andy Kolstad, P20 Strategies LLC
Dr. Susan Loomis, Consultant
Dr. Barbara Plake, University of Nebraska-Lincoln
Dr. Laurie Wise, HumRRO

Governing Board Staff:
Ms. Michelle Blair
Dr. Sharyn Rosenberg
Dr. Lisa Stooksberry

HumRRO:
Dr. Sunny Becker
Dr. Monica Gribben
Dr. Hillary Michaels
Dr. Sheila Schultz
Dr. Arthur Thacker

NCES:
Dr. Enis Dogan

ETS (NAEP Design, Analysis, and Reporting Contractor):
Dr. Mary Pitoniak

On the phone (for part of the meeting):
Dr. Marianne Perie, University of Kansas

Read-ahead Materials

1. National Academies of Sciences, 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. **(see chapters 1, 5, 8)**

   A free PDF can be downloaded at: https://www.nap.edu/catalog/23409/evaluation-of-the-achievement-levels-for-mathematics-and-reading-on-the-national-assessment-of-educational-progress

2. National Assessment Governing Board (2016). *Response to the National Academies of Sciences, Engineering, and Medicine 2016 Evaluation of NAEP Achievement Levels.*

3. Egan, K., Michaels, H., Thacker, A. and Schultz, S. (2018). *Reporting achievement level descriptors for the National Assessment of Educational Progress* (2018 No. 040). Alexandria, VA: Human Resources Research Organization.

4. Perie, M. (2018). *Validating achievement level descriptors. A white paper developed for the National Assessment Governing Board.*

5. Kolstad, A. (2018). *The basis of scale anchoring in item mapping: Some issues of concern. A white paper developed for the National Assessment Governing Board.*

6. Loomis, S. (2018). *Anchor studies for analysis of NAEP achievement levels. A white paper developed for the National Assessment Governing Board.*

7. Pitoniak, M., Dion, G. and 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.* Prepared under contract to and in conjunction with the National Assessment Governing Board. Princeton, NJ: Educational Testing Service. **(example of previous NAEP anchor study)**

Uses of NAEP
For August 3, 2018 COSDAM Discussion

Over the past couple of years, COSDAM has had several discussions about the need to explicitly state how NAEP results (in general, and achievement levels in particular) are intended to be used, and then to focus dissemination efforts on increasing the most appropriate and impactful uses of NAEP. The very first standard (Standard 1.0) of the *Standards for Educational and Psychological Testing* states: "Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided" (AERA, APA, & NCME, 2014; p. 23).

The Governing Board's Strategic Vision includes a goal to expand the availability, utility, and use of NAEP resources, in part by creating new resources to inform education policy and practice (SV #3). COSDAM activities to address this goal include: conducting research on how NAEP results are currently used (both appropriately and inappropriately) by various stakeholders; developing a statement of the intended and unintended uses of NAEP data (in conjunction with NCES); and working with NCES to produce documentation of validity evidence in support of the appropriate uses of NAEP.

One of the major recommendations (Recommendation #5) from the recent evaluation of NAEP achievement levels is: *"Research is needed to articulate the intended interpretations and uses of the achievement levels and to collect validity evidence to support these interpretations and uses. In addition, research is needed 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"* (National Academies of Sciences, Engineering, and Medicine, 2017, p. 13). The proposed revised policy on Developing Student Achievement Levels for NAEP references an interpretative guide that would accompany NAEP reports and include specific examples of appropriate and inappropriate interpretations and uses of the results (Principle 3h).

As part of the Technical Support contract, the Human Resources Research Organization (HumRRO) has been conducting research on how NAEP results have been used by various audiences, including: federal, state, and local policymakers; educators; media; education researchers; and the general public. The first phase of this work (currently underway) is to analyze existing artifacts produced by these various audiences; a potential follow-up activity is to

conduct interviews and/or focus groups to gather additional information that cannot be gleaned from existing artifacts, if warranted.

As part of their work conducting the evaluation, the National Academies of Sciences, Engineering, and Medicine (2017) also conducted some research on how the achievement levels are being used; the evaluation report includes a summary of uses, interpretations, and actions for the NAEP achievement levels (p. 192-193).

Using preliminary findings from the research efforts referenced above, along with their own knowledge of common uses and interpretations, Governing Board staff developed two high level lists to support the August COSDAM discussion. The first is a list of primary uses (how different types of NAEP results are used); the second indicates secondary uses (common interpretations and actions based on those uses). The lists do not attempt to differentiate appropriate versus inappropriate uses and interpretations.

During the upcoming August Board meeting, COSDAM members will discuss how to use this information to inform next steps for: 1) developing a statement about appropriate uses of NAEP; and 2) developing an interpretative guide for communicating how the NAEP achievement levels should be used.

Discussion Questions

- What are the general principles and considerations for developing a statement of appropriate and inappropriate uses of NAEP?
- What are the general principles and considerations for developing an interpretative guide for communicating achievement level results?
- In order to develop these documents, is it necessary to gather additional information about how NAEP is used by conducting interviews and/or focus groups? If so, what are the priority audiences and questions to be answered?

References

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

National Academies of Sciences, 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.

Common Uses, Interpretations, and Actions Based on NAEP Data

Primary Uses

- Compare NAEP scale scores and/or achievement levels at a single point in time across states, districts (TUDA), and/or student groups
- Compare NAEP scale scores and/or achievement levels over time (trends) for the nation, states, districts (TUDA), and/or student groups
- Rank order states or districts in terms of NAEP scale scores and/or achievement levels overall and/or for a specific student group
- Analyze performance gaps in NAEP scale scores and/or achievement levels between two student groups at a single point in time
- Analyze changes in performance gaps in NAEP scale scores and/or achievement levels between two student groups over time (gap trends)
- Validate performance or changes in performance on state tests
- Analyze the relationship between contextual variables and NAEP scale scores and/or achievement levels
- Describe the context in which students learn from information gathered by student, teacher, and school questionnaires
- Compare NAEP scale scores and/or achievement levels across subject areas
- Compare NAEP scale scores and/or achievement levels across grades
- Compare NAEP scale scores and/or achievement levels before and after a program or policy is implemented
- Estimate the percentage of students who are academically prepared for college by the end of high school
- Show examples of what students know and can do through sample items and item maps
- Establish a common scale for linking state tests and comparing results across all school districts (e.g., Stanford Education Data Archive)
- Link other assessments to NAEP to provide state-level results on other assessments that were not administered at the state level (e.g., TIMSS)
- Establish a common scale for comparing the rigor of state standards to each other and to NAEP Proficient
- Compare the percentage of students at or above each achievement level on NAEP and on other assessments, including state and international assessments
- Serve as a benchmark of performance at NAEP Proficient to inform standard settings on other assessments

Secondary Uses

- To evaluate whether current programs and policies are effective
- To support the need for new programs and policies
- To influence decisions about funding for educational policies and programs
- To influence legislation
- To determine whether the nation, states, and/or TUDAs are making progress for students overall and/or selected student groups
- To evaluate the quality of education at a single point in time and/or over time
- To claim that some states and/or districts are doing a better job educating students based on their rankings on NAEP
- To identify where there are large performance gaps and/or interventions are needed
- To identify states and/or TUDAs who are doing something extraordinary so that best practices can be shared
- To criticize states for lying about the percentage of students at or above Proficient if it varies substantially from NAEP
- To generate and test hypotheses about factors related to student achievement (education research)
- To claim that students should do more of X because X is correlated with higher performance
- To determine whether U.S. students will be internationally competitive
- To call for higher standards
- To call for more accountability systems
- To claim that the majority of students lack basic skills (or are faring well)
- To make claims about the percentage of students who are performing “on grade level”
- To inform the development of state content standards

HumRRO
HUMAN RESOURCES RESEARCH ORGANIZATION

2017 No. 089

Memorandum #1: Considerations Related to the Validation of NAEP Achievement Levels

Prepared for:
Sharyn Rosenberg
National Assessment Governing Board
800 North Capitol Street N.W., Suite 825
Washington DC 20002
Phone: (202) 357 6940
Email: Sharyn.Rosenberg@ed.gov

Prepared under:
ED NAGB 17 C 0002
Contracting Officer: Munira Mwalimu
Phone: (202) 357 6906
Email: Munira.Mwalimu@ed.gov

Authors:
Arthur A. Thacker
Tanya Longabach

Date:
February 5, 2018

Headquarters: 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314 | Phone: 703.549.3611 | humrro.org

Memorandum #1: Considerations Related to the Validation of NAEP Achievement Levels

Table of Contents

List of Tables

List of Figures

Memorandum #1: Considerations Related to the Validation of NAEP Achievement Levels

Introduction

One common characteristic of educational assessments is the need to make broader inferences about students' knowledge and abilities from specific behaviors (Mislevy, Almond, & Lukas, 2003). Since we cannot directly see the knowledge and abilities we wish to measure, or to observe them in full, our measurement of those constructs is a proxy measurement. Therefore, we need to justify the inference that the observable behavior is a manifestation of the unobservable construct we are trying to measure. The ways that we interpret the score that students receive on an assessment depends on the inferences we make between the observed student behavior and the unobserved construct.

Validity is a property of the interpretations assigned to scores, and these interpretations are considered valid if they are supported by convincing evidence. In order to evaluate the plausibility of a test score interpretation, it is necessary to be clear about what the interpretation claims. That is, a claim should be made explicitly and directly about the inferences we intend to make. The interpretive argument specifies a network of inferences leading from the scores to the conclusions we intend to make based on those scores, as well as the assumptions supporting these inferences. In assembling and organizing evidence for the interpretive argument, we are developing a validity argument, the goal of which is to show that the interpretive argument is plausible (Kane, 2001). The process of developing the validity argument is known as validation. If the proposed interpretation of test scores is limited, as it is for some observable attributes, the requirements for validation can be very modest. If the proposed interpretations are more ambitious, as they are for traits and theoretical constructs, more evidence and more kinds of evidence are required for validation (Kane, 2013).

The *Standards for Educational and Psychological Testing* (AERA et al., 2014) place great importance on validity, calling it "the most fundamental consideration in developing tests and evaluating tests" (p.11). Specifically, Standard 1.0 states that "clear articulation of each intended test score interpretation for a specific use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided" (p.23). The associated standard cluster 1, including standards 1.1-1.7, elaborate on various aspects of validity that are essential to support assessment uses and interpretations.

Argument-based validation, as described by Kane (2006; 2013), primarily involves supporting the intended inferences that can be drawn from assessment scores. We typically begin by identifying the persons or groups that are expected to draw inferences from the test scores and we then describe those inferences in as much detail as possible. Once we understand the expected inferences, we can generate evidence to support the use of the test scores for those specific purposes. The National Assessment of Educational Progress (NAEP) is a very complex assessment system that does not produce individual students' scores. Many of the inferences that NAEP supports are quite different from most other student assessments.

The National Assessment Governing Board's (Governing Board) recent Strategic Vision[1] identifies policymakers, educators, researchers and business leaders, the media, and the general public as stakeholders who are expected to use NAEP results. The Strategic Vision is not so specific as to describe how each group is expected to use NAEP results, but it does indicate that they should be informed "about what America's students know and can do in various subject areas and compare achievement data over time and among student demographic groups." The Strategic Vision also states that NAEP should "inform education policy and practice."

The Governing Board is working towards developing a statement of intended and appropriate uses for both scale scores and achievement levels. HumRRO is currently conducting a research study to determine how various audiences have used and interpreted NAEP results. However, the current lack of specificity in the inferences each indicated group might make represents a substantial challenge for validation. For that reason, we will approach the creation of this section of the validity argument in two ways. First, we will address some of the most straightforward interpretations of NAEP results. These interpretations are well-described on the website[2] and are most commonly associated with the Nation's Report Card. We will not provide an exhaustive list of these interpretations and inferences here, but we will demonstrate a claim structure that might be used to support them. Then we will seek out inferences the identified groups have actually made from NAEP results. We will then describe how those inferences were supported and discuss additional claims and evidence that might be necessary for validation of those inferences.

Note that this memorandum is not comprehensive. Our goal is to provide guidance on how NAEP achievement levels might be validated for making specific inferences. The number of potential inferences that might be made and the amount of documentation available to potentially support those inferences is well beyond the scope of this memorandum. The examples we include in this memorandum, while important, do not necessarily represent the most important validation issues or interpretations of NAEP levels rather, they were chosen to be illustrative of the range of inferences. Where possible, we summarize the literature related to common claims, but these summaries do not represent an exhaustive literature review.

Summary of Achievement Level Descriptors Use and Interpretation.

Achievement level descriptors (ALDs) are the descriptions of knowledge, skills, and abilities of students at specific achievement levels. ALDs often include input from policymakers, stakeholders, and content experts. Egan, Schneider, and Ferrara (2012) identify three major uses of ALDs: standard setting guidance, test development, and score interpretation.

Some researchers identify standard setting as a primary use of ALDs. For example, Bourque (2000) said that the most important function of ALDs is considered to be providing "a mental framework or structure for standard setting panelists" (p.8). The clarity of ALDs is essential for setting meaningful cut scores (Kane, 2001): if ALDs are unclear, panelists cannot confidently determine how to sort examinees into groups based on achievement and set the cut scores. ALDs highlight what examinees need to accomplish to meet performance standards (Hambleton, Pitoniak & Copella, 2012).

[1] See https://www.nagb.gov/content/nagb/assets/documents/newsroom/press-releases/2016/nagb-strategic-vision.pdf.

[2] See www.nationsreportcard.gov.

Using ALDs to guide test development has been a topic of some debate. Some researchers suggest that ALDs can be used as a tool to guide the development of test blueprints, item specifications, and items themselves (Egan, Schneider, & Ferrara, 2012). While this idea makes sense, it is predicated on the ability of item writers to not only make judgments regarding the specific content that the item assesses, but also of the item difficulty, so that a wide range of items can be created that probe different ability levels as described in the ALDs. This use of ALDs may be challenging until it becomes clearer what factors affect item difficulty (Schneider, Huff, Egan, Tully, & Ferrara, 2010).

ALDs are an essential instrument of score interpretation; they were introduced in NAEP standard setting with the specific goal of making scale score interpretation easier and more meaningful (Kane, 2001; Bourque, 2009; Egan, Schneider, & Ferrara, 2012). Referencing performance categories (e.g. advanced, proficient, basic) used to divide a score reporting scale into ordered score intervals – rather than referencing the test scores themselves – may be a more understandable way of communicating test results (Hambleton, Pitoniak & Copella, 2012). With ALDs providing the descriptions of what the students at each of the performance categories know and can do, the stakeholders can easily see what abilities are associated with a scale score. ALDs give meaning to the cut scores established during a standard setting session.

National Academies of Sciences, Engineering, and Medicine (2017) outline the following purposes for having achievement standards:

- to be able to summarize students' present achievement and track their progress;
- to mark disparities between what we expect students to know and what they actually know;
- to stimulate policy conversations about educational achievement (and possibly discussions about methods of achieving the levels we want the students to be at);
- to identify content areas of high and low performance, as well as student subgroups of high and low performance; and
- to inform policy interventions and reform measures to improve student learning.

These uses of ALDs can at times be challenging to reconcile (Egan, Schneider, & Ferrara, 2012). For example, when ALDs are first created prior to a standard setting (so they can guide standard setters), they may be mainly aspirational; that is, they may articulate the policymaker's vision of the goals and rigor of achievement and answer the question "what should the students at specific achievement levels know and be able to do?" Later on, after the assessment data are collected and student scores are being reported by proficiency levels, the question being answered may change to "what do the students actually know?"

The validity of the assessment score inferences and ALD validity are interrelated. In an ideal situation, ALDs would guide the development of the test, so that the test is aligned with the construct of interest. The ALDs describe the degree to which students at each performance level possess this construct. The ALDs could then guide item writers in creating items that are aligned with this construct and elicit the knowledge that is aligned with the construct of interest. ALDs could also guide standard setters so they create cut scores with the same construct concept in mind as the item writers. Because the test is aligned with ALDs, and ALDs describe the degree to which the student possesses the construct of interest, the test assesses appropriate content. The ALDs used in score reporting, in turn, are aligned with test items and

represent the observed skills of students at a particular performance level. However, this process is seldom followed in reality (Egan, Schneider, & Ferrara, 2012). The disconnects between ALDs, cut scores, and the assessment itself, including assessment framework, items, and scoring, at different stages of the process may challenge the validity of ALDs.

Answers to the following questions would support the validity of the standards.

- Are the standards reasonable (based on a common understanding of what students should know and be able to do in the subject area)?
- Are the standards informative to the public?
- Can the public understand what students are expected to know and do?
- Do the standards lead to appropriate interpretations?

These general and typical purposes described above are consistent with the intended purposes of the NAEP ALDs as described in the Governing Board's *Strategic Vision*. The typical questions asked as part of the validation of standards are also applicable to the NAEP ALDs. After reviewing information related to the creation and use of the NAEP ALDs, we identified several issues that may represent challenges for their validation. These include:

- There is disagreement and/or confusion among stakeholders about how to interpret the meaning of "proficient" described by the NAEP ALDs.
- There has been disagreement from the beginning of NAEP administration regarding what the achievement levels should be; they have been declared "trial" and continue to have this status.
- The achievement levels are considered to be unreasonably high by some people.
- There is little guidance on how the achievement levels should be used and interpreted.

Our summary is very similar to validation challenges described by National Academies of Sciences, Engineering, and Medicine (2017): It remains challenging to find guidance on the intended interpretations and uses of NAEP achievement levels for stakeholders, including educators, administrators, and the public. The support for the uses of the achievement levels—the way that NAEP audiences use the results and the decisions they base on them – cannot be easily found. The guidance offered to users varies widely and is often delivered piecemeal, with important details spread across different web pages and reports. Users can obtain NAEP information at three separate websites: the Governing Board site (http://www.nagb.org); the National Center for Education Statistics (NCES) site (http://nces.ed.gov/nationsreportcard/); and a third called "The Nation's Report Card" (http://www.nationsreportcard.gov). There is some overlap across the three sites in the information available about NAEP, and all have links that take the user from one site to another. But interpretative guidance is uneven across the three, and it can be quite challenging to locate information about the achievement levels (Edley & Koenig, 2017).

Inferences from Various Stakeholders

Policymakers

For purposes of this memorandum, we define policymakers as national and state legislators, board and committee members at the federal, state, and district level who make policy and/or recommendations for policy in education, and other individuals who make or influence educational policy (e.g., congressional staffers, lobbyists). These individuals are responsible for policy across educational institutions and have considerable power to influence curriculum, instruction, assessment, teacher professional development, and other factors. They must address information regarding what students know and can do, and whether students are prepared for their next experiences, as policymakers strive to improve the state of American education.

Policymakers use NAEP scores and performance level descriptors for the following purposes:

- making comparisons to other districts, states, and the nation;
- making within-state subgroup comparisons;
- analyzing state achievement trends;
- suggesting changes to state assessments and to aid in defining levels of student performance;
- validating state standards and building the case for educational reform and change in their states (Zenisky, Hambleton, & Sireci, 2009); and
- building arguments for new or amended legislation and for requesting funding related to education (Edley & Koenig, 2017).

NAEP is well-structured in many ways for policymakers, who tend to be most interested in aggregate reports of student performance rather than individual student scores. NAEP is designed to generate comparable results across states and demographic groups. NAEP maintains a scale across years and allows for tracking of trends. However, when policymakers use NAEP to justify changes to state assessments or state performance definitions, build a case for educational reforms, or for requesting funding, they must support those uses based on their own understanding of NAEP and their judgements about NAEP's suitability for those purposes.

Educators

For purposes of this memorandum, we define educators as those persons who work most directly with students. They are responsible for instruction and for implementing curriculum and assessments. Educators include teachers, teachers' support personnel, content area specialists, academic coaches, etc. We also include school principals in this category, although there is some overlap with policymakers, since principals greatly influence policy within their particular schools.

Because NAEP does not produce results for individual students or at the school level, score interpretations are of limited use for educators. The ALDs and the frameworks, however, may provide considerable useful information. The frameworks indicate the content that students are expected to know in specific subjects at specific grades. The ALDs indicate how students will be categorized based on the level of their knowledge and skill related to that content. The ALDs help educators better understand how student performance is differentiated.

Educators receive their information about NAEP from various sources, including the three main NAEP websites mentioned earlier. They receive much of their information from their state education agency's website and the media. NCES also supports a NAEP state coordinator in each state who serves as a liaison between the state department of education and the NAEP programs. They are available to assist in the interpretation of NAEP results. We reviewed a sample of state websites as part of preparing this memorandum. We selected websites to reflect either high or low performance on NAEP to highlight any qualitative differences in the information presented to educators.

The three lowest performing states on NAEP 4th and 8th grade reading and mathematics and the three highest performing states based on 2015 results[3] are shown in Table 1. The state Department of Education (DOE) websites and state education agency websites were searched to determine whether and how the states use NAEP data. We specifically searched for information on using NAEP for standard setting purposes.

Table 1. Highest and Lowest Performing States on 2015 NAEP Reading and Mathematics, Grades 4 and 8

Subject/Grade	High Performing	Low Performing
Mathematics		
Grade 4	MA MN NH	AL NM MS
Grade 8	MA MN NH	AL CA MS
Reading		
Grade 4	MA NH VT	NM CA AK MS
Grade 8	NH MA VT	MS NM LA

There were both differences and similarities in how the low and high performing states referred to the available NAEP data. The low performing states provided much less information about participating in NAEP and the purposes of NAEP, in general, compared to the high performing states. High performing states, on the other hand, were more likely to provide details about student performance and participation on NAEP. Many state DOE websites include links to the state NAEP results on the Nation's Report Card website. Some state websites made a statement that comparisons can be made of how students from different states performed on NAEP, or reference studies that linked state standards to the NAEP standards. However, both low and high performing states provided little information about the explicit uses of the NAEP data for the purposes of creating state level ALDs and informing the determination of cut scores at the state level.

The websites did not include any explicit reference to whether or how NAEP standards may inform state performance standards, or how NAEP data may serve as impact data in state standard settings. The most explicit statement of the connection between state assessment and NAEP was found on the MA DOE website: "…NAEP has taken on a greater prominence under the No Child Left Behind Act and serves to externally confirm results of state assessments, such as the Massachusetts Comprehensive Assessment System (MCAS)" (National Assessment of Educational Progress Frequently Asked Questions, 2017)." The state of Vermont makes

[3]For more information see the website https://www.nationsreportcard.gov/profiles/stateprofile?chort=2&sub=RED&sj=AL&sfj=NP&st=MN&year=2015R3.

another explicit comparison between the structure of its own state science test and the NAEP science assessment standards: “The tests were designed to measure different standards, or frameworks, on separate scoring scales, but both assessments address similar skills and content areas. These assessments provide a way to reference national, state and local science achievement” (*Vermont Students Score among Best in the Nation on the National Assessment of Educational Progress*, 2016). The state also points out some similarities in the pattern of scores on both the state assessment and NAEP.

Among the state websites studied, most high performing statues reported:

- trends or comparisons of successive cohorts;
- comparison of the percentage of students at or above Proficient on NAEP to the percentage of students at or above Proficient on a state test;
- point-in-time comparisons across states, districts, or population groups (e.g., VT included information showing an increase in the performance of students of low SES);
- performance on subscales (e.g. algebra, vocabulary, etc.)
- rank ordering of states or districts;
- comparisons across population groups to examine performance gaps; and
- comparisons across subject areas.

Lower performing states tended to mention NAEP reports less often. However, we did find some information in the comments of school administrators to the media that NAEP results were used as an indication that the current state education system was in need of reform. For example, in 2013 the then-superintendent of Louisiana, John White, “used the [NAEP state achievement] report to reiterate his push for the Common Core national education standards. ’The growth this year was moderate. If we want to see something beyond incremental growth, we've got to raise our standards, and the Common Core standards is the best way to do that,’ he said“ (Bacon-Blood, 2013).

Researchers and Business Leaders

For purposes of this memorandum, researchers and business leaders include persons conducting educational research and individuals from private industry with an interest in elementary and secondary student performance. Currently, NAEP data use and interpretation research by these stakeholders may take the following directions (Edley & Koenig, 2017):

- track trends in and compare the performance of successive cohorts,
- make point-in-time comparisons across states and school districts,
- compare the performance of population groups within and across states (performance gaps),
- rank order the performance of states and compare state to national performance;
- compare performance across tested subject areas,
- examine relationships among student performance and selected student/school/family variables, and

- compare states' standards for proficient performance in reading and mathematics by placing them on a common scale defined by NAEP scores ("mapping studies").

Beginning with NAEP results from 2003, NCES conducted a series of studies that mapped each state's grade 4 and 8 reading and mathematics proficiency levels to the NAEP scale. This mapping was designed as a mechanism to evaluate the extent to which state standards reflected the same rigor as NAEP standards, and it was used as a policy lever to encourage states to set challenging standards for their students (Edley et al., 2017). In the mapping study report by Bandeira de Mello, Bohrnstedt, Blankenship, & Sherman (2015), the NAEP score that corresponds to a state's standard (i.e., the NAEP scale equivalent score) is determined by a direct application of equipercentile mapping. For a given subject and grade, the percentage of students reported in the state assessment to be meeting the standard in each NAEP school is matched to the point on the NAEP achievement scale corresponding to that percentage. The percentage of students passing the state standard was mapped onto the NAEP scores. The results are then aggregated over all of the NAEP schools in a state to provide an estimate of the NAEP scale equivalent of the state's threshold for its standard (Bandeira de Mello et al., 2015).

Peterson and Ackerman (2015) took a different approach to the comparison of state achievement scores and NAEP scores. They calculated the difference between the percentage of students considered "proficient" by both the state and NAEP assessments. The magnitude of the difference was considered to indicate how rigorous the state standards are as compared with NAEP standards.

These examples indicate that some researchers and policymakers do consider NAEP achievement levels to be a standard that states should strive toward. At the same time, some researchers caution against using NAEP as an infallible measure of state educational achievement due to fundamental differences between the state and NAEP frameworks and standards (e.g., Ho & Haertel, 2007). It is important to remember that determining the score equivalency between NAEP scale and state scale does not say anything about the equivalency or lack thereof in knowledge and skills associated with the score. The NAEP and state assessments may or may not measure the same knowledge and skills. An alignment study would need to be conducted to assess the extent to which the two assessments measured the same construct.

Many studies focused on validity evidence based on relationships with external variables, that is, setting benchmarks on NAEP that are related to concurrent or future performance on measures external to NAEP. Examples are academic preparedness for college; international tests; state tests and their alignment with NAEP (Edley et al., 2017). The studies indicate that there is considerable correspondence between the percentages of students at NAEP achievement levels and the percentages on other assessments (Gattis et al., 2016; Jia et al., 2014; Lim & Sireci, 2017; Neidorf, Binkley, Gattis, & Nohara, 2006; Phillips, 2014a, 2014b; Poland & Plevyak, 2015; Provasnik, Lin, Darling, & Dodson, 2013). These studies show that the NAEP achievement-level results (the percentage of students at the advanced level) are generally consistent with the percentage of U.S. students scoring at the reading and mathematics benchmarks on the Programme for International Student Assessment (PISA), the mathematics benchmarks on Trends in International Mathematics and Science Study (TIMSS), and at the higher levels for College Board Advanced Placement (AP) exams. For example, a report by Fields (2014) states that the content of the 12th grade NAEP reading and mathematics assessments was found to be similar to widely recognized tests used for college admission and placement. A linking study by Moran, Freund, & Oranje (2012) determined that there is a higher correlation between NAEP and SAT mathematics scores than between NAEP and SAT reading

scores The SAT reading benchmark, however, was closer to the NAEP Proficient score than the SAT math benchmark. Several studies investigated the relationship between NAEP Proficient and college and career readiness (Moran, Oranje, & Freund, n.d.; Schneider, Kitmitto, Muhusani, & Zhu, 2015), but the relationship was found to be fairly weak. Additional research in this area was proposed.

During the August 2016 Governing Board quarterly meeting, researchers provided the following recommendations regarding the use of NAEP data.

- Panelists urged the Governing Board to enable linkages from NAEP data to state-level or national-level to conduct research about the long-term effects of educational policies.
- All panelists agreed that while NAEP data describe trends in student achievement, the data do not support conclusions about the reasons for these trends. Additional research is needed to discover factors that can improve schools and student learning.
- It was suggested that the NAEP data be used to compare the performance of districts with similar demographic characteristics, such as poverty levels. NAEP data may be used to guide best practices on what works in the improvement of educational achievement.

The Media

While academic and research articles provide scientific, well-reasoned rationales for or against the specific interpretations of NAEP, articles by the media present a different side. They tell the story of those who are trying to use information under real-life conditions from the assessments that the academics are studying, and the real-world challenges and issues experienced by practitioners in the field.

Articles in publications like *Education Week* illustrate that there is a large degree of confusion accompanying the application and interpretation of NAEP standards. While many researchers and even state officials may assume the debate about the application of NAEP standards is resolved, magazine and newspaper articles question whether it is appropriate for states to incorporate NAEP standards into the standards of the state, and what the appropriate uses for NAEP scores are in general.

One point of argument is lack of clarity on the meaning of “proficient” and the application of that meaning to state standards. Not all media representatives consistently clarify for the public that NAEP Proficient is not grade-level proficiency and that NAEP Proficient is intended to be an aspirational standard. What makes this matter more complicated is that under the No Child Left Behind Act (NCLB), states had to create achievement levels that were grade-specific and most states chose to adopt the ALD title of “Proficient.” Reconciling these sets of standards causes additional conflict and confusion when states are trying to create their achievement levels and communicate them to the public. One suggestion to make the situation more understandable is for policymakers to explain to the stakeholders “what are good goals for educational purposes compared to what is appropriate for accountability when establishing cut scores on their state assessments” (Hull, 2008), why they may be different, and which performance levels are more appropriate for each specific purpose.

Many researchers are concerned that information from NAEP gets misinterpreted by the media and politicians, sometimes to serve the interests of specific groups. Various misinterpretations of NAEP results are frequently used by the politicians and media, giving rise to the term "misnaepery" (Sawchuk, 2013). One prominent example of this inappropriate interpretation includes tying an increase in state NAEP scores to some specific policy or intervention implemented by the state, and a decrease – to a policy that was proposed by an organization, but then not implemented. In practice, it is very challenging to make these causal connections. Organizations that are using NAEP scores to bolster claims about the effects of a specific policy are likely not interpreting the NAEP scores correctly (Chingos & Blagg, 2015).

A number of misinterpretations come from the misunderstanding of NAEP's definition of "proficient", with some reporters claiming that being below proficient means being "below grade level." Yet another source of confusion comes from comparing state assessment scores with NAEP scores and arriving at opposing conclusions. Comparing the achievement of different student population groups is often fraught with misinterpretations as well (e.g., treating the NAEP achievement scale as continuous between grades and comparing achievement of one population at a higher grade to the achievement of another population at a lower grade).

At least in part, these misinterpretations arise from a lack of readily available or accessible information on how the NAEP scores should be interpreted, what the appropriate uses of these scores are, and what conclusions are appropriate to make. Educational researchers call for using caution in deciphering which claims are appropriate, and discouraging the propagation of false claims about NAEP data interpretation (Polikoff, 2015a, 2015b).

The General Public

The general public may not have sufficient knowledge and training to deeply understand the intent and the meaning of state or national assessments, and may have a difficult time interpreting and critically evaluating information coming from various, often conflicting, sources. The media may make the situation in education appear more critical or negative than it really is. For example, if a state performs as one of the best on NAEP, but there is no growth in scores, the general public may see headlines like "Public education test results are dismal. Schools are failing NH children" (Levell, 2016). In addition, as mentioned earlier, the information provided by the media may not be completely objective, and score interpretations may be promoting a specific political agenda.

There is some confusion among the general public regarding why their state may have high scores on the state assessments, but low scores on NAEP (Weiss, 2016; Dillon, 2005). This may occur if the state set standards lower than NAEP standards, or if the state simply has different content standards. There may also be conflicting information on exactly how the state standards compare to NAEP standards; this may cause one study to claim that a state has low standards, and another study – that the state is either lagging behind others, or low on scores from some other perspective. A study by Achieve[4], describes several NAEP objectives at grade 4 contrasted with the grade those same objectives are introduced in several states' standards documents. The objective "Use simple ratios to describe problem situations," is typically introduced in grade 6 in many states. Discrepancies like this add complexity to potential comparisons between NAEP results and state testing results.

[4] See https://www.achieve.org/files/16-149_Achieve_NAEP%20math%20report.pdf.

One potential goal would be for the general public to be able to use state and national assessments to make decisions about whether children are getting the best education in their particular state. It is likely impossible to make such inferences at the school or even classroom level from state and national assessments. The media, however, may make it sound like those conclusions are appropriate and necessary. The same article by Levell (2016) that proclaimed the failure of New Hampshire public education, for example, suggests that, based on the fact that there was little to no growth in the student scores on state assessments or NAEP, the parents should "[e]ngage your local school board and question why they are using College and Career Readiness Standards and tests that are not providing a better education for our children;" consider a transfer to a charter or private school; or refuse to have their child take a state assessment. It may be helpful for the general public to have access to a source of clear, easy to understand, reliable information on the kinds of inferences that can legitimately be made from state and national assessments.

Approaching Validation of the NAEP Performance Levels Using a Validity Argument

A strong validity argument relies upon a foundation of thorough and specific definitions of the various purposes of the assessment. These purposes are typically illustrated via a ***Theory of Action*** *(*TOA) document or graphic. The TOA indicates the intended uses and expected impact of the assessment system. As depicted in Figure 1, the TOA can inform testable claims related to the interpretation of test scores. These testable claims represent the ***interpretive argument***. Every use or interpretation of an assessment score relies on meeting specific claims and the various assumptions that justify them. The evidence supporting those assumptions represents the ***validity argument***. The NAEP assessments represent a large number of potential interpretations/uses for test scores.

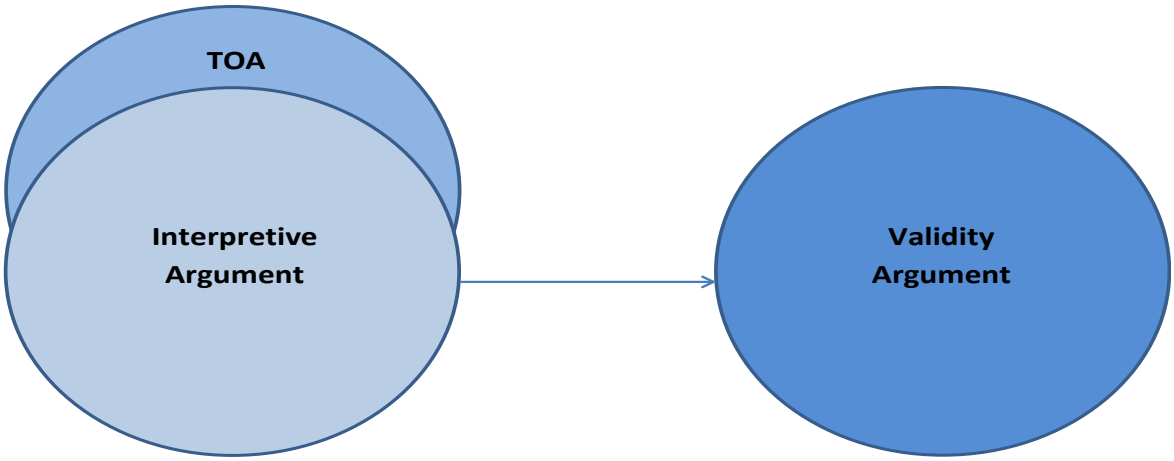

Figure 1. Relationships among theory of action (TOA), interpretive argument, and validity argument.

The Governing Board's Strategic Vision indicates that NAEP results should inform stakeholders "about what America's students know and can do in various subject areas and compare achievement data over time and among student demographic groups" (p. 1). The ALDs provide context for that goal by helping stakeholders interpret student performance in the various subject areas. Estimates of the proportions of students who would be classified as below Basic,

Basic, Proficient, or Advanced for each state, for select large school districts, and for demographic groups of students within them are reported. Reports are generated based on the performance of representative groups of students within those states and districts.

The subject matter content tested by NAEP and the ways student mastery of that content are operationalized in the achievement levels are described in the frameworks documents. These documents are vital to the TOA and to the interpretive argument. They describe what is tested on each of the NAEP subject tests and help us differentiate student performance into meaningful categories. If we were to construct a chain of logic, as is typically done in a TOA, the following assertions might be included.

> The subject area content included in the frameworks represents important key knowledge, skills, and concepts students should know at the indicated grade level.

1. The ALDs differentiate important differences in students' mastery of the content included in the frameworks.
2. NAEP assessments allow for strong estimates regarding the proportions of students scoring in each of the performance categories.
3. Score reports, or report cards, can be referenced to the frameworks and ALDs to interpret what students within a given state or large district know and can do.
4. Comparisons across states, large districts, and demographic groups allow stakeholders to identify gaps in terms of what students know and can do.
5. Stakeholders use NAEP performance information to better understand student achievement in their efforts to improve the education of American students.

The next step toward constructing the validity argument is to use the chain of logic from the TOA to describe how inferences from test scores are used by stakeholders in the process of achieving the goals of the testing program. When we consider the interpretive argument, we are forced to imagine the role of the various stakeholders. As an example, if we were to assume the role of a state education agency stakeholder, we might interpret NAEP results in the following ways, among others.

1. My state NAEP scores provide a snapshot of student performance for the current year's students' performance in the tested subjects.
2. My NAEP scores represent student achievement for the academic content the students are expected to learn, as described in the NAEP framework for each subject.
3. My state scores can be directly compared to other states and those comparisons will tell me if my state is preparing students as well as other states.
4. Demographic groups of students can be compared to each other for my state, and those comparisons give me information about performance gaps among those groups.
5. By comparing demographic group performance across states, I can determine if my state's performance gaps are larger or smaller than the gaps in other states.
6. The proportions of students from my state in each performance level are in those levels because of differences in their preparation related to knowledge, skills, and abilities as described in the ALDs.

7. I can directly compare my NAEP results this year to prior year's results to determine if students in my state are improving, declining, or staying at about the same level in the tested subjects and grades.

The next step in the process of building a validity argument would be to support the inferences described above through a claims and evidence structure. The claims are usually written as a series of "if…then" statements. The claims support the specific inference described in the interpretive argument. If we take #6 from the list of inferences above "The proportions of students from my state in each performance level are in those levels because of differences in their preparation related to knowledge, skills, and abilities as described in the ALDs," the claims might include the following.

1. If NAEP test items are designed to differentiate the skills associated with the knowledge, skills, and abilities described in the ALDs, then NAEP scores may relate directly to the ALDs.
2. If NAEP content is sufficiently similar to the content educators teach in schools, then NAEP scores may reflect students' preparation in schools.
3. If student preparation in schools improves, then NAEP scores should also improve.

There are other claims that might be needed to support this inference, but these provide an example of the structure of the validity argument. The claims are then arranged in a structure or graphic that indicates their interconnected nature and dependencies. Failure to support one claim may undermine all subsequent claims that depend on it. For example, the frameworks define the NAEP assessment content. If that content were substantively different from the content taught in schools within a state, NAEP's validity for determining if the students were improving from year to year would be compromised. The students might be improving greatly on content extraneous to NAEP. All inferences related to subgroup performance or subgroup gains would also be undermined. Comparisons to other states, with content similar to that tested on NAEP, would also be undermined.

For the final step, one would simply summarize the evidence supporting each of the claims and determine if the claim is supported, not supported, or if there is insufficient evidence to draw a conclusion. For many claims, previously collected evidence can simply be referenced. For other claims, new investigations may be needed or updates to existing research may be required to account for changes in the American education system, contextual variables that threaten validity, or other factors.

The validity argument might be structured in any number of ways, but a simple approach is to generate tables that include claims, assumptions, evidence, and support. Table 2 provides one example of how a portion of a NAEP validity argument related to the achievement levels might look. The claims are abbreviated from the list of "if…then" statements above and are leftmost in the table. The next column contains the assumptions that underlie this claim. The third column lists evidence that might be used to support the assumptions. The final column is for a summary judgement regarding whether the evidence is supportive (S), non-supportive or counter to the assumption (N), or inconclusive (I). Mock values for this final column are provided in Table 2 to illustrate one way that the validity argument might be constructed. These values do not represent an evaluation of the evidence available.

Table 2. Test Design Claims, Assumptions, and Evidence

Claim	Assumptions	Evidence	Summary Judgement
1. Items Differentiate NAEP Achievement Levels	Items were written to reflect NAEP achievement levels.	Item writing guidelines, instructions, and documentation reflect achievement levels.	S
		Item coding in metadata is linked to achievement levels	S
		Each of the achievement levels is well represented in the item pool for all content categories.	I
		ALD classification accuracy is acceptably high.	I
	Item and test statistics support classification of students.	Metadata supports classification (e.g., the most difficult items reflect the descriptions in the higher achievement levels).	N
		Documentation from standards-setting activities indicate appropriate processes were followed.	S
2. NAEP tests the content taught in schools	Content from NAEP Frameworks largely coincide with state academic standards.	Alignment studies indicate substantial correspondence of content.	N
	The depth described in the NAEP ALDs is similar to the depth described in state performance level descriptors.	Alignment studies show similar ranges of depth of knowledge (DOK) for NAEP ALDs and state performance level descriptors.	N
	Schools teach the main categories of content described by the Frameworks	Review of course syllabi shows correspondence to NAEP Frameworks.	I
3. Improvements in student preparation are reflected on NAEP	NAEP results are sensitive to major changes in educational practice.	Analysis of trend data tracks the timing of major state reform efforts.	S
	NAEP gains/losses are reflected in similar measures of student performance.	Comparisons of gains scores on NAEP are consistent with gains on statewide assessments, ACT, SAT, etc.	I

Contextual Factors that Represent Challenges for Constructing a Validity Argument for NAEP Achievement Levels

One of the most challenging aspects of validation for NAEP ALDs is the context in which NAEP scores are interpreted. The ALDs differentiate students into "Proficient" versus "not-Proficient" categories, and those labels are common with federal requirements for state assessments. It is common for the media to compare state results to NAEP results. When states declare a larger proportion of students to be proficient than NAEP, that finding is often taken as evidence that the state's standards are less rigorous. When NAEP reports that a substantive proportion of students score lower than proficient, those results can be characterized as indicative that students are not on grade level, or that they are unprepared for the next stage in their educational experiences.

These inferences are not supported by NAEP's official documentation, but they are so common that it might be beneficial to consider them when constructing a validity argument. It may be beneficial to characterize the NAEP achievement levels in the context of other common metrics or common understanding of terms. For example, there are multiple indicators of readiness for college (e.g., ACT and SAT benchmarks, specific high school course grades, placement tests, etc.). Many of these indicators have been validated based on outcome criterion (e.g., college course grades, advancement from year 1 to year 2 in college, or attainment of a degree). Providing context related to the NAEP achievement levels that reference similar information may help with interpretation. NAEP is not designed as a college entrance exam, nor as a specific indicator of college readiness. However, indicating that students who score in a particular category tend to also meet other indicators of college readiness could help stakeholders make more sense of their scores.

Another key way that the achievement levels are used by educators is as a guide for what content students are expected to learn and to what degree they are expected to learn that content. The frameworks and the achievement levels provide guidance on expectations for educators, especially in subjects other than mathematics and reading/English language arts, where there may not be clear state standards documents. The frameworks may be used less for mathematics and reading because all states were required to adopt standards for those subjects by federal mandate under the No Child Left Behind Act. Later, most states adopted the Common Core State Standards[5] (CCSS), either in their entirety or with minor editing. These CCSS now serve to guide much of the content taught in American schools. States typically individually worked to characterize performance in relation to the CCSS, so despite common content standards, performance standards vary substantially by state. The NAEP Frameworks and achievement levels are secondary indicators of what students should know and be able to do. If there are important differences between the two standards documents, it could undermine the validity of NAEP scores. If performance is categorized differently by the state for the CCSS than for NAEP, it becomes a challenge for educators to reconcile the differences. Depending on how the states define "Proficient" in reference to the CCSS, educators may not be striving toward "Proficient" as defined by the NAEP achievement levels even if the content of the state assessment and NAEP are largely the same.

There are other contextual factors that should be considered related to the NAEP achievement levels. These factors represent a challenge when drawing inferences from NAEP results and may foster misunderstandings and misuses of data. Their impact can be attenuated by clear guidance regarding the inferences that are supported and those that are not.

[5] See http://www.corestandards.org/.

Using NAEP Achievement Levels to Inform Statewide Testing Standards

One way that NAEP achievement levels have been used by state policymakers is to inform cut scores during standards setting for their statewide achievement tests. States are required to test students in reading/English language arts (ELA) and mathematics in grades 3-8 and high school under the Every Student Succeeds Act (ESSA). Many states also have statewide tests for science and social studies in selected grades. States are required to report results in terms of the proportion of students scoring at the "Proficient" level or above. The level of reporting and the use of the common performance category "Proficient" leads many stakeholders to make comparisons between statewide testing results and NAEP results. States may be criticized if a much greater proportion of students are classified as proficient in grade 4 mathematics on the statewide test than are classified as proficient on NAEP. One of the ways that some states avoid this criticism is to include NAEP achievement levels as part of their standards setting procedures.

While there are several ways that states might include NAEP results in their standards setting, we will consider two here. The first is to use NAEP results as impact data. This use of NAEP may or may not impact cut scores set for state assessments. NAEP results are often used as part of a set of impact data—so the proportions of students in each achievement level on NAEP are considered in conjunction with other information (e.g., the proportion meeting college benchmarks, the proportion in each of the state's reporting categories for a prior assessment, etc.) prior to assigning final cut scores. This typically occurs after standards setting panelists have completed at least one round of assigning cut scores. Impact data is used as a "reality check" to determine if the state cut scores will create controversy in light of other information.

Using NAEP achievement levels to generate impact data requires little in the way of validity evidence, as long as the standards setting facilitators make clear that no direct relationship is expected between NAEP and state assessment results. If, however, the facilitators do not make clear that NAEP achievement levels do not imply grade level performance, college readiness, or other inferences, this impact data can have a much more significant impact on the state's cut scores. If such inferences were intended, a great deal of validity evidence would be needed to support them. Some standards setters guard against making sweeping changes during later rounds of the process, when impact data a reviewed, by placing limits on how far the cut scores can be moved at each stage. This prevents panelists from basing their cut scores on impact data to the exclusion of the performance level descriptors and/or test items.

On the other end of the spectrum, states could create cut scores for their assessments that mirror NAEP achievement levels. This could be accomplished through an equipercentile process without using panelists. It is more likely that the equipercentile solution is presented to panelists as a starting point for standards setting. Then, based on the state's performance level descriptors and/or items, panelists might move the cut scores in one direction or the other to better align with the state's overall assessment system. Limits might be placed on how far the cut scores could deviate to ensure that the proportions of students in each classification category were similar to NAEP. This process would assure that state assessments had similar rigor to NAEP and would allow for more coherent comparisons between the state system and NAEP.

The validity evidence needed to support using NAEP achievement levels in this way would be much more stringent. First, the state would need to ensure that the content of the two tests were sufficiently similar to support consistent cut scores. This would likely require an alignment study. Then, the state would need to establish that the performance level descriptors for the statewide

assessment and for NAEP captured much the same kinds of performance and referenced similar differentiators for each performance category. If not, students might exhibit qualitatively different skills on the assessments, despite scoring similarly.

Evaluating NAEP's Achievement Levels for an Evolving Educational Landscape

NAEP tests students in specified grades in several subjects. Reading and mathematics are tested every other year, while other subjects are tested less often. NAEP's achievement levels for math and reading were established in the early 1990s, while achievement levels for some of the other subjects (e.g., writing, science) have been set or revised more recently. It is important to consider the claims and assumptions that led to the creation of NAEP achievement levels and to verify that those claims and assumptions continue to be relevant and supported as education in America evolves. It is important to verify that NAEP continues to measure the most important content for the tested subjects, that those subjects are the most relevant for stakeholders, and that the knowledge, skills, and abilities described in the achievement levels still represent the most important differentiators for student achievement. A strong validity argument is not static, but routinely tests its claims and assumptions as the inferences stakeholders draw from test information change.

Summary: Steps Toward Developing a Validity Framework for NAEP Achievement Levels

The most important step toward validation of the NAEP achievement levels is to explicitly state the inferences that are expected to be made. These inferences will guide the creation of the specific validity claims, which in turn will help the Governing Board organize and present evidence to support the use of the Achievement Levels for their designated purposes. This priority is in line with the Governing Board's Strategic Vision and is explicit in its response to the achievement levels evaluation (National Assessment Governing Board, 2017).

Once the inferences are made explicit, the next step in the validation process is to investigate the utility of the Achievement Levels for their intended purposes. We know that one of those purposes is to help define what students know and can do within the tested subjects. The ALDs describe student performance within specific ranges on the scale. Users of NAEP data are provided with the proportions of students expected to be at each performance level, which they interpret in conjunction with the ALDs. It would be beneficial to sample from these interpretations to ascertain if the information provided is meeting the needs of key stakeholders, and to determine if those stakeholders are making unsupported interpretations from the data.

This process will provide key input into the next step in establishing a validity framework, the creation of an interpretive guide for NAEP achievement levels. Such a guide would indicate the key inferences stakeholders are expected to make, caveats and limitations on those interpretations, and warnings about common potential misinterpretations or misuses of the NAEP Achievement Levels or achievement level data. The interpretive guide should not be limited to achievement levels, but also include information on the use of scale scores, comparisons across jurisdictions (e.g. states or large districts), and it should describe when it is most appropriate to use achievement levels versus scale scores.

Once the interpretive guide is complete, it can be used to guide the remainder of the validity argument. For example, if the interpretive guide characterizes the content in the Achievement Level for fourth grade Science at Basic as the content that the typical student scoring at that level has mastered, validity evidence would be needed to support that statement. The content described for the Basic level of fourth grade science might be compared with the content of the

NAEP test items that best discriminate within the Basic range of the scale. If the item content essentially matched the content described in the ALD, that finding would represent support for the interpretation. There is, of course, other evidence that might also be used to support such an interpretation. The inference would be considered valid if the preponderance of this evidence was supportive and no evidence directly contradicted the inference.

This process would be repeated for each of the inferences described in the interpretive guide until all the inferences were addressed to the satisfaction of assessment validity experts, several of whom serve on the Governing Board. For many of the intended inferences, it will be possible to simply reference research that has already been completed. For other inferences, it may be necessary to conduct additional research in order to bring appropriate evidence to bear. If any of the inferences is unsupported by evidence or if the evidence that is available is negative, either the interpretation must be altered or the test information bolstered in some way. The evidence included in the validity argument may need to be revised or updated any time the NAEP assessments are revised or altered, any time there is a significant shift in the national educational landscape, and when there are concerns that the evidence is so dated that it may no longer be applicable.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-597). Washington, DC. Council on Education.

Angoff, W. H. (1988). Proposals for theoretical and applied development in measurement. *Applied Measurement in Education*, I, 215-222.

Bacon-Blood, L. (2013, November 7). Louisiana students score near bottom on national tests. Retrieved from http://www.nola.com/education/index.ssf/2013/11/louisiana_students_score_near.html

Bandeira de Mello, V., Bohrnstedt, G., Blankenship, C., & Sherman, D. (2015). *Mapping State Proficiency Standards onto NAEP Scales: Results from the 2013 NAEP Reading and Mathematics Assessments* (NCES 2015-046). National Center for Education Statistics. Washington, DC: U.S. Department of Education.

Bourque, M. L. (2009). A History of NAEP Achievement Levels: Issues, Implementation, and Impact 1989-2009. National Assessment Governing Board.

Chingos, M., & Blagg, K. (2015, October 28). How do states really stack up on the 2015 NAEP? Retrieved from https://www.urban.org/urban-wire/how-do-states-really-stack-2015-naep

Dillon, S. (2005, November 26). Students ace state tests, but earn D's from U.S. Retrieved from http://www.nytimes.com/2005/11/26/education/students-ace-state-tests-but-earn-ds-from-us.html#story-continues-1

Egan, K.L., Schneider, M.C., & Ferrara, S. (2012). Performance level descriptors: History, practice and a proposed framework. In G.J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (pp. 79-106). New York: Routledge.

Fields, R. (2014). Towards the National Assessment of Educational Progress (NAEP) as an Indicator of Academic Preparedness for College and Job Training. Washington, DC: National Assessment Governing Board. Retrieved from http://ed.sc.gov/scdoe/assets/File/tests/middle/naep/NAGB-indicator-of-preparedness-report.pdf

Gattis, K., Kim, Y., Stephens, M., Dager, L., Fei, H., & Holmes, J. (2016). A Comparison Study of the Program for International Student Assessment (PISA) 2012 and the National Assessment of Educational Progress (NAEP) 2013 Mathematics Assessments. AIR-NAEP Working paper #02-2016.

Hambleton, R., Pitoniak, M., & Copella, J. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G.J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (pp. 47-76). New York: Routledge.

Ho, A. D., & Haertel, E. H. (2007). Apples to apples? The underlying assumptions of state-NAEP comparisons. *Paper commissioned by the Council of Chief State School Officers*. Retrieved from https://scholar.harvard.edu/files/andrewho/files/ho_haertel_apples_to_apples.pdf

Hull, J. (2008, June 17). The proficiency debate. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/The-proficiency-debate-At-a-glance/The-proficiency-debate-A-guide-to-NAEP-achievement-levels.html

Jia, Y., Phillips, G., Wise, L.L., Rahman, T., Xu, X., Wiley, C., & Diaz, T.E. (2014). 2011 NAEP-TIMSS Linking Study: Technical Report on the Linking Methodologies and Their Evaluations (NCES 2014-461). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Kane, M. (2001). So much remains the same: conception and status of validation in setting standards. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp.53-88). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. (2006). *Validation*. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, v50(1), pp. 1-73.

Levell, M. (2016, November 4). Public education test results are dismal. Schools are failing NH children. Retrieved from http://nhpoliticalbuzz.org/public-education-test-results-are-dismal-schools-are-failing-nh-children/

Lim, H., & Sireci, S. G. (2017). Linking TIMSS and NAEP assessments to evaluate international trends in achievement. Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, (25), 1-25.

Loomis, S. (2012). Selecting and training standard setting participants. In G. J. Cizek (Ed.), Setting performance standards: Concepts, methods, and perspectives (pp.107-134). Mahwah, NJ: Lawrence Erlbaum.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. ETS Research Report Series, 2003(1).

Moran, R., Freund, D., & Oranje, A. (2012). Analyses relating Florida students' performance on NAEP to preparedness indicators and postsecondary performance. Washington, DC: Author. Retrieved from https://www.nagb.org/content/nagb/assets/documents/what-wedo/preparedness-research/statistical-relationships/Florida_Statistical_Study.pdf

Moran, R., Oranje, A., & Freund, D. (n.d.). NAEP 12th grade preparedness research: Establishing a statistical relationship between NAEP and SAT. Retrieved from http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparednessresearch/statistical-relationships/SAT-NAEP_Linking_Study.pdf

National Academies of Sciences, 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.

National Assessment Governing Board. (2017). National Assessment Governing Board's Response to the National Academies of Sciences, Engineering, and Medicine 2016 Evaluation of NAEP Achievement Levels. Provided by the National Assessment Governing Board November 2017. Author.

National Assessment of Educational Progress Frequently Asked Questions. (2017, October 11). Retrieved from http://www.doe.mass.edu/mcas/natl-intl/naep/faq.html?section=overview

Neidorf, T., Binkley, M., Gattis, K., & Nohara, D. (2006). Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments. NCES 2006-029.

Norcini, J.J., & Shea, J.A. (1992). The reproducibility of standards over groups and occasions. Applied Measurement in Education, 5, 63-72.

Peterson, P., & Ackerman, M. (2015). States raise proficiency standards in math and reading. Retrieved from file:///C:/NAEP/ednext_XV_3_peterson%20ackerman%202015.pdf

Phillips, G. W. (2007). Expressing International Educational Achievement in Terms of US Performance Standards: Linking NAEP Achievement Levels to TIMSS. American Institutes for Research.

Phillips, G. W. (2014a). International Benchmarking: State and National Education Performance Standards. American Institutes for Research.

Phillips, G. W. (2014b). Linking the 2011 National Assessment of Educational Progress (NAEP) in Reading to the 2011 Progress in International Reading Literacy Study (PIRLS). American Institutes for Research.

Poland, S., & Plevyak, L. (2015). US student Performance in science: A review of the four major science assessments. *Problems of Education in the 21st Century*, *64*.

Polikoff, M. (2015a, October 6). Friends don't let friends misuse NAEP data. [Blog post]. Retrieved from https://morganpolikoff.com/tag/naep/

Polikoff, M. (2015b, October 28). My quick thoughts on NAEP. [Blog post]. Retrieved from https://morganpolikoff.com/tag/naep/

Provasnik, S., Lin, C., Darling, D., & Dodson, J. (2013). A comparison of the 2011 Trends in International Mathematics and Science Study (TIMSS) assessment items and the 2011 National Assessment of Educational Progress (NAEP) frameworks. *National Center for Education Statistics*.

Sawchuk, S. (2013, July 24). When bad things happen to good NAEP data. Retrieved from https://www.edweek.org/ew/articles/2013/07/24/37naep.h32.html

Schneider, M. C., Huff, K. L., Egan, K. L., Tully, M., & Ferrara, S. (2010). Aligning achievement level descriptors to mapped item demands to enhance valid interpretations of scale scores and inform item development. In *annual meeting of the American Educational Research Association, Denver, CO*.

Schneider, M., Kitmitto, S., Muhusani, H., & Zhu, B. (2015). Using the National Assessment of Educational Progress as an Indicator for College and Career Preparedness. Washington, DC: Author. Retrieved from http://www.air.org/sites/default/files/downloads/report/Using-NAEP-as-an-Indicator-College-Career-Preparedness-Oct-2015.pdf

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). Setting performance standards for student achievement: A report of the National Academy of Education Panel on the Evaluation of the 1992 Achievement Levels. Stanford, CA: National Academy of Education.

Vermont students score among best in the nation on the National Assessment of Educational Progress. (2016, November 3). Retrieved from http://education.vermont.gov/sites/aoe/files/documents/edu-press-release-naep-necap-scores.pdf

Vinovskis, M. A. (1998). Overseeing the Nation's Report Card: The Creation and Evolution of the National Assessment Governing Board (NAGB).

Weiss, J. (2016, January 27). Report says Texas school standards are worst in nation. Retrieved from https://www.dallasnews.com/news/education/2016/01/27/report-says-texas-school-standards-are-worst-in-nation

Zenisky, A., Hambleton, R.K., & Sireci, S.G. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education, 22*(4), 359-375.

Communication and Interpretation of Achievement Levels

At the November 2018 Governing Board meeting, COSDAM and the Reporting and Dissemination Committee will hold a joint meeting to discuss the two committees' work on achievement levels for the National Assessment of Educational Progress (NAEP).

Background

From 2014 to 2016, the National Academies of Sciences, Engineering, and Medicine evaluated the NAEP achievement levels in mathematics and reading, which are the responsibility of the Governing Board. In their evaluation, the National Academies noted eight common uses of NAEP achievement levels, specifically:

- Trends or comparisons of successive cohorts, e.g., the percentage of students at or above Proficient in reading has increased over time;
- Comparison to a state assessment;
- Point-in-time comparisons across states, districts, or population groups, e.g., more students in state A who are at or above Proficient in reading compared to state B;
- Rank ordering states or districts;
- Comparison across population groups to examine performance gaps;
- Comparison across subject areas, e.g., more students perform at or above Proficient on mathematics than in reading;
- Comparison of before and after an action or policy implementation; and
- Relationships among achievement results and contextual data.

The evaluation recognized the usefulness and value of the achievement levels but made several important recommendations, most of which focus on the work of COSDAM as well as two that also address the work of the R&D Committee:

> 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.

Since the release of these recommendations in November 2016, Governing Board staff and COSDAM members have started working to fulfill these recommendations. The draft revision of

the Board policy on developing student achievement levels (planned for full Board action in November 2018) establishes an

> "interpretative guide [which] shall accompany NAEP reports, including specific examples of appropriate and inappropriate interpretations and uses of the results" (Principle 3h).

This guide is intended for inclusion on the Nation's Report Card website and on specific report card webpages. The guide will target stakeholders, such as media, policy advocates, members of the general public, educators, and policymakers. These groups may be familiar with both NAEP and achievement levels, but their understanding, interpretation, and use of achievement levels could be informed and improved with guidance from the Governing Board.

The Reporting and Dissemination Committee will collaborate with COSDAM on the development of this interpretative guide. The overarching question of the joint meeting will focus on the general approach the interpretative guide should take. This joint meeting also will elicit feedback on several specific features of the guide:

(1) the scope—what should be covered and what should not;
(2) the content—uses of achievement levels, value and usefulness of achievement levels;
(3) the language—non-technical, accessible; and
(4) the delivery—how the guide will be included with report cards.

If there is time, the conversation may extend to initial discussions of a statement on both the uses and usefulness of NAEP generally, not only of achievement levels specifically.

In addition, the committees should deliberate together on how to engage stakeholders on improving their use and interpretation of NAEP and achievement levels beyond the interpretative guide.

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 (e.g., framework update processes) or ad hoc committee (i.e., exploring new approaches to postsecondary preparedness) also have not been included below.

The Governing Board's Assistant Director for Psychometrics, Sharyn Rosenberg, will answer any questions that COSDAM members have about ongoing or planned activities.

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 *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*	Ongoing linking studies include: national NAEP-ACT linking study; longitudinal studies at grade 12 in MA, MI, TN; longitudinal studies at grade 8 in NC, TN; NAEP-TIMSS linking study; NAEP-HSLS linking study; NAEP Validity Studies (NVS) studies Informational update on current studies was provided in the March 2018 COSDAM materials As of October 2018, analyses are currently underway for the national NAEP-ACT linking study, with presentation to COSDAM tentatively planned for March 2019	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
SV #3: Expand the availability, utility, and use of NAEP resources, in part by creating new resources to inform education policy and practice *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)* *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)*	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 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 A joint discussion of COSDAM and the Reporting & Dissemination Committee was planned for November 2018 but has been postponed to March 2019 to allow time for focused discussion on achievement levels instead	Use research to draft short document of intended and appropriate uses for COSDAM discussion (March 2019) NCES produces documentation of validity evidence for intended uses of NAEP scale scores Governing Board produces documentation of validity evidence for intended uses of NAEP achievement levels	Board adopts formal statement or policy about intended uses of NAEP. The goal is to increase appropriate uses and decrease inappropriate uses (in conjunction with dissemination activities to promote awareness of the policy statement)
Disseminate information on technical best practices and NAEP methodologies, such as training item writers and setting achievement levels	This idea was generated during the August 2017 COSDAM discussion of the Strategic Vision activities	Work with NCES and R&D to refine list of technical topics for dissemination efforts	Stakeholders benefit from NAEP technical expertise

Strategic Vision Activity	Current Status and Recent Work	Planned Next Steps	Desired Outcome
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 *Consider new approaches to creating and updating the achievement level descriptors and update the Board policy on achievement levels*	Input for the policy revision was provided through a panel of standard setting experts, a literature review on considerations for creating and updating achievement level descriptors (ALDs), and a technical memo on developing a validity argument for the NAEP achievement levels (early 2018) COSDAM discussed the policy revision during the May and March 2018 Board meetings Full Board discussed the draft revised policy during the August 2018 Board meeting Public comment was sought from August 30 – October 15, 2018; Board calls to discuss the comments took place in October Additional discussion of the draft revised policy will take place during the upcoming November 2018 Board meeting	Board action on revised policy statement (planned for November 2018)	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

Strategic Vision Activity	Current Status and Recent Work	Planned Next Steps	Desired Outcome
SV# 7: Research policy and technical implications related to the future of NAEP Long-Term Trend assessments in reading and mathematics *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*	White papers commissioned, symposium held in Washington, DC (March 2017), and follow-up event held at American Educational Research Association (AERA) annual conference (April 2017) Full Board and Executive Committee discussions (March, May, and August 2017) and webinar on secure LTT items and p-values from 2012 administration (October 2017) The NAEP budget in Fiscal Year 2019 has been increased by $2 million with a goal of moving up the next administration of LTT (Discussion in November 2018 Executive Committee meeting)	Per the discussion and next steps at the March 2018 Executive Committee meeting, COSDAM will discuss design considerations for the next administration of LTT. Pending the outcome of discussions in the Executive Committee meeting in November, additional information about design considerations will be shared with COSDAM at the March 2019 meeting.	Determine whether changes to the NAEP LTT schedule, design and administration are needed (led by Executive Committee and NCES)
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 *Pending outcomes of stakeholder input (ADC activity), evaluate the technical implications of combining assessments, including the impact on scaling and trends*	COSDAM presentation and discussion on initial considerations for combining assessments During the past year, there have been several full Board presentations and discussions on the assessment schedule Initial draft schedule and budget to be discussed in November 2018	Board action on the NAEP Assessment Schedule tentatively scheduled for March 2019	Determine whether new assessment schedule should include any consolidated frameworks or coordinated administrations

Strategic Vision Activity	Current Status and Recent Work	Planned Next Steps	Desired Outcome
SV# 10: Develop new approaches to measure the complex skills required for transition to postsecondary education and career *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*	Several studies are ongoing (see activities under SV# 2) Per COSDAM discussion at August 2017 meeting, additional studies are on hold until at least November 2018 pending Board decision on how to move forward with findings from Ad hoc Committee on Measures of Postsecondary Preparedness	Decide whether additional research should be pursued at grade 8 to learn more about the percentage of students "on track" to being academically prepared for college by the end of high school or whether additional research should be conducted with more recent administrations of NAEP and other tests Decide whether Board should make stronger statement and/or set "benchmarks" rather than using "plausible estimates"	Statements about using NAEP as an indicator of academic preparedness for college continue to be defensible and to have appropriate validity evidence