National Assessment Governing Board
Committee on Standards, Design and Methodology
Friday, March 2, 2018
10:00 am – 12:30 pm
AGENDA

10:00 – 10:05 am	Welcome and Review of Agenda *Andrew Ho, COSDAM Chair*	
10:05 – 10:45 am	Best Practices in Achievement Levels Setting (SV #5)	Attachment A
	• Summary of Expert Panel Meeting *Thanos Patelis, HumRRO*	Attachment B
10:45 am – 11:25 am	• Literature Review on Achievement Level Descriptions (ALDs) *Karla Egan, EdMetric*	Attachment C
	• Technical Memo on Considerations for a Validity Framework for the NAEP Achievement Levels *Art Thacker, HumRRO*	Attachment D
11:25 am – 12:25 pm	Identifying Revision Goals for the Board Policy on Achievement Levels Setting (SV #5) *Andrew Ho* *Sharyn Rosenberg, Assistant Director for Psychometrics*	
12:25 – 12:30 pm	Questions on Information Items	
	Writing Grade 4 Achievement Levels Setting Update	Attachment E
	Update on Implementing the Strategic Vision (SV#2-10)	Attachment F
	Summary of Ongoing NAEP Linking Studies (SV #2)	Attachment G

Best Practices in Achievement Levels Setting (SV #5)

Background

Over the past year, COSDAM members discussed the need to revise the 1995 Governing Board policy on Developing Student Performance Levels for NAEP (attached). The Board's formal response to the November 2016 evaluation of the NAEP achievement levels (attached) noted that several of the report recommendations would be addressed through a revision of the Board policy. In particular, the Board's response stated that the updated policy will specify a process and timeline for conducting regular recurring reviews of the achievement level descriptions (ALDs) and will be explicit about the conditions that necessitate consideration of a new standard setting. In addition, one of the planned activities for the implementation of the Strategic Vision is to consider new approaches to creating and updating the achievement level descriptions in the revision of the Board policy on achievement levels.

Given that the policy is over 20 years old, there is also a need to revisit the policy more generally to ensure that it reflects current best practices in standard setting. COSDAM members have acknowledged the need to seek input from multiple stakeholders throughout the process of revising the policy. To get an initial sense of the potential scope of recommended revisions to the policy, Assistant Director for Psychometrics Sharyn Rosenberg conducted informal conversations with several standard setting experts in spring 2017. Feedback from those conversations (excerpted below from the May 2017 COSDAM materials) was shared with COSDAM in May 2017 and informed the additional work that has been performed since then.

Key takeaways from expert conversations conducted during March/April 2017

- All references to publications and some references to organizations need to be updated
- Achievement-levels setting processes should be elaborated, and procedures institutionalized over time should be made explicit in the policy (e.g., use of split panels, use of feedback and impact data, roles and qualifications of content/process facilitators)
- Some word choices are not quite accurate or appropriate (e.g., "judges" is no longer a common term and should be replaced by "panelists")
- The response probability (RP) criterion of 0.50 for identifying exemplar items is not ideal and does not match the criteria used in reporting of NAEP item maps
- The following aspects of the achievement level descriptions (ALDs) should be revisited:
    - What is meant by "preliminary ALDs"
    - How and when the preliminary ALDs are finalized in the standard setting process
    - The extent to which the preliminary ALDs do and should inform item development
    - Whether the ALDs refer to the full range of the level or performance at the threshold/borderline
    - Whether shorter, more concise versions of the ALDs should be developed for reporting

- Public comment should be limited to the design/methodology (or perhaps only specific novel elements) and should not refer to the results, which are embargoed prior to release
- Description of methodology should not be limited to the Angoff method
- Composition, qualifications, and size of the panels should be revisited (e.g., definition of general public panelists, how the panel size relates to the standard errors of cut scores)
- There should be explicit guidance for when and how to revisit the achievement level descriptions and cut scores, but this should be balanced by acknowledging the value of stability in the standards since they acquire meaning over time
- Procedures for conducting the standard setting process and quality control processes should be updated to reflect the shift to digital-based assessments
- Consider including information about primary ways the achievement levels should or should not be used
- Validation should be characterized as an ongoing process, and the approach, timing, and types of evidence collected should be reconsidered
- Achievement levels should not be the "initial and primary means" of reporting NAEP

March 2018 COSDAM discussion

As part of the Technical Support contract, the Governing Board requested that the Human Resources Research Organization (HumRRO) undertake several activities to inform the revision of the Board policy on setting achievement levels for NAEP. The following efforts will be discussed at the upcoming COSDAM meeting:

Expert panel meeting on NAEP achievement levels setting

A two-day panel meeting of standard setting experts took place on January 10-11, 2018 to discuss and provide input to the Governing Board regarding best practices for setting and maintaining achievement levels. The panel was composed of the following members: Dr. Michael Bunch (Measurement Inc), Dr. Karla Egan (EdMetric, LLC), Dr. Steve Ferrara (Measured Progress), Dr. Ed Haertel (Stanford University), Dr. Ron Hambleton (University of Massachusetts-Amherst), Dr. Laura Hamilton (RAND), Dr. Marianne Perie (University of Kansas), and Dr. Barbara Plake (University of Nebraska-Lincoln). The meeting minutes are provided in Attachment B. During the March 2018 COSDAM meeting, Dr. Thanos Patelis of HumRRO will present a few key takeaways from the expert panel meeting and will answer any questions that Board members may have. One of the expert panelists, Dr. Karla Egan, will also be in attendance at the COSDAM meeting.

Literature review on achievement level descriptions (ALDs)

Several of the issues raised in the evaluation of NAEP achievement levels and the preliminary conversations with standard setting experts in spring 2017 were related to the development, use, and review of achievement level descriptions. In addition, one of the Strategic Vision activities is to consider new approaches to creating and updating the ALDs as part of the effort to revise the Board policy on achievement levels setting. Under subcontract to HumRRO, Drs. Karla Egan and Anne Davidson of EdMetric performed a literature review (Attachment C) on best practices related to ALDs, including the question of whether and how multiple versions of ALDs should be used. Some of the issues discussed in the literature review, such as the use of ALDs in the item development process, have implications beyond the Board policy on achievement levels setting; exploration of this idea in particular would require follow-up conversations with the National Center for Education Statistics (NCES) and the Assessment Development Committee (ADC). During the March 2018 COSDAM meeting, Dr. Karla Egan will present a few key takeaways from the literature review and will answer any questions that Board members may have.

Technical memo on building a validity framework for the NAEP achievement levels

Over the past year, the Board has discussed the need to articulate the intended uses of NAEP scale scores and achievement levels and to provide specific validity evidence in support of those uses. There are several Strategic Vision activities related to this goal, consistent with the recommendations from the recent evaluation of NAEP achievement levels. To provide advice on how to approach the construction of a validity argument for the NAEP achievement levels, Drs. Arthur Thacker and Tanya Longabach of HumRRO developed a technical memorandum. During the March 2018 COSDAM meeting, Dr. Thacker will present a few key takeaways from an excerpt of that memorandum (Attachment D) and will answer any questions that Board members may have.

Revision goals for the Board policy on setting achievement levels

Following the presentations and discussions on various aspects of best practices in achievement levels setting, COSDAM members will spend the last hour of the Committee meeting identifying revision goals for the Board policy statement and considering whether any additional information is needed. In advance of the March 2018 Board meeting, COSDAM Chair Andrew Ho will send an email to COSDAM members with proposed discussion questions.

Adopted: March 4, 1995

National Assessment Governing Board

Developing Student Performance Levels for the National Assessment of Educational Progress

Policy Statement

Foreword
A policy on setting achievement levels on the National Assessment of Educational Progress (NAEP) was first adopted in 1990 and amended several times thereafter. The present policy, adopted in 1995, contained introductory and explanatory text, principles, and guidelines. Since 1995, there have been several changes to the NAEP authorizing legislation (currently, the NAEP Authorization Act: P.L. 110-279). In addition, related legislation has been enacted, including the No Child Left Act of 2001. Consequently, introductory and other explanatory text in the original version of this policy, no longer germane, has been deleted or revised to conform to current legislation. The Principles and Guidelines remain in their original form except for Principle 4, from which the reference to the now decommissioned Advisory Council on Education Statistics has been deleted. (Foreword added August 2007.)

Principles for Setting Achievement Levels

Principle 1

The level setting process shall produce for each content area, three threshold points at each grade level assessed, demarcating entry into three categories: *Basic, Proficient, and Advanced.*

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

Basic.	*This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.*
Advanced.	*This level signifies superior performance beyond proficient.*

Principle 2

Developing achievement levels shall be a widely inclusive activity of the Board, utilizing a national consensus approach, and providing for the active participation of teachers, other educators (including curriculum specialists and school administrators at the local and state levels), and non-educators including parents, members of the general public, and specialists in the particular content area.

The development of achievement levels shall be conducted in two phases. In phase 1, the assessment framework development process shall yield preliminary descriptions of the achievement levels *(Basic, Proficient, and Advanced),* which shall subsequently be used in phase 2 to develop the numerical standards (cut scores) and to identify appropriate examples of assessment exercises that typify performance at each level. The levels will be updated as appropriate, typically when the assessment frameworks are updated.

Principle 3

The Governing Board shall incorporate the student performance levels into all significant elements of NAEP, including the subject area framework development process, exercise development and selection, and the methodology of the assessment. The achievement levels shall be used to report the results of the NAEP assessments so long as such levels are reasonable, valid and informative to the public.

Principle 4

In carrying out its statutory mandate, the Governing Board will *exercise its policy judgment in setting the levels.* The Board shall continually seek better means of setting achievement levels. In so doing, the Board may seek technical advice as appropriate from a variety of sources, including external evaluations provided by the Secretary, the Commissioner, and other experts. Proposed achievement levels shall be reviewed by a broad constituency, including consumers of NAEP data, such as policymakers, professional groups, the states and territories. In carrying out its responsibilities, the Board will ordinarily engage the services of a contractor who will prepare recommendations for the Board's consideration on the levels, the descriptions, and the exemplar exercises.

Guidelines for Setting Achievement Levels

Each guideline presented below is accompanied by a rationale and a summary of the implementation practices and procedures to be followed in carrying out the principle. It should be understood that the full implementation of this policy will require the

contractor, through Governing Board staff, to provide assurances to the Board that all aspects of the practices and procedures for which they are responsible have been completed successfully. These assurances will be in writing, and may require supporting documentation prepared by the contractor and/or Governing Board staff.

Summary of Guidelines

Guideline 1

The level setting process shall produce for each content area, three threshold points at each grade level assessed, demarcating entry into three categories: *Basic, Proficient, and Advanced.*

Guideline 2

The level setting process shall be a widely inclusive activity of the Board, carried out by a broadly representative body of teachers, other educators (including curriculum specialists and local and state administrators), and non-educators including parents, concerned members of the general public, and specialists in the particular content area; this process and resulting products shall be reviewed by a broad constituency.

Guideline 3

The level-setting process shall result in achievement level cut scores for each grade and level, expanded descriptions of the content expected at each level based on the preliminary descriptions provided through the national consensus process, and exemplar exercises that are representative of the performance of examinees at each of the levels and of the cognitive expectations for each level described.

Guideline 4

In carrying out its statutory mandate, the Board will *exercise its policy judgment in setting the levels.* However, in so doing, they will seek technical advice from a variety of sources, but especially from the contractor who will prepare the recommendations on the levels, the descriptions, and the exemplar exercises, as well as from consumers of NAEP data, including policymakers, professional groups, the states, and territories.

Guideline 5

The achievement levels shall be the initial and primary means of reporting the results of the National Assessment of Educational Progress at both the national and state levels.

Guideline 6

The level-setting process shall be managed in a technically sound, efficient, cost-effective manner, and shall be completed in a timely fashion.

Guideline 1

The level setting process shall produce for each content area, three threshold points at each grade level assessed, demarcating entry into three categories: *Basic, Proficient, and Advanced.*

Rationale

The Board is committed to describing the full range of performance on the NAEP scale, for students whose performance is in the mid-range, as well as for those whose performance is below and above the middle. It is highly desirable to endorse realistic expectations for all students to achieve no matter what their present performance might be. Three benchmarks on the NAEP scale suggest realistic expectations for students in all regions of the performance distribution. Likewise, the Board is committed to preserving trend results in NAEP. Three achievement levels accommodate growth (and possible declines) in all ranges of the performance distribution.

Practices and Procedures

Policy Definitions

The following policy definitions will be applied to all grades, 4, 8, and 12, and all content areas in which the levels are set. It is the Board's view that the level of performance referred to in the policy definitions is what students *should be able to know and do,* and not simply the current academic achievement of students or that which today's U.S. schools expect.

Proficient.	*This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter.*
Basic.	*This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.*
Advanced.	*This level signifies superior performance beyond proficient.*

From Policy Definitions to Content Descriptions

In the course of applying the policy definitions to the level-setting process, it will be necessary to articulate them in terms of the specific content and sequence (now called descriptions) appropriate for the grades in which the levels are being set. This will be completed on a preliminary basis through the process which develops the assessment

frameworks. These preliminary descriptions will be used to initially guide the work of deriving the advice that will assist the Board in setting the levels. Throughout the process of obtaining such advice, however, these descriptions may be refined, expanded, and edited to more clearly reflect the specific advice on the levels.

Training of Judges

In training the judges for the level-setting activity, it is necessary that all arrive at a common conceptualization of *Basic, Proficient, and Advanced* based on the policy definitions of the Board. Such conceptualizations must be within the scope of the assessment framework under consideration and capable of being applied at the individual item level (Reid, 1991.)

Judges must also be trained in the specific model that will be used to generate the rating data. At the very least, they need to understand the purposes for setting the levels, the significance of such an activity, the NAEP assessment framework for the subject area under discussion, elements that make particular exercises more or less difficult, and the rating task itself.

Judges shall be trained by individuals who are both knowledgeable in the subject matter area and are experienced, capable trainers in a large-group setting. Presentations shall be prepared, rehearsed, and piloted before implementation.

Judges shall be provided comprehensive, user-friendly training materials, adequate time to complete the task, and the appropriate atmosphere in which to work, one that is quiet, pleasant, and conducive to reaching the goals of the level-setting activity. It is also required that judges take the assessment under the same NAEP-like conditions as students, that is, using the NAEP student booklets, having all manipulatives and ancillary materials, and timed.

Guideline 2

The level setting process shall be a widely *inclusive* activity of the Board, carried out by a broadly representative body of teachers, other educators (including curriculum specialists and local and state administrators), and non-educators including parents, concerned members of the general public, employers, scholars, and specialists in the particular content area. This process and resulting products shall be reviewed by a broad constituency.

Rationale

The spirit of the legislative mandate of the Board is one of moving toward a national consensus on policy issues affecting NAEP. The Board has historically involved broad audiences in its deliberations. The achievement levels are no different. Further, the Board views the level-setting activity as an extension of the widely inclusive effort to derive the assessment frameworks and scope and sequence of each assessment. Finally, the magnitude of the decisions regarding *what students should know and be able to do is*

simply too important a decision to seek involvement from professionals alone; it must have the benefit of the collective wisdom of a broadly representative body, educators and non-educators alike.

Practices and Procedures

Sample of Judges

The panel of judges will be composed of both educators and non-educators. About two-thirds of the panel will represent teachers and other educators; one-third will represent the public, non-educator sector, for example, scholars, employers, parents, and professionals in occupations related to the content area. They will be drawn from a national sampling frame and will be broadly representative of various geographic regions (Northeast, Southeast, Central, West, and the territories) types of communities (urban, suburban, rural), ethnicities, and genders.

Individual panel members shall have expertise in the specific content area in which the levels are being developed, expertise in the education of students at the grades under consideration, and a general knowledge of assessment, curriculum, and student performance. The composition of the panels should be such that they meet the requirements of the *Standards (1985).*

The size of the panels should be responsive to what the research demonstrates regarding numbers of judges involved (see Jaeger, 1991). While it may not be practical or beyond the resources available, every effort should be made to empanel a sufficient number of judges to reduce the standard error of the cut score. While there is no absolute criterion on the magnitude of the standard error of the cut score, a useful rule of thumb is that it should not exceed the *combined* error associated with the standard error of measurement on the assessment and the error due to sampling from the population of examinees.

Review Procedures

Throughout the process and particularly at critical junctures, groups that have a legitimate interest in the process will be involved. During the planning process interested groups and individuals will be encouraged to participate and share their experiences in the area of setting standards. These groups might include professional societies, *ad hoc* advisory groups, standing advisory committees to the Governing Board or its contractor(s) and NCES and its contractor(s) and grantees. Documents (such as the Design Document and Interim Reports) will be disseminated in sufficient time to allow for a thoughtful response from those who wish to provide one.

Proposed levels will be widely distributed to major professional organizations, state and local assessment and curriculum personnel, business leaders, government officials, the Planning and Steering Committees of the framework development process, the Exercise Development panels, and other groups who may request them.

When it is deemed useful by the Board, public hearings and forums will be conducted in Washington, D.C. and other parts of the country to encourage review and input on a broad regional and geographic basis.

Guideline 3

The resulting products of the level-setting process shall be (1) achievement level scores marking the threshold score for each grade and level, (2) expanded descriptions of the content expected at each level based on the preliminary descriptions provided through the national consensus process, and (3) exemplar exercises that are representative of the performance of examinees at each of the levels and of the cognitive expectations for each level described. These three products form the basis for reporting the results of all future NAEP assessments.

Rationale

The NAEP scale, while useful for aggregating large amounts of information about student performance in a single number, requires contextual information about the specific content and the sequencing of that content across particular grades, in order to be truly beneficial to users of NAEP data. In order to make the NAEP data more useful, descriptions of each level which articulate content expectations and exemplar exercises taken from the public release pool of the most current NAEP assessment must accompany the benchmarks or cut scores for each level. The descriptions and exemplars are intended to be illustrative of the kind of content that is represented in the levels, as well as an aid in the interpretation of the NAEP data.

Practices and Procedures

Methodology

The methodology to be used in generating the levels will depend upon the specific assessment formats for the content area in which the levels are being set. Historically, in the case of multiple choice exercises and short constructed response formats, a modified Angoff (1971) procedure has been employed. In the case of extended constructed response formats, a paper-selection procedure has been employed. Neither of these is without its disadvantages. As the assessment formats of future assessments become more complex and employ more performance-type exercises, it is quite likely that alternate procedures will be needed. The Board will decide these on a case-by-case basis, looking for advice from those who have had experience in dealing with these alternative assessment formats. In any case, the design for carrying out the process must be carefully crafted, must be appropriate to the content area and philosophy of the assessment framework, and must have a solid research base.

The procedures will generally be piloted prior to full implementation. The purpose of the pilot would be to test out the materials used with the judges, the training procedures, the feedback information given to the judges during the process, and the

software used to complete the initial analyses. Procedures would be revised based on the pilot experience and evaluation evidence.

Whatever methodology is used, all aspects of the procedures will be documented for the purposes of providing evidence of procedural validity for the levels being recommended. This evidence will be made available to the Board at the time of deliberations about the levels being set.

Quality Control Procedures

While there are numerous points in a complex process for mistakes to occur, there are at least three important junctures where quality control measures need to be in place. First, is the point of data entry. Ideally, judges' ratings should be scanned to reduce manual errors of entry. However, if the ratings are entered manually, then they shall be entered and 100% verified using a double-entry, cross-checking procedure. Second, software programs designed to complete initial analyses on the rating data must be run with simulated data to de-bug, and provide assurances of quality control. The programs should detect logical errors and other kinds of problems that could result in incorrect results being generated. Finally, the production of cut scores on the NAEP scale is the final responsibility of the NAEP operations contractor. Only final cut scores, mapped onto the properly weighted and equated scale, received in writing from the operations contractor, will be officially communicated to the Board, or others who have a legitimate need to know. *Once the accuracy of the data has been ensured by the level-setting and operations contractors, the Board shall make a policy determination and set the final achievement levels, informed by the technical process of the level-setting activity.*

Descriptions of the Levels

The preliminary descriptions developed through the framework development process will be the starting point for developing recommendations for the levels under consideration. The preliminary descriptions are *working descriptions* for the panels while doing the ratings. These may be expanded and revised accordingly as these panels conduct the ratings, examine empirical performance data, and work to develop their final recommendations on the levels. The recommended descriptions will be articulated in terms of what students *should know and should be able to do.* They shall be coherent within grade, and consistent across grades, and will reference performance within the three regions created by the cut scores. No descriptions will be done for content below the *Basic* level.

Exemplar Exercises

The exemplars chosen from the released pool of exercises for the current NAEP assessment will reflect as much as possible performance both in the *Basic, Proficient, and Advanced* regions of the scale, as we1l as at the threshold scores. Exemplars will be selected to meet the rp = *.50* criterion, and will demonstrate the range of performance possible within the regions. They will likewise reflect the content found in the final descriptions and the range of item formats on the assessment. Evidence will be provided for the degree of congruence between the content of the exemplars and that of the descriptions. There will be at least three exemplars per level per grade identified.

Guideline 4

In carrying out its statutory mandate, the Board will *exercise its policy judgment in setting the levels.* However, in so doing, they will seek technical advice from a variety of sources, but especially from the contractor, who will prepare the recommendations on the levels, the descriptions, and the exemplar exercises, as well as from consumers of NAEP data, including policymakers, professional groups, the states and territories.

Rationale

Setting achievement levels is both an *art* and a *science.* As an *art,* it requires judgment. It is the Board's best policy judgment what the levels should be. However, as a *science,* it requires solid technical advice based on a sound technical process. The Board is committed to seeking such technical advice from a variety of sources.

Practices and Procedures

Technical Advice throughout the Process

The Board seeks to involve persons who have had experience in standard-setting at the state level, and from those who are users of the NAEP results. Regular presentations will be given to standing committees who advise on NAEP matters such as the Education and Information Advisory Committee (EIAC) of the CCSSO, and the NAEP NETWORK. Their counsel will be sought on matters of substance as the work of the Board progresses. The EIAC and other similar constituencies may also be invited to send a representative to all standing technical advisory committees of the Board's contractor(s) which deal with the level-setting process.

The Board will also seek advice from the technical community throughout the level-setting process. Efforts will be made to ensure that presentations are made regularly to such groups as the American Educational Research Association (AERA), the National Council for Measurement in Education (NCME), and the professional groups in the content areas such as the International Reading Association (IRA), the National Science Teachers Association (NSTA), and other similar organizations. The Board will seek to engage technical groups available to them, including the Technical Review Panel, the National Academy of Education, their own contractor(s), and NCES and its contractor(s), in constructive research studies focused on providing information on the technical aspects of NAEP related to level-setting (e.g., scaling, weighting, mapping ratings to the scale, etc.)

Validity and Reliability Evidence

The Board will examine and consider all evidence of reliability and validity available. These data would include, but need not be limited to, procedural evidence such as the selection and training of judges and the materials and methods used in the process, reliability evidence such as intra-judge and inter-judge consistency data, and finally, internal and external validity data. Such data will help to inform *the Board's policy decision as they set the levels.*

Procedural evidence, while informative, is not necessarily sufficient evidence for demonstrating the validity of the levels. Therefore, the conduct of the achievement level-setting process shall be implemented so that a series of both internal and external validation studies shall be conducted simultaneously. To the extent possible, in order to realize maximum efficiencies in the use of resources, validation studies shall be included in the design of the level-setting data collection activities. Such studies may include, but shall not be limited to, convergent and divergent validation efforts, for example, conducting alternate standard-setting methods or conducting cross-validation level-setting activities, as well as exploring alternate methods for refining and expanding the preliminary achievement levels definitions, and empirically examining various technical decision rules used throughout the process.

As part of the validation task, additional evidence as to the suitability and appropriateness of identifying the subject area content of the recommended achievement levels ranges and cut-scores will be gathered. This evidence may include, but need not be limited to, data resulting from behaviorally anchoring the ranges and/or cut-scores, or data resulting from some other alternative procedures that employ a more global approach other than the item content of the particular assessment. The results of these studies will provide a clear indication of what students know and can do at the levels.

The results from these validation efforts shall be made available to the Board in a timely manner so that the Board has access to as much validation data as possible as it considers the recommendations regarding the final levels. Kane (1993) suggests that an "interpretive argument would specify the network of inferences leading from the score to the conclusions drawn about examinees and the decisions made about examinees, as well as the assumptions that support these inferences." An interpretative argument which articulates the rationale for interpreting the levels shall accompany the presentation of proposed levels to the Board.

Again, to maximize the efficient use of resources and to minimize duplication of effort, it is highly desirable for contractors to coordinate the design of such studies with other agencies responsible for evaluating the level-setting activities.

Guideline 5

The achievement levels shall be the initial and primary means of reporting the results of the National Assessment of Educational Progress at both the national and state levels.

Rationale

In an effort to improve the form and use of NAEP the Board seeks to make the results of NAEP more accessible and understandable to the general public and to policy makers. The Board also supports the movement from norms-based assessments to standards-based assessments. Reporting the results of NAEP using the achievement levels accomplishes these ends to a greater degree than heretofore possible.

Practices and Procedures

Reporting What Students Know and Can Do

The purpose of most NAEP reports, but particularly those published under the auspices of the National Center for Education Statistics, is to report to the American public and others on the performance of students—that is, to report on *what students know and can do.* The purpose of the achievement levels is to identify for the American public what students *should know and should be able to do,* and to report the actual performance of students in relation to the achievement levels. Therefore, NAEP reports incorporate elements of both of these aspects of performance.

Clarity of interpretation of the NAEP data can be achieved by ensuring that the descriptions of performance for the levels and the exemplar exercises reflect what the empirical data show for a given assessment. This may be achieved by the modified procedures of *scale anchoring* [1] or by new procedures developed specifically for the purposes of providing elements of the content of the frameworks in the reporting mechanisms.

Reporting Student Performance

In describing student performance using the levels, terms such as *students performing at the Basic level* or *students performing at the Proficient level* are preferred over *Basic students* or *Proficient students.* The former implies that students have mastery of particular content represented by the levels, while the latter implies an inherent characteristic of individual students.

In reporting the results of NAEP, the application of the levels of *Basic, Proficient, and Advanced* applies to the three regions of the NAEP scale generated when the appropriate cut scores are mapped to the scale. However, three cut scores yield, in fact, four regions. The region referenced by content which falls below the *Basic* cut score will be identified by descriptors that are not value-laden.

Interpreting Student Performance

When interpreting student performance using the levels, one must diligently avoid over interpretations. For example, each of the NAEP subject areas are scaled independently of each other, even though each scale uses the same metric, i.e., scores ranging from 0 to 500. Because the metrics are identical, it does not follow that comparisons can be made across subjects. For example, a *Proficient* cut score of *235* in reading should not be interpreted to have the same meaning as a *Proficient* cut score of *235* in U.S. history. Neither should unwarranted comparisons be made in the same subject area from one assessment year to the next, unless the data for the two years have been equated and we have reason to believe that the scale itself has not changed from time 1 to time 2.

Guideline 6

The level-setting process shall be managed in a technically sound, efficient, cost-effective manner, and shall be completed in a timely fashion.

Rationale

Since a contractor(s) is conducting technical advisory and assistance work for the Board, it is critical that such work be performed to meet high quality standards, including efficiency, cost-effectiveness, timeliness, and adherence to sound measurement practices. *However, in the final analysis, it is the Governing Board that makes the policy decision regarding the levels, not the contractor.*

Practices and Procedures

The contractor(s) shall prepare a fully detailed Planning Document at the onset of the level-setting work. This document will guide the progress of the work, serve as a monitor, and be the basis for staff and Board supervision. The Planning Document will outline milestone events in the process, provide a chronology of tasks and subtasks, as well as a monthly chronology of all activities across all tasks, and detail all draft and final documents that will be produced, the audience for such reports, and the number of copies to be provided by the contractor.

Procedures adopted by a contractor(s) to carry out the level-setting process must encourage and support national involvement by the relevant and required publics. Such meetings will also be conducted in a physical environment which is conducive to work and planning. To the extent possible, current technology shall be used in all areas of the level-setting process to increase efficiency and to reduce error.

The contractor(s) shall work closely and in a professional manner with the NAEP operations contractor in striving to fulfill the requirements of the level-setting process by (1) making all requests for information and data in a timely manner, (2) providing all requested information and data in a timely manner, (3) adhering to all predetermined deadlines so as not to impede the work of the operations contractor, and (4) advising the operations contractor of all unusual findings in the data so that a concerted effort can be mounted to resolve the problem or issue at hand.

The contractor(s) shall develop the initial level-setting design adhering to sound measurement principles and ensure that the various components of the design (e.g., selection of judges) are congruent with current standard-setting research. In the implementation of such designs, they shall employ state-of-the-art training strategies and measurement practices.

The contractor(s) shall produce documents in a timely manner and make oral presentations upon request. Presentations may include, but need not be limited to, the Board's quarterly meetings, relevant Board committees, and professional and lay groups.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *(1985). Standards for Educational and Psychological Testing.* Washington, DC: APA.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement (2nd ed., pp. 508-600).* Washington, DC: American Council on Education.

Jaeger, R.M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10,* 3-6, 10, 14.

Kane, M. *(1993).* The validity of performance standards. Unpublished manuscript.

National Academy of Education (1992). *Assessing student achievement in the states. The first report of the National Academy of Education panel on evaluation of the NAEP trial state assessment: 1990 trial state assessment.* Stanford, CA: Author.

National Assessment of Educational Progress Authorization Act, (P.L. 110-279).

Reid, J.B. (1991). Training judges to generate standard-setting data. *Educational Measurement: Issues and Practice, 10,* 11-14.

Endnotes

1. The traditional scale anchoring procedures anchored at the 200, 250, 300 350 points of the scale (± 12.5 points), using a p = *.65,* and a discrimination of .30 with the next lower level. The modified anchoring procedures (tried in reading for 1992) anchored at the achievement levels cut scores ***(±. 12.5),*** using a p = .65, and no discrimination criterion.

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.

Attachment B

Expert Panel Meeting on NAEP Achievement Levels

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:	Wade Buckland Monica Gribben D. E. (Sunny) Becker	Date:	February 2, 2018

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

Expert Panel Meeting on NAEP Achievement Levels

Table of Contents

Expert Panel Meeting on NAEP Achievement Levels

National Assessment Governing Board Technical Support Project
January 10–11, 2018

As part of its efforts to update the policy statement, *Developing Student Performance Levels for the National Assessment of Educational Progress,* the National Assessment Governing Board (hereafter, "Governing Board") directed the Human Resources Research Organization (HumRRO) to convene a two-day meeting of recognized experts in standard setting. The Governing Board's current policy on setting achievement levels is now more than 20 years old and is undergoing revision to ensure it reflects current best practices in standard setting. In addition, the recent evaluation of NAEP achievement levels (National Academies of Sciences, Engineering, and Medicine, 2017) contained several recommendations for improving procedures and practices for setting and communicating the NAEP achievement levels. Some of those recommendations have implications for the policy revision, while other recommendations will be addressed through additional activities beyond the scope of the policy.

We were fortunate to assemble a stellar panel: **Dr. Michael Bunch** (Measurement Inc.), **Dr. Karla Egan** (EdMetric, LLC), **Dr. Steve Ferrara** (Measured Progress), **Dr. Ed Haertel** (Stanford University), **Dr. Ron Hambleton** (University of Massachusetts), **Dr. Laura Hamilton** (RAND), **Dr. Marianne Perie** (University of Kansas), and **Dr. Barbara Plake** (University of Nebraska-Lincoln). Four of the Experts participated in the recent evaluation of NAEP achievement levels (National Academies of Sciences, Engineering, and Medicine, 2017) — Dr. Egan and Dr. Hamilton as Committee members, Dr. Plake as an external reviewer of the report, and Dr. Haertel as Report Coordinator.

The meeting was held on January 10 and 11, 2018 in Alexandria, Virginia. Advance materials were provided in mid-December and included: a description of the meeting purpose and goals; an agenda with guiding questions and list of participants (see Appendix A); the current Governing Board policy on achievement levels setting; the recent evaluation of NAEP achievement levels (National Academies of Sciences, Engineering, and Medicine, 2017); and the Governing Board's formal response to each recommendation in the evaluation.

The expert panel (hereafter, "Experts") discussed various aspects of achievement levels setting to generate advice to the Governing Board's Committee on Standards, Design and Methodology (COSDAM) as it considers policy revisions. The meeting was organized around several broad topics: foundational issues including participants, procedures, and methodology; issues related to achievement level descriptions (ALDs); recommendations for revisiting performance standards over time; interpretation and use of achievement level results; and other issues. Each topic was introduced with a set of guiding questions, intended to prompt, but not to limit, discussion. This document is organized around those major topic areas.

Dr. Sunny Becker (HumRRO) opened the meeting with introductions and briefly presented the purpose and rationale for the meeting. Dr. Sharyn Rosenberg (Governing Board staff) then provided a detailed overview of NAEP achievement levels setting, including: historical information; the current policy and procedures; institutionalized procedures not reflected in the policy statement; planned minor updates; recommendations from the recent evaluation of NAEP achievement levels (National Academies of Sciences, Engineering, and Medicine, 2017); and the Governing Board's formal response to each recommendation contained in the evaluation. Dr. Rosenberg's presentation is included in Appendix B.

Experts quickly agreed that the policy statement should include high level guidance related to goals for what should be done in an achievement levels setting, but that specific details about how to carry out each goal should be relegated to a new "processes and practices manual."

The rich discussions resulted in several considerations for COSDAM. Most suggestions were not unanimously agreed upon (unless otherwise noted). Where Experts expressed differing opinions, the stated advantages and disadvantages of each approach are presented. Discussion topics have been grouped by theme and do not always reflect the order of conversation.

Participants, Procedures & Methodology

Selection of Standard Setting Panelists

The current policy (p. 6) specifies that two-thirds of the panel will represent teachers and other educators, and the remaining one-third will represent the public, non-educator sector. The policy also calls for broad representation by geographic region, urbanicity, ethnicity, and gender. In her opening presentation, Dr. Rosenberg noted that the guideline on panelist composition has been operationalized as 55 percent teachers in the given content area and grade; 15 percent non-teacher educators (e.g., university professors in the content area); and 30 percent professionals in the content area (e.g., children's book authors and editors for setting achievement levels on the 2017 NAEP Writing assessment at grade 4). The achievement levels setting contractor typically performs extensive outreach to nominators and then ranks potential panelists based on their qualifications and the desire to assemble a broadly representative panel according to the criteria included in the policy.

Experts discussed several potential criteria for individuals selected to participate in achievement level setting panels, in addition to those that appear in the current policy:

- Experience with the content and students in the grade above the grade of the assessment (e.g., grade 5 for grade 4 assessment and post-secondary for grade 12)
- Teachers of students with disabilities (SWD) who take NAEP
- Teachers of English Language Learners (ELL) who take NAEP
- Teachers of ethnically diverse students
- Representation from schools across performance levels (i.e., low, middle, and high performing schools)
- Teachers of students across ability levels (e.g., teachers of low-achieving and high-achieving students)
- Unrelated individuals (i.e., teachers from different schools who do not know each other)

Some Experts suggested that members of the general public either should not be included or should be reduced to only one or two individuals. They expressed concern that members of the general public often do not have the skills needed to contribute in a meaningful way (i.e., content or child development expertise). In their experience, teachers often need to educate the

public members on achievement level setting panels. Some Experts did note that states tend to include only one or two participants representing the public in standard setting panels for state assessments, such as tutors who are familiar with the content and students. On the other hand, one Expert noted that the inclusion of true general public panelists (i.e., parents or other individuals who are not necessarily professionals in the content area) may increase credibility, and that this has been common practice for many tests. That is, it may be harder to explain results to the public if the lay public is not represented in the process.

Experts recommended that specific percentages of types of individuals should be removed from the policy. They suggested creating detailed descriptions of the relevant expertise required on the panel when seeking nominations. The criteria for who should be included on the panel should be driven in part by the claims that the Governing Board wishes to make from NAEP results. For example, if the achievement levels are intended to indicate preparedness for college, then the standard setting panels should include postsecondary representatives.

Additionally, some Experts stated that people with a stake in where cut scores are set do not belong on standard setting panels. States often use a two-part process for setting achievement levels, where the first panel consists of content experts, and then a second panel includes some overlap with the first but may include some non-educators. In the case of NAEP, the Governing Board already serves a role similar to the second panel by taking the content-based recommendations and discussing any additional considerations in their role as the policy body for NAEP.

Number of Standard Setting Panelists

The current policy (p. 6) does not specify how many standard setting panelists should comprise each subject and grade but indicates that the size of the panel should be informed by current research and should aim to reduce the standard error of the cut score. In her opening presentation, Dr. Rosenberg noted that the number of panelists has been institutionalized as 20–22 for field trials and pilot studies, and 30–33 for operational standard setting panels. Each table typically consists of 5-6 people.

Some Experts suggested that the reference to Jaeger (1991) in the current policy could be updated with Raymond and Reid (2001), which recommends 12–15 panelists per subject and grade. From a group management perspective, most Experts thought that 30 is the maximum feasible with two groups of 15 participants each. Experts noted a need to include a sufficient number to allow for independent subgroups (to facilitate estimation of variability arising from group interaction during the standard setting process). They agreed that it is favorable to have a large number for a national assessment where goals include representation as well as reducing the standard errors of the cut score judgments, but they did not think it was necessary for the policy to specify an exact number.

The number of panel members for a pilot study would depend on the purpose of the pilot. A full "dress rehearsal" would require more participants than a small feasibility study.

Training and Role of Facilitators

The current policy (p. 5) specifies that facilitators be knowledgeable in the subject area and experienced, capable trainers in a large-group setting. Although not explicitly stated in the policy, in practice NAEP has used separate content and process facilitators for each subject and grade, guided by training and a detailed facilitator handbook. In her opening presentation, Dr.

Rosenberg noted that typically panelists have been divided into two rating groups (group A and B, within each subject and grade) to cover the entire item pool. There are some items common across both groups, and a subset of items is unique to each group. Both groups conduct their activities in the same room and share content and process facilitators.

Experts suggested that the policy should reference separate content and process facilitators and additional information about facilitator qualifications. There was clear consensus on the need to train facilitators and use scripts to guide the process, especially for introducing and interpreting impact data. In addition, Experts agreed that facilitators must present a neutral position and should be reminded that their role is not to persuade the participants. They noted that the guidance of the facilitators can strongly affect the outcome of deliberations. Further, if independent groups are used, facilitators and participants in each group should avoid cross-group discussions to maximize independence of the results.

There was extensive discussion about whether groups A and B should in fact share content and process facilitators, or whether each group should have their own facilitators and conduct most activities completely independently in separate rooms. However, there was no consensus on whether there should be one or two sets of facilitators. There are pros and cons for each option. With one set of facilitators, there is no way to quantify potential facilitator effects. Each group receives the same instructions, guidance, and level of assistance. Although the two groups work on separate groups of items, there is no way to estimate the true standard error since they are not truly independent. With two sets of facilitators, there was a common understanding that both groups would receive a shared introduction and general instruction from one of the facilitators and then break into two groups, each with its own set of facilitators. Using this procedure, any facilitator effects associated with differences among leaders would be reflected in the variability of group outcomes. Even with training and a script to guide the process, there may be facilitator effects due to individual differences in how facilitators handle questions and guide discussion. A benefit of two sets of facilitators is the ability to calculate a true standard error since the groups are independent. An alternate suggestion was to include three process facilitators, one person to oversee the two group leaders. Some Experts maintained that the decision to use one or two sets of facilitators should be based on the balance between the goals of improving the achievement level setting process versus accurately estimating the precision of the resulting cut scores.

Standard Setting Methodology

The current policy (p. 7) states that the methodology should be appropriate to the assessment format for the content area and have a solid research base. The policy references the modified Angoff procedure, which was used by NAEP for all achievement level settings conducted between 1990 and 1998. In her opening presentation, Dr. Rosenberg noted that since 2005, the Governing Board has used a modified Bookmark method for most achievement levels setting, with the exception of the Body-of-Work method for NAEP Writing. The policy does not specify a number of rounds, but typically 3 rounds have been used (4 rounds for the Mapmark procedure implemented for 2005 grade 12 mathematics).

Experts stated that selection of an appropriate methodology should depend on the assessment design, including but not limited to: item types; number of items; type of scoring; mode of assessment; and number of achievement levels. Experts agreed that the method should be flexible to allow different approaches and should focus on the principles that the standard setting is designed to achieve (e.g., consistency of panelist judgments) rather than outcomes.

Experts agreed that current best practice calls for multiple rounds of standard setting with clearly defined objectives for each round. The objectives will vary by assessment and process used for creating achievement level descriptors, therefore the policy should not prescribe an approach. Experts allowed for the possibility that future methodological improvements may lead to best practices not currently envisioned, such as eliminating the need for multiple rounds. Experts suggested flexibility in the policy regarding methodology so that new and better methodologies are not ruled out.

Use of Feedback and Impact/Consequences Data

The current policy does not contain information about feedback and impact/consequences data. In her opening presentation, Dr. Rosenberg noted that institutionalized practices include feedback data after each round (how the panelists' judgments compare to others in the same group and/or the full group), and impact/consequences data after rounds 2 and 3 (how the recommended cut scores would translate into percentages of students at or above each achievement level). In addition, panelists typically complete a Consequences Questionnaire at the end of the process which gives them an opportunity to disagree with the panel recommendation and provides the Governing Board with their rationale for alternate recommendations.

The use of impact data was discussed but there was no consensus about whether or how it should be used. Most Experts said that the policy should require impact data but others advised against introducing impact data. Those in support of using impact data indicated that the policy needs to specify why impact data are needed. It should focus on the principle behind the impact data, on *why* it is being used but not *what* is being used or *how* it is used. One Expert who felt strongly that impact data should not be used argued that it has the potential to undermine content-based judgments, and that it is the Governing Board's role to make judgments about impact.

In considering impact data, Experts noted a need to follow a logical progression of assessment design. What claims are we trying to make? What evidence do we need to make those claims? They suggested using pilot studies to assess the need for impact data, including which data to include, where in the process to consider it, how to apply impact data, the number of rounds to use, and the wording of directions. To use impact data, the sample must be large enough to answer the questions being asked.

Experts noted that best practices for using impact data typically include overall percent proficient at a possible cut score, by demographic groups. Given how achievement levels are used and the aspirational nature of the NAEP proficient cut point, one Expert suggested that it might be more useful to look at different data, such as impact data for top performing schools.

Role of Pilot Study

The current policy (p. 7–8) specifies that a pilot study should be conducted prior to the operational meeting to try out the materials, procedures, feedback, and software. In her opening presentation, Dr. Rosenberg explained that the pilot study has been institutionalized as a full "dress rehearsal" with only minor changes intended to take place for the operational meeting. If a new element of the process is in need of testing (e.g., a standard setting software being used for the first time), a field trial is performed in advance of the pilot study. A pilot study is performed for every subject and grade, but with a slightly smaller group of panelists (typically 20-22 for a pilot study, compared to 30-33 for an operational meeting). If procedures are not

modified substantially between the pilot study and operational meeting, then the Governing Board often considers the consistency of pilot study results and operational study results in its final deliberations when setting cut scores.

Experts discussed several aspects of the pilot study. They agreed that NAEP should have a rigorous process and that the pilot study should serve as a dress rehearsal, but one Expert noted that there is still a difference between a "dress rehearsal" and "opening night". That is, the operational meeting is intended to have more weight in the Governing Board's deliberations than the pilot study.

Some Experts noted that pilot results may have evidentiary value, but that a decision about whether or not they should be compared to the operational results should be made prior to conducting the operational meeting. If the pilot and operational assessments yielded similar results, the Governing Board's confidence in its decision on cut scores could be bolstered. However, Experts noted that it would not be appropriate to use the pilot study results in other ways, such as combining results across both meetings or presenting pilot study results to operational participants. It would not be appropriate, after the operational standard setting, to frame the relevance of the pilot results one way or another depending upon how closely they matched the operational results.

Calculation of Panelist Judgments and Variation of Judgments

The current policy does not include any information about how panelist cut scores should be summarized. Some Experts suggested using the median, instead of the mean, to reduce the effects of outliers. Others recommended allowing flexibility for robust statistics, such as using regression to find the midpoint, but noted that a discussion of median versus mean is too fine-grained for a policy statement.

The current policy does note that the standard error of the cut score "should not exceed the combined error associated with the standard error of measurement on the assessment and the error due to sampling from the population of examinees" (p. 6). Most Experts did not find this to be a meaningful rule of thumb. Some Experts noted that there is no "true" cut score, but the consistency of panelist judgments is part of procedural validity evidence. Panelist evaluations and information about the consistency of their judgments do address whether the process worked as intended.

One Expert recommended that information about the standard error should be reported to the Governing Board in terms of uncertainty in the percent at or above each achievement level rather than on the scale score. For example, the percent of student at or above Proficient might range from 30 to 50 percent when considering results within two standard errors of the recommended cut scores.

Public Comment

The current policy (p. 6) calls for public comment throughout the process, including on the Design Document and the proposed levels. In her opening presentation, Dr. Rosenberg noted that it has never been feasible to collect public comment on the achievement level results because those data are considered embargoed until they are officially released by the Commissioner for Education Statistics.

Experts suggested that the current language be revised to include comment only on ALDs for content and clarity. They did not advocate requesting comment on cut scores. Some Experts referenced Smarter Balanced and PARCC, who invited public comment on methodology only. Experts did not think it was advisable to convene a group of stakeholders to comment on the results via non-disclosure agreements (NDAs) because this activity has the potential to undermine the content-based judgment process.

Experts suggested seeking guidance from other government agencies (e.g., the Environmental Protection Agency) about how public comment is solicited or consulting a public relations firm to obtain more input.

Issues Related to Achievement Level Descriptions

The current policy includes general definitions for *Basic*, *Proficient*, and *Advanced*; there is no description for below *Basic* because it is not an official NAEP achievement level. The policy refers to two phases for developing content descriptions: preliminary ALDs developed with the assessment framework and final ALDs developed during the achievement levels setting process. In her opening presentation, Dr. Rosenberg noted that although the policy does not specify whether the ALDs should be finalized prior to or during the achievement levels setting process, since 1998 the common practice has been to finalize the ALDs prior to convening an achievement levels setting panel. The current policy does not address how and when ALDs should be reviewed or revised over time, and the recent evaluation of NAEP achievement levels (National Academies of Sciences, Engineering, and Medicine, 2017) recommended that such guidance be developed. In practice, the ALDs have been reviewed and revised using anchoring studies, only in response to framework changes.

Best Practices for Developing and Updating ALDs

Dr. Karla Egan kicked off the discussion on this topic by summarizing her recent work to document best practices for developing and maintaining ALDs. She stated that there is no comprehensive set of best practices for how to start writing ALDs. She explained that the literature at large has scattered recommendations on developing ALDs. Experts thus brainstormed best practices based on their experiences, including:

- Recruit and train good facilitators
- Specify how ALDs should be used
- Base ALDs on the claims expected to be made with assessment results
- Identify the desired grain size for ALDs
- Reference content actually measured by the assessment
- Do not rely on adverbs such as “some” or “most of the time” to distinguish levels

Panel members described the ALD process as iterative. They indicated that grain size should be based on claims to be made (i.e., how the ALDs will be used). Guidance regarding appropriate grain size might be included in the processes and practices manual, rather than specified in the policy document.

Similarly, Experts did not agree on whether the ALDs should be written in terms of "can" versus "should" statements. The NAEP ALDs currently use "should" statements. One Expert explained that *can* is after the fact, meaning that someone with a specific score can do stated tasks, whereas *should* refers to what someone should be able to do to receive a specific score.

Experts suggested using 4-5 people who will dig into the data for several days to write ALDs rather than a large group. Additional considerations for developing ALDs include:

- Write to midpoint of levels
- Hire cognitive theorists
- Focus on what students can do rather than what they cannot do (e.g., at below *Basic* level)
- Use item mapping for information about types of items students are likely to get correct, even in the below *Basic* category

Experts did not agree on whether borderline or threshold ALDs should be developed in advance and provided to standard setting panelists, or whether the standard setting panelists should be the group to develop them when needed (e.g., when using a modified Bookmark method). On the one hand, developing them in advance ensures that adequate time can be devoted to this task and guarantees consistency of borderline ALDs across both the pilot and operational meetings. Alternatively, the exercise of having panelists wrestle with developing borderline ALDs enhances their understanding of the tasks they are to perform. Experts agreed that if borderline ALDs are developed in advance, it would be important to implement other activities aimed at having standard setting panelists internalize them. One Expert suggested that research could be performed to address this question of whether or not it is preferable to develop borderline ALDs in advance.

Use of ALDs in the Item Development Process

Although Pellegrino, Jones, and Mitchell (1999) in *Grading the Nation's Report Card* and Hambleton et al.'s response (2000), both agreed that ALDs should inform item development, the Experts indicated that there is no evidence that item writers are able to effectively use ALDs. They suggested that item writers be involved from the beginning of the ALD writing process to shape the descriptors into something that they can use. Generally, more precise terminology is needed if ALDs are intended for item writing. Items often need to meet many criteria, so it is difficult to align with ALDs as well as other dimensions. Further, more goes into the difficulty of an item than the content being assessed (e.g., item type, stimulus, response type), so ALDs do not always help item writers predict the achievement level of an item. Therefore, it is not clear that ALDs would be useful to item writers.

Experts noted that some of the NAEP ALDs (i.e., mathematics grades 4 and 8) have not been revised since 1992 and are out of date. If ALDs are intended to be used for item development, they must be updated.

Multiple Types of ALDs

Currently, NAEP has two versions of achievement level descriptions: policy level and range. The Experts agreed that different types of ALDs should be thought of as an interrelated set of

tightly aligned descriptors with different uses. For example, threshold ALDs should be a subset of range ALDs. Reporting ALDs should be shorter, more general, and easier to understand. Experts noted that some standard setting methods do not use threshold ALDs. The Experts disagreed on whether threshold ALDs should be reported; some considered them primarily a tool for panelists' use during the standard setting process. Panels of content experts could develop or finalize ALDs prior to standard setting for range, threshold, and reporting, with two groups – one developing reporting ALDs and another crafting the range and threshold descriptors. Returning to the debate between *should* and *could* statements, Experts noted that *should* is most appropriate for the ALDs used in standard setting, whereas *can* is best used for ALDs for reporting.

Whether or not the Governing Board should consider additional types of ALDs depends on the intended uses of the NAEP achievement levels.

Consideration of a below Basic ALD

Experts engaged in extensive debate about whether the Governing Board should consider developing ALDs for the below Basic category. Many states, but not all, do have descriptions for the lowest category of performance on their assessments. Most Experts felt it was not necessary for NAEP to develop ALDs for below Basic. Given the purpose of NAEP and the lack of individual scores, there are no student or teacher score reports. In addition, it is difficult to describe what students know and can do in the below Basic category because there is no policy definition describing what these students should know and be able to do, and because their performance can range from falling just below the Basic cut score to not answering any items correctly. The NAEP item maps do include items below Basic so this is an alternative way of representing what students at a given score point in the below Basic range are likely able to do.

Two Experts felt strongly that NAEP should contain descriptions for below Basic. Even though there are no individual scores on NAEP, they noted that some jurisdictions do have large numbers of students in the below Basic category. ALDs for this category could be written in terms of what students may be able to do, or could describe the midpoint of the category.

Reviewing and Revising ALDs Over Time

Experts engaged in a rich discussion of this topic. There was consensus that specific details of when and how often to review and update ALDs is not policy, but practice. Suggestions for inclusion in a processes and practices manual included periodically: updating the language; reviewing to assess the need for change; studying the impact of accessibility and technology changes on ALDs and standards; considering scale drift; and examining potential changes in dimensionality. Experts suggested using an event-based change system. For example, if there has been a change in curriculum or a change to the framework, then a review should be completed. Changes to item types, composition of students being assessed, and instructional techniques would trigger a review. A review might, but would not necessarily, lead to changes to the ALDs. In addition to the primary event-based system, a secondary time-based system should be in place too. If no events trigger a review, Experts suggested looking at the linkage, ALDs and cut scores once every 5–10 years or every other administration for assessments that are administered less frequently than reading and mathematics.

Additional suggestions included periodic reviews of standard setting policy and ALDs by psychometricians. This could be accomplished by the measurement experts on the Governing Board, COSDAM, or a special Technical Advisory Committee (TAC).

Revising Performance Standards Over Time

The recent evaluation of NAEP achievement levels (National Academies of Sciences, Engineering, and Medicine, 2017) included a recommendation for the Governing Board to implement a regular cycle for considering the desirability of conducting a new standard setting. The current policy does not include any information about the need to revisit cut scores.

Some Experts felt that performance standards should only be changed when the underlying assessment construct has changed and all evidence fails to link the old and new assessments. Experts agreed that changes to scales and achievement levels cause confusion and backlash from users and stakeholders, so should be undertaken with caution. They suggested establishing a panel to conduct standards validation and verification and review of cut scores to ensure they are appropriate, when the underlying assessment has been changed (e.g., a change to the framework and ALDs). Measurement and content experts should be included. Experts observed the tendency for measurement experts to err toward maintaining trend (by using equating to maintain trend and cut scores) while content specialists are more likely to see new constructs being measured that necessitate changes to scales and cut scores. The Experts suggested that only when bridge studies and all other evidence fails to link a previous assessment to the new one, should NAEP cut scores be changed.

One Expert who participated in the recent evaluation of NAEP achievement levels interpreted the recommendation on revisiting performance standards as not necessarily being time-based. Instead, it was intended to reflect a need to revisit the entire assessment system to ensure that it is still defensible to continue using the current cut scores as other aspects of the system have changed.

Pilot bridge studies can serve as an early warning of the need to break trend. If these studies point to an inability to accurately link the old and new assessments, then there may be a need to initiate work on setting new cut scores.

Many states use NAEP achievement levels for benchmarking when setting cut scores on their own assessments, so a change to the NAEP achievement levels could trigger actions by states, as well.

Interpretation and Use of Achievement Level Results

The recent evaluation of NAEP achievement levels (National Academies of Sciences, Engineering, and Medicine, 2017) includes a recommendation to conduct research, articulate, and provide validity evidence related to the intended interpretations and uses of the achievement levels. The current policy does not specify how achievement level results are intended to be used, although there is some general information about reporting (p. 11).

Experts engaged in a lively discussion of use and interpretation of achievement level results. They agreed that there must be an easily accessible (i.e., does not require many clicks from the home page) “interpretive guide” that makes interpreting the data easy. Experts recommended that the policy refer to a need for an interpretative guide but that the policy not attempt to delineate the appropriate uses of NAEP achievement levels, since this may change over time and can be specific to a given subject. The guide should accompany the release of the Nation’s Report Card. Some information can be the same for each assessment, while other information would be customized. Common misuses should be included with a rationale for why certain uses are inappropriate, although caution should be used in highlighting misuses so that they do

not reflect negatively on NAEP. The Experts recommended a simple version of results and interpretation, something that the lay person can understand.

Many misuses of NAEP data occur when people make inappropriate causal conclusions, interpret NAEP Proficient as representing grade level performance, or construe gap trends using achievement level results. The Experts discussed the need for additional guidance in tracking gaps. They suggested referencing Ho and Haertel's (2007a, 2007b) policy brief on using NAEP for the type of guidance to provide to school districts when they are trying to reduce gaps and use NAEP to monitor gap trends. Ho (2008) discusses how using scale scores is a better approach than achievement levels when comparing gaps over time.

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

Other Issues

Throughout the meeting, several issues of a more general nature were raised.

- Achievement levels and labels (i.e., Basic, Proficient, and Advanced) have value and should change only if mandated by legislation. States co-opted the term proficient when it became a requirement in NCLB. They use the same term even if it does not have the same meaning. **There was strong consensus that confusion over the meaning of NAEP Proficient should be addressed through communication efforts, not by changing the terminology.**

- Consider establishing a standing Technical Advisory Committee (TAC) for the Governing Board on achievement levels and/or frameworks, including validation of achievement levels. Currently any TACs used by the Governing Board are project-specific and do not consider overarching issues. If such a group were formed, it would be important to keep its role distinct from the panels that advise NCES by only focusing on technical issues in the Governing Board's domain.

- Uncertainty about the standard setting process is relevant primarily for the Governing Board; it is not important to report with the results. It is not a meaningful question to think about how close your result is to some true value since this does not exist.

- The current policy does not provide much guidance about documentation, but it is important to document and release details of the standard setting activities in a timely matter.

- Many testing programs hire independent observers or evaluators to attend standard setting meetings and write a report. NAEP does have TACSS members observe standard setting meetings, but they do not write a formal report. The Governing Board

could consider formalizing the current process or hiring an external entity to serve this role.

- Exemplar items should continue to be used to provide meaning to the achievement level results, but the specific details of how to select them (e.g., RP value) do not belong in the policy statement.

References

Hambleton, R.K., Brennan, R.L., Brown, W., Dodd, B., Forsyth, R.A., Mehrens, W.A., Nellhaus, J., Reckase, M., Rindone, D., van der Linden, W.J., & Zwick, R. (2000). A response to "setting reasonable and useful performance standards" in the National Academy of Sciences' Grading the Nation's Report Card. *Educational Measurement: Issues and Practice, 19*(2), 5-14.

Ho, A. D., & Haertel, E. H. (2007a). (Over)-interpreting mappings of state performance standards onto the NAEP scale. *Paper commissioned by the Council of Chief State School Officers*.

Ho, A. D., & Haertel, E. H. (2007b). Apples to apples? The underlying assumptions of state-NAEP comparisons. *Paper commissioned by the Council of Chief State School Officers*.

Ho, A. D. (2008). The problem with "proficiency": limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, *37*(6), 351-360.

Jaeger, R.M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice*, *10*(2), 3-6, 10, 14.

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. Doi: 10.17226/23409.

Pellegrino, J.W., Jones, L.R., & Mitchell, K.J. (Eds.). (1999). *Grading the Nation's Report Card: Research from the Evaluation of NAEP.* Washington, DC: National Academy Press.

Raymond, M.R., & Reid, J.R. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119-157). Rahway, NJ: Lawrence Erlbaum.

Appendix A: Meeting Agenda and Attendees

Expert Panel Meeting on NAEP Achievement Levels
National Assessment Governing Board Technical Support Project
January 10 –11, 2018 | Agenda

DAY 1

9:00 – 9:30	**Welcome, Introductions, and Meeting Goals**	Dr. Sunny Becker
9:30 – 10:30	**Overview of NAEP Achievement Levels Setting**	Dr. Sharyn Rosenberg
10:30 – 10:45	**Review & Revise Agenda Topics**	Dr. Sunny Becker
10:45 – 11:00	**Break**	
11:00 – 12:30	**Foundational Issues: Participants, Procedures & Methodology**	Dr. Sunny Becker

Guiding Questions:

1. What should the policy include in terms of the type of panelists and their qualifications and experience?
2. What guidance should be included about the selection of a standard setting methodology?
3. Should the policy specify recommended numbers of panelists for the pilot study and operational meeting?
4. How should the policy address the type and timing of impact data provided to panelists?
5. What procedures should be used to evaluate and describe the consistency of panelist judgments?
6. How should the policy address issues related to precision of cut scores?
7. How should the participants and/or procedures for seeking public comment be modified from the current policy?

12:30 – 1:30	**Break for lunch**	
1:30 – 2:30	**Foundational Issues: Continuation of Discussion**	
2:30 – 2:45	**Best Practices for Developing and Updating ALDs**	Dr. Karla Egan
2:45 – 5:00[1]	**Issues Related to Achievement Level Descriptions**	Dr. Thanos Patelis

Guiding Questions:

1. What are best practices for developing ALDs?
2. How should the ALDs be used in the item development process?
3. Should NAEP consider using multiple versions of ALDs – e.g., range ALDs, threshold ALDs, reporting ALDs?
4. How and when should the ALDs be reviewed and updated over time?

6:00 **Meet for optional group dinner:** *a la Lucia* (315 Madison St, Alexandria, VA 22314).

[1] Fifteen minute break approximately 3:15 – 3:30

DAY 2

9:00 – 9:15	**Review of Previous Day and Plan for Today**	Dr. Sunny Becker
9:15 – 10:15	**Recommendations for Revisiting Performance Standards over Time**	Dr. Thanos Patelis

Guiding Questions:

1. What information is needed to evaluate the need for revision over time?
2. What criteria should trigger a review of performance standards for possible revision?
3. Should performance standards be revisited with a specific minimum frequency?
4. What implications should be kept in mind when performance standards are revised?

10:15 – 12:15[2]	**Interpretation and Use of Achievement Level Results**	Dr. Art Thacker

Guiding Questions:

1. Should the policy provide guidance on communicating the meaning of NAEP Basic, Proficient, and Advanced and how they may differ from other common uses of those terms?
2. What should be included in the section on validation?
3. Should the policy include a general statement of appropriate uses for the achievement levels, or should it describe a process for determining appropriate uses for a given assessment?
4. How should the policy provide guidance on helping users to determine the inferences that are best made with achievement levels and those best made with scale scores?
5. How should exemplars be selected and reported with the NAEP results?

12:15 – 1:15	**Break for lunch**	
1:15 – 2:45	**Other Issues**	All

Guiding Questions:

1. Are there other important elements missing from the policy?
2. Are there other aspects from the evaluation of NAEP achievement levels that should be addressed by the policy?

2:45 – 3:00	**Wrap-up**	Dr. Sunny Becker

[2] Fifteen minute break approximately 11:00 – 11:15

Attendees

Expert Panelists:
Dr. Michael Bunch, Measurement Inc.
Dr. Karla Egan, EdMetric, LLC
Dr. Steve Ferrara, Measured Progress
Dr. Ed Haertel, Stanford University
Dr. Ron Hambleton, University of Massachusetts
Dr. Laura Hamilton, RAND
Dr. Marianne Perie, University of Kansas
Dr. Barbara Plake, University of Nebraska-Lincoln

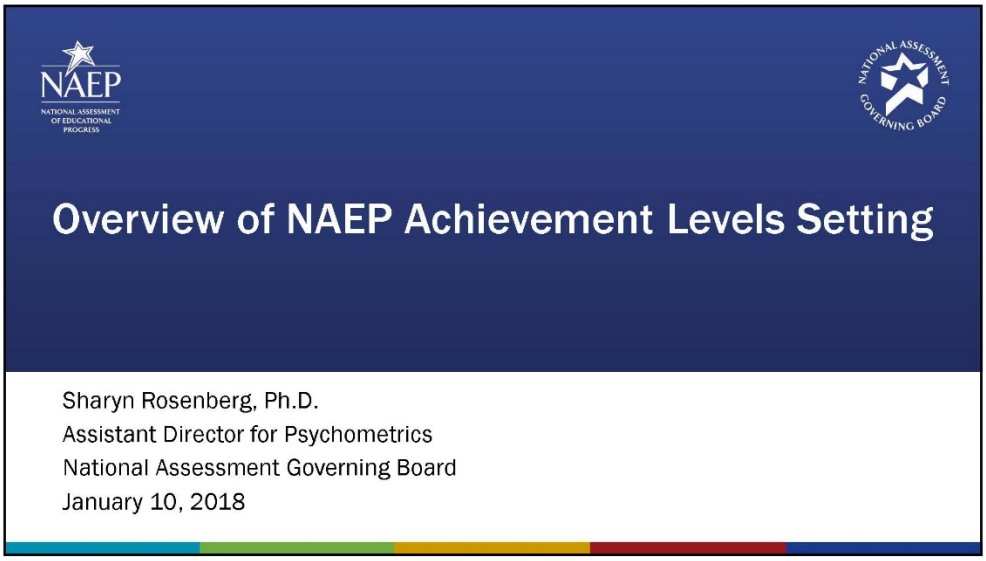
NAGB Staff:
Ms. Michelle Blair
Dr. Sharyn Rosenberg
Dr. Lisa Stooksberry

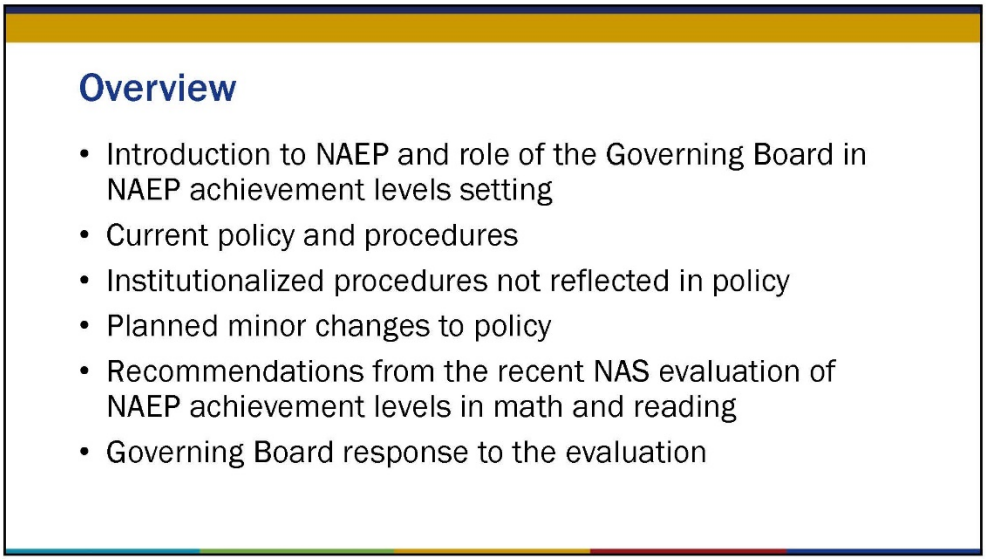
HumRRO:
Dr. Sunny Becker
Mr. Wade Buckland
Dr. Monica Gribben
Dr. Thanos Patelis
Dr. Arthur Thacker

NCES:
Dr. Enis Dogan

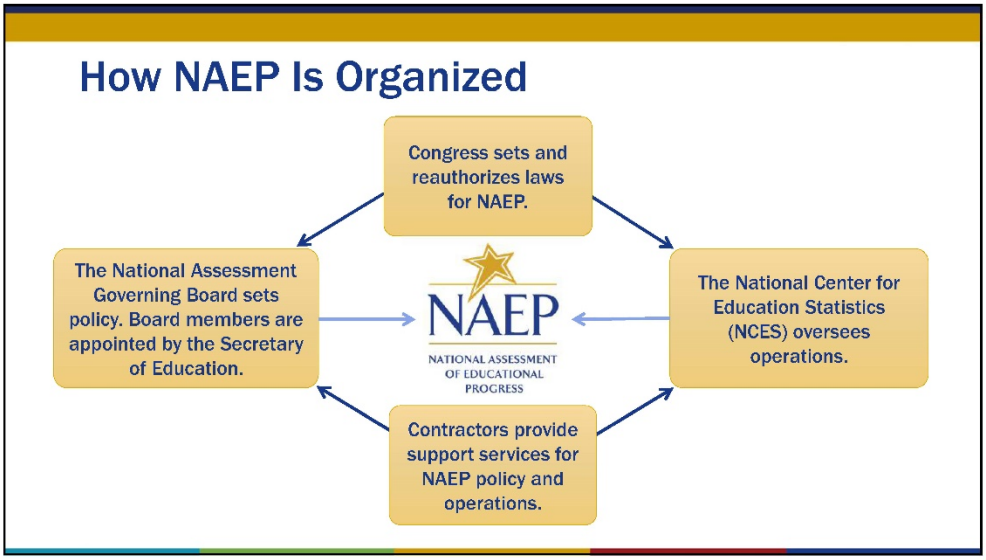
ETS (NAEP Design, Analysis, and Reporting Contractor):
Dr. Mary Pitoniak

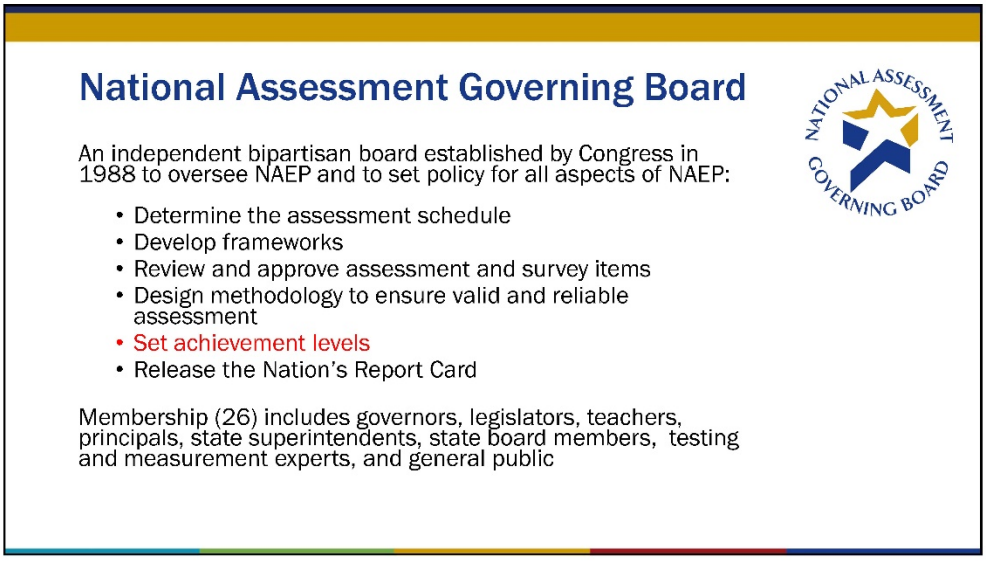
Appendix B: Presentation Slides





1


How NAEP Is Organized
Congress sets and reauthorizes laws for NAEP.
The National Assessment Governing Board sets policy. Board members are appointed by the Secretary of Education.
NAEP
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS
The National Center for Education Statistics (NCES) oversees operations.
Contractors provide support services for NAEP policy and operations.


National Assessment Governing Board
An independent bipartisan board established by Congress in 1988 to oversee NAEP and to set policy for all aspects of NAEP:
• Determine the assessment schedule
• Develop frameworks
• Review and approve assessment and survey items
• Design methodology to ensure valid and reliable assessment
• Set achievement levels
• Release the Nation's Report Card
Membership (26) includes governors, legislators, teachers, principals, state superintendents, state board members, testing and measurement experts, and general public

2

Congressional Authorization

- 1988 – P.L. 100-297: Achievement goals
- 1994 – P.L. 103-382: Performance levels
- 2001 – P.L. 107-110: Achievement Levels

1994 and 2001 authorizations state that the levels should be used on a developmental (trial) basis until the Commissioner determines as a result of an evaluation that such levels are reasonable, valid, and informative to the public

Achievement Levels Policies

- Original policy (adopted May 11, 1990)
  - Focus on initial effort to set achievement levels on 1990 mathematics
- Current policy (adopted March 4, 1995)
  - Based on general policy revision from 1993
  - Minor changes to references and addition of foreword in 2007

Some differences between 1990 and 1995 in **policy definitions**, methodology, inclusion of specific guidelines

3

Achievement Levels Policy Definitions

Original Policy Level Definitions (1990) (Wording remaining in current definitions.)	Current Policy Level Definitions (1995) (Additional wording since 1990)
Proficient. This central level represents solid academic performance for each grade tested—4, 8, and 12. It will reflect a consensus that students reaching this level have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling. For grade 12, the proficient level will encompass a body of subject-matter knowledge and analytical skills, cultural literacy, and insight that all high school graduates should have for democratic citizenship, responsible adulthood, and productive work.	Proficient. This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.
Advanced. This higher level signifies superior performance beyond proficient grade-level mastery at grades 4, 8, and 12. For grade 12, the advanced level will show readiness for rigorous college courses, advanced technical training, or employment requiring advanced academic achievement. As data become available, it may be based in part on international comparisons of academic achievement and may also be related to Advanced Placement and other college placement tests.	Advanced. This level signifies superior performance beyond Proficient.
Basic. This level, below Proficient, denotes partial mastery of knowledge and skills that are fundamental for proficient work at each grade tested. For grade 12, this will be higher than minimum competency skills (which normally are taught in elementary and junior high schools) and will cover significant elements of standard high school level work.	Basic. This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.

Guideline 1: Current Policy Definitions

- **Proficient**: This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter.
- **Basic**: This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.
- **Advanced**: This level signifies superior performance beyond proficient.

4

Development of Achievement Level Descriptions

Policy specifies two phases of development for content-specific ALDs:

- **Preliminary ALDs**: To be developed as part of Framework (to inform item development)
- **Final ALDs**: ALDs are updated to set achievement levels

Policy does not specify whether or not ALDs should be finalized prior to or during the ALS process

- Since 1998, common practice has been to finalize the ALDs in a separate step (by convening a panel of content experts) prior to beginning the achievement levels setting process

Updating Achievement Level Descriptions

- Policy does not address how and when ALDs should be reviewed or revised over time

- In practice ALDs have been reviewed and revised using anchoring studies, only in response to framework changes
  - Math ALDs reviewed in 2003 for grades 4 and 8 but not revised
  - New Math ALDs for grade 12 in 2005 with new framework
  - Math ALDs for grade 12 reviewed and revised to reflect revisions to 2009 NAEP Math Framework at grade 12
  - New NAEP Reading Framework (all 3 grades) in 2009 led to new ALDs but the trend lines and cut scores were maintained based on bridge studies

5

Panelist Training

- Policy specifies that participants receive training on assessment framework, standard setting methodology, and specific judgment task
- Policy specifies that facilitators be knowledgeable in subject area and experienced, capable trainers
- Explicit requirement that panelists take NAEP assessment under the same conditions as students

Guideline 2: Widely Inclusive Activity Composition and Size of Panels

- Policy specifies 2/3 teachers and other educators, and 1/3 representing "the public, non-educator sector, for example, scholars, employers, parents, and professionals in occupations related to the content area"
  - In practice, this has been operationalized as 55% teachers, 15% non-teacher educators, and 30% professionals in the content area (e.g., published authors and editors for Writing)
- Policy specifies that size of panels should be responsive to research with the goal of reducing the standard error of the cut score
  - In practice, field trials and pilot studies conducted with 20-22 panelists and operational studies conducted with 30-33 panelists at each grade

6

Review Procedures

- Policy specifies that Design Document and interim reports be disseminated to groups with legitimate interest in the process
- Explicit requirement that proposed levels be "widely distributed to major professional organizations, state and local assessment and curriculum personnel, business leaders, government officials, the Planning and Steering Committees of the framework development process, the Exercise Development panels, and other groups that may request them."
- In practice, public comment on Design Document, limited engagement of NAEP panels when appropriate, achievement level results treated as embargoed data prior to official release

Guideline 3: Resulting Products Methodology

Not specified but emphasis on modified Angoff in 1995 policy

- Appropriate to content area and assessment framework
- Solid research base
- In practice, modified Angoff was used through 1998 standard settings
- Since 2005, variations of modified Bookmark/Mapmark and BOW have been used
- Since 2011, standard setting process conducted on computers

7

Methodology

- Requirement for pilot study – has been institutionalized as:
  - Full dress rehearsal – only minor changes intended
  - New elements first tried out in field trial
  - Pilot study is conducted for every subject/grade
  - Results from pilot study serve as replication of operational results (with caveats)
- Quality control procedures
  - Policy very specific but does not account for shift to computerized processes
- Final cut scores determined by the Board, not the contractor

Descriptions of the Levels

- Reference to ALDs mostly redundant with Principle 1
  - Policy allows for possibility that standard setting panel will revise/finalize ALDs
  - Since 1998 this is separate step prior to convening ALS panel
  - Exception is if ALS panel identifies a major issue/concern with using ALDs in standard setting process
- Exemplars for NAEP reporting
  - Selected from pool of released items
  - Should reflect performance at each level and threshold scores
  - Should be selected to meet rp = 0.50 criterion
  - (But item maps on Nation's Report Card use rp = 0.65, rp = 0.72)

8

Guideline 4: Advice from Technical Community

- Technical Advisory Committee on Standard Setting (TACSS)
  - Operationalized as 6 members, including:
    - 1 member of the NAEP Design, Analysis and Reporting contractor (typically has been Mary Pitoniak)
    - 1 member of a state testing agency
  - Specific to a given standard setting, consultants to contractor
  - Other groups/NAEP committees engaged as needed
- Validity and reliability evidence to inform Board action
  - Procedural evidence
  - Intra-judge and inter-judge consistency data
  - Internal and external validity data

Guideline 5: Reporting Practices and Procedures

- Achievement levels shall be the initial and primary means of reporting NAEP results at the national and state levels
- Emphasis on what students **should** know and be able to do
- Use of exemplar items (also noted in Principle 3)
- Promoting appropriate interpretations and use
  - "When interpreting student performance using the levels, one must diligently avoid over interpretations"

9

Guideline 6: Guidance for Level-Setting Contractor

- Prepare documents and make presentations, including a Planning Document to guide the project
- Encourage and support national involvement
- Use current technology
- Work closely with NCES
- Develop a sound design

Institutionalized Procedures Not Explicitly in Policy

- ALS meetings are usually 4-5 days, with very thorough training
- Typically 3 rounds of ratings
- Use of multiple rating groups per grade (typically 2; 3 for TEL)
  - Entire item pool is used and divided among two groups
  - Subset of common items
  - Groups A and B are in same room with one process and one content facilitator
  - Tables of 5-6 panelists

10

Institutionalized Procedures Not Explicitly in Policy

- Use of evaluations throughout the process (following each major activity and end of day)
  - Feedback reviewed in real time to address issues/concerns
  - Panelists asked to indicate whether they are ready to proceed to ratings
  - Final questionnaire asks whether panelists would be willing to sign a statement (after reading it of course) recommending the use of the cut scores from the ALS process

Institutionalized Procedures Not Explicitly in Policy

- Use of feedback and impact/consequences data
  - Feedback after each round
  - Since 1998, impact/consequences has been presented to panelists
  - Current practice is to present impact data after rounds 2 and 3
  - Final consequences questionnaire
- Technical report of ALS procedures is made available to the public following the release of NAEP results

11

Planned Minor Changes

- Update professional references and citations
- Update references to external groups and legislation
- Update and clarify terminology (e.g., change "judges" to "panelists" and "growth" to "changes across the full range")
- Re-organize principles and remove duplicative information
- Reflect shift to digital-based assessments
- Ensure that policy reflects current practices

Evaluation of NAEP Achievement Levels for Math and Reading

- To address requirement in NAEP legislation
- Contract to perform this work was let by the National Center for Education Evaluation and Regional Assistance (NCEE)
  - Independent of the Governing Board and NCES
- Work was performed by the National Academies of Sciences, Engineering, and Medicine via an expert panel
- Report was issued in November 2016
- NAEP legislation requires Board to formally respond with 90 days
- Governing Board response sent to the Secretary of Education and Congress in late December 2016

12

Evaluation of NAEP Achievement Levels for Math and Reading

- Focus of evaluation was math and reading achievement levels
  - Current Reading achievement levels for grades 4, 8, and 12 were set on the 1992 assessments using a modified Angoff methodology
    - New NAEP Reading Framework in 2009 resulted in updated ALDs (from anchoring study) but not new scale or achievement levels
  - Current Math achievement levels for grades 4 and 8 were set on the 1992 assessments using a modified Angoff methodology
    - New NAEP Math Framework in 2005 resulted in new scale, ALDs, and achievement levels at grade 12 (using a Mapmark methodology)
    - Minor updates to NAEP Math Framework in 2009 resulted in new ALDs at grade 12 to reflect academic preparedness for postsecondary endeavors
- Current practices and procedures have evolved since 1992

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.

13

Board Response to Recommendation #1

- Board will issue a procurement to conduct studies that evaluate whether the ALDs should be revised, based on their alignment with the NAEP framework, item pools, and cut scores
- Board will first update the achievement levels policy (in conjunction with response to Recommendation #3) to address the larger issue of specifying a process and timeline for conducting regular recurring reviews of the ALDs
- Board felt strongly that cut scores for reading and math should not be changed in 2017 and has a Resolution on this topic

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.

14

Board Response to Recommendation #2

- This decision is in the purview of the NCES Commissioner
- The Board is committed to providing the NCES Commissioner with the necessary information to make this determination

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.

15

Board Response to Recommendation #3

- 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.

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.

16

Board Response to Recommendation #4

- NCES and the Governing Board have been working on statistical linking studies in several grades and subjects
- The Governing Board's Strategic Vision includes an explicit goal to increase opportunities for connecting NAEP to other national and international assessments and data.

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.

17

Board Response to Recommendation #5

- Research currently underway 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.

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.

18

Board Response to Recommendation #6

- 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.
- 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.

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.

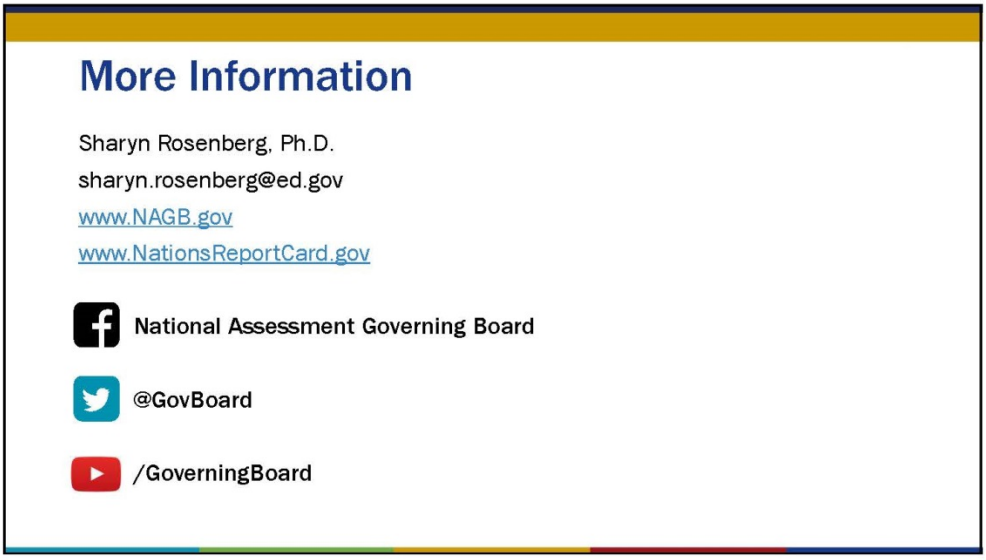
19

Board Response to Recommendation #7

- 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.

Questions?

20


More Information
Sharyn Rosenberg, Ph.D.
sharyn.rosenberg@ed.gov
www.NAGB.gov
www.NationsReportCard.gov
National Assessment Governing Board
@GovBoard
/GoverningBoard

21

Toward Coherence in Assessment Systems through Achievement Level Descriptors Using the NAEP Example

Karla Egan, Ph.D.
Anne Davidson, Ed.D.
EdMetric, LLC[1]

SECTION I. INTRODUCTION

Achievement level descriptors (ALDs[2]) are widely used in K-12 assessment programs as they provide meaning to test scores by defining the knowledge, skills, and abilities of students at specified levels of performance. The use of ALDs has been widely associated with standard setting and score reporting; however, another use has been lurking in the literature for well over two decades—the use of ALDs to guide test design and development (e.g., Hansche, 1998; Pellegrino, 2014). Indeed, this use is found in the policy documents concerning item development (National Assessment Governing Board, 2002) and concerning standard setting (National Assessment Governing Board, 1995) for the National Assessment of Educational Progress (NAEP). Even so, the use of ALDs to guide test design and development has not gained traction until recently. With the rise of principled assessment design (e.g., evidence-centered design) and validity formulated as an evidentiary argument, the field has begun to explore the use of ALDs to guide test design efforts.

Granularity refers to the degree of specificity the ALD provides and relates to the purpose and intended use of the ALD. The ALDs that guide test design and development will necessarily

[1] This paper was produced under subcontract to the Human Resources Research Organization (HumRRO) as part of contract number ED-NAG-17-C-0002 with the National Assessment Governing Board: Technical Support in Psychometrics, Assessment Development, and Preparedness for Postsecondary Endeavors.

[2] Achievement level descriptors are also called performance level descriptors (PLDs) in the literature.

be detailed explications of the content area. As such, they will have a different look and feel than the ALDs often associated with assessment programs which have traditionally come out of the process of standard setting and are placed on score reports. When ALDs are used for stakeholder understanding (e.g., placed on individual score reports), this necessitates a less refined grain size for ALD development and reporting purposes. A more refined level of granularity is required for ALDs used to guide item writing than those ALDs written to describe achievement levels.

The measurement field has indiscriminately applied the term "ALD" to all uses. This is a bit like asking for a screwdriver without specifying if a Phillips-head or flat-head is needed. Increased precision in language regarding ALDs improves understanding and communication. To this end, Egan, Schneider, and Ferrara (2012) proposed a typology of ALDs, shown in Table 1, that are interrelated but vary in their intended uses, audiences, and (we add) granularity.

TABLE 1. TYPES OF ACHIEVEMENT LEVEL DESCRIPTORS WITH INTENDED USE, AUDIENCE, AND GRANULARITY

ALD Type	Intended Use	Audience	Granularity
Policy	High-level description of expected performance in each achievement level	Policy makers	Least refined level. A single set is created for the testing program.
Range	Details the knowledge and skills *expected of and/or demonstrated by* students across the range of achievement within a performance level. These descriptors are typically written for each content strand.	Educators, item writers	Most refined level. Usually created for the level of the content standards for which items will be written.
Target	Describes the knowledge and skills *expected of* students right at the cut score.	Standard setting panels	Mid-level refinement. Usually written at the level of the reporting category.
Reporting	Summarizes the knowledge and skills *demonstrated by* students right at the cut score	Parents, educators	Mid-level refinement. Usually written at the level of the reporting category.

In brief, the policy ALD guides the development of the other three ALD types. Policy ALDs are developed as one of the first steps in a testing program. They set the tone for policy expectations of the testing program. Range ALDs are developed either in conjunction with the content frameworks or immediately after the content frameworks. They are written at the same level for which items will be written. If items are written for each content strand, then Range ALDs should be developed at the strand level. The Target ALDs are written prior to standard setting. Starting with the Range ALDs, the Target ALDs aggregate the knowledge, skills, and abilities (KSAs) that best discriminate performance near the cut scores. The Target ALDs are operationalized through the standard setting process. The Reporting ALDs are based on cut score placement, and they are written immediately after the cut scores are accepted by the sponsoring agency.

This paper examines the potential use of the ALD framework in the context of NAEP and how the framework might impact the National Assessment Governing Board (Governing Board) policies related to standard setting and to test design and development. In particular, we want to examine how the delineation of ALD type may affect the Governing Board's policies on:

- Using multiple types of ALDs, including ALDs for item writing
- Using "should" versus "can" in ALDs
- Writing descriptors for the lowest achievement level category

To do this, we first examine the framework in more detail in the second section of this paper. We survey the ALD literature, specifically as it relates to test design and development in the third section of this paper. In the context of this paper, the generic term "ALD" is used when examining the literature. In many cases, the authors did not refer to the granularity of the ALD even when recommending the ALDs be used for test design and development. In the fourth section, we look at the way that ALDs are currently developed for and used by NAEP and other entities. In the fifth section, we examine potential uses for ALDs.

SECTION II. THE ALD FRAMEWORK

Figure 1 shows the development pathway of the ALD types as well as the relationship between the four ALD types. We discuss each type of ALD in more detail within this section. It is important to understand that Policy, Range, and Threshold ALDs are created based on inputs from educators, cognitive scientists, and other learning theorists, and they reflect our best theories on how students demonstrate knowledge. The Reporting ALDs are based on how students perform on the test.

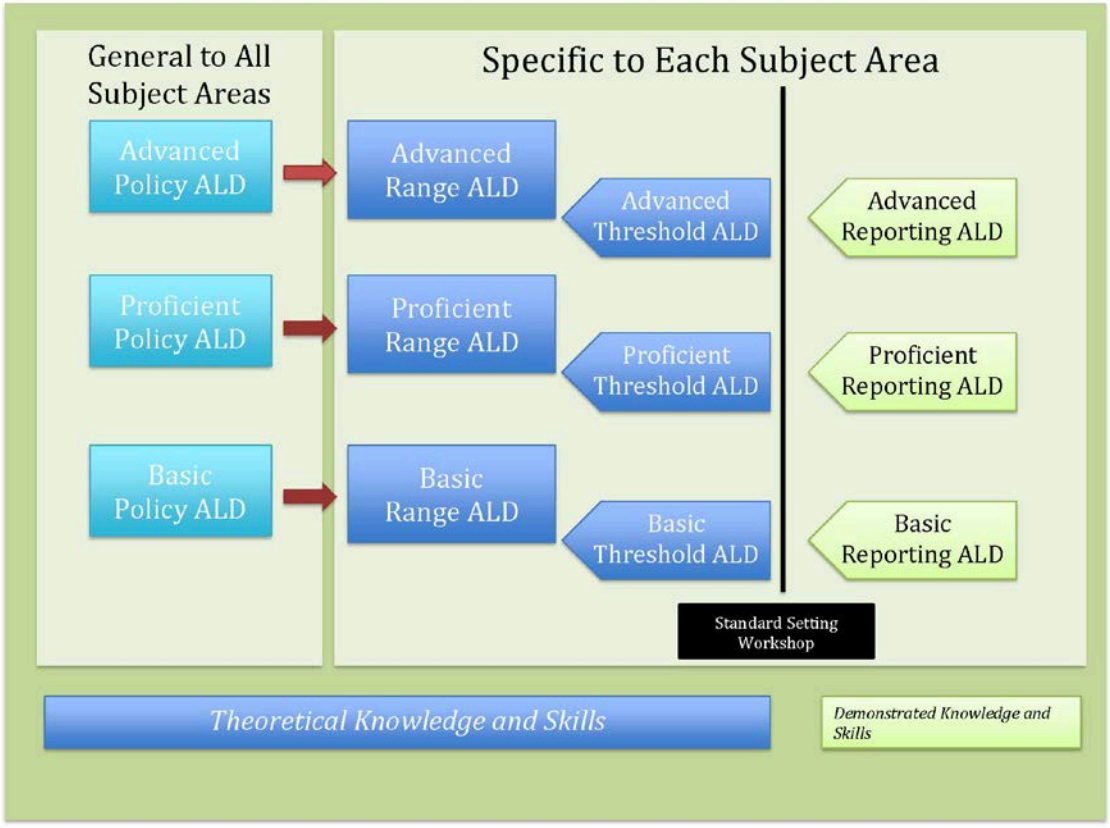


FIGURE 1. LINKED SYSTEM OF ALDS (BASED ON EGAN, SCHNEIDER, & FERRARA, 2012)

Policy ALDs. These descriptors are generally short, high-level statements describing student performance. Policy ALDs are used by policy makers as a way of communicating the intent of the achievement level. They do not focus on content-related knowledge and skills;

rather, they are statements about student performance that are general to the testing program or are general to a content area within a testing program. Table 2 shows the policy descriptors for both NAEP and Smarter Balanced. The NAEP policy descriptor is general to the program, and the Smarter Balanced policy descriptors are general to the content area within the testing program. They are presented to show levels of performance side-by-side and allow for comparison across four levels in terms of substantive meaning as well as text characteristics.

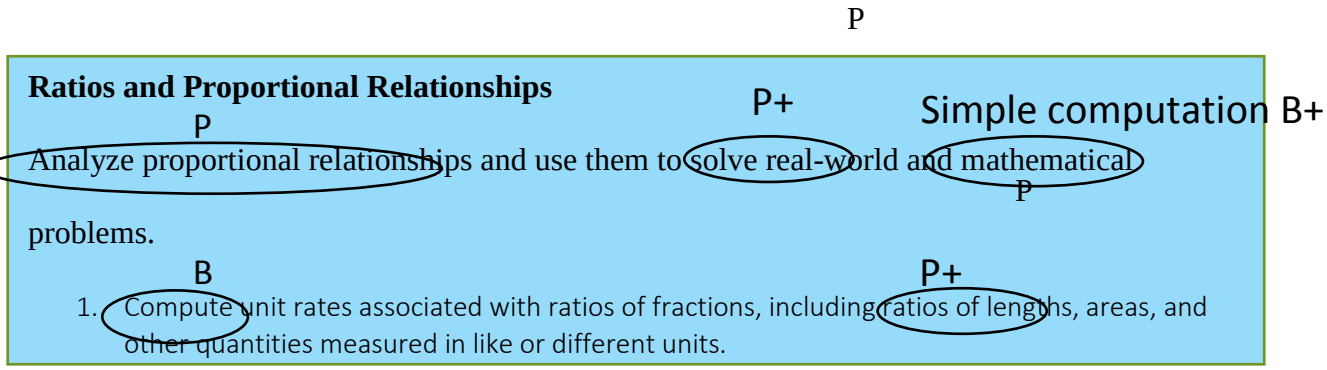
TABLE 2. POLICY DESCRIPTORS FOR NAEP AND SMARTER BALANCED

NAEP	Smarter Balanced
Advanced. Superior performance beyond proficient.	**Level 4**. Student demonstrates thorough understanding of and ability to apply the knowledge and skills associated with college content-readiness.
Proficient. Solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.	**Level 3.** Student demonstrates adequate understanding of and ability to apply the knowledge and skills associated with college content-readiness.
Basic. Partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.	**Level 2.** Student demonstrates partial understanding of and ability to apply the knowledge and skills associated with college content-readiness.
Below Basic. No descriptor.	**Level 1.** Student demonstrates minimal understanding of and ability to apply the knowledge and skills associated with college content-readiness.

The importance of the policy descriptor rests with its "defining phrase" and its label. These aspects of the policy descriptor set the tone for the testing program. Within the NAEP descriptors, the defining phrases are "solid academic performance" and "competency over challenging subject matter" for Proficient, "superior performance" for Advanced, and "partial

mastery" for Basic. Within the Smarter Balanced descriptors, the defining phrase is "adequate understanding" for Level 3, "thorough understanding" for Level 4, "partial understanding" for Level 2, and "minimal understanding" for Level 1.

The Governing Board chose to use descriptive labels for the NAEP achievement labels. Smarter Balanced, on the other hand, chose numeric labels. It may be that numeric labels engender fewer connotations than words. Burt and Stapleton (2010) showed that word labels (e.g., Proficient) connote certain inferences for panelists. At that point, the most widely-used label was "Proficient" (Egan et al., 2012). It is unknown how the widespread use of that label influenced respondents. As numeric labels are used, it will be interesting to see if certain numbers (e.g., 3) connote the same inferences as words do.



Range ALDs. These ALDs "articulate the intended construct so that items are written and tests developed to align with expected achievement from the very beginning of the test development process," (Egan et al., 2012, p. 103). It is important to emphasize the intended granularity of these ALDs.

FIGURE 2. DECONSTRUCTED CONTENT STANDARD

The Range ALDs are written by deconstructing content standards in terms of expected student performance. To do this, we first identify the KSAs within the content standards that

would be expected of the *Proficient* student. We then adjust those KSAs to reflect the other achievement levels. Figure 1 shows an example of this parsing activity. Here, *P*- represents the skills of the student entering Proficient, *P* represents the skills of the average Proficient student, and *P*+ represents the skills of the highly-Proficient student. The *B*s represent the KSAs expected of the Basic student.

The deconstruction process serves as a jumping off point for Range ALD authors. The task of parsing out the content standards is finite and approachable (albeit not easy). It gives writers source material from which to make initial judgments. Once the writers have parsed out the content standards, then they must use their own knowledge and experiences to adjust the expectations for the remaining performance level. This deconstruction process serves to closely align the Range ALDs to the content standards.

As the writers construct the Range ALDs, they will write statements that reflect their expectations for student performance. They will articulate the knowledge and skills that students *should* be able to demonstrate in each achievement level. These *should* statements will be based on the expertise of the Range ALD writers, who may be educators, cognitive scientists, and/or content experts.

Target ALDs. These ALDs focus on the specific area of the Range ALDs that will be used for standard setting. For example, the Target ALDs will describe the knowledge and skills of the students right at the cut score in a Bookmark (Lewis, Mitzel, Mercado, & Schulz, 2012) or Angoff (1971) standard setting. Target ALDs are developed prior to standard setting, and Target ALD writers select the knowledge and skills from the Range ALDs that they believe best differentiate between achievement levels. At this point, the knowledge and skills are still based

on expectations of student performance rather than actual student performance. As such, Target ALDs are still written as *should* statements.

Reporting ALDs. These summary statements are intended to communicate the meaning of student performance to stakeholders. These ALDs may be found on individual student reports or on an assessment-related website. By their nature, they are not as broad as the Policy Descriptors. Reporting ALDs are developed once the cut scores have been finalized. They will be specific to a grade/content area. When using item-mapping procedures, they will be based on items that students in a particular level were able to answer correctly. The Reporting ALDs are written as *can* statements because they are based on evidence provided by items that students within a group answer correctly. Ideally, Reporting ALDs should not be based on a single test form; instead, Reporting ALDs should be based on several forms or on an item bank.

The Governing Board's current subject-matter ALDs most closely resemble ALDs in the Reporting category, except that the NAEP ALDs discuss the knowledge and skills that students *should* do instead of the knowledge and skills that students *can* do. In addition, the NAEP subject-matter ALDs are developed prior to standard setting.[3]

ALDs and the Validity Argument

This "linked system of (A)LDs … serve to define the construct that is being measured and describe what students should know and be able to do in relation to the construct. When a clear definition of the target of measurement exists, a more fully aligned assessment system is created" (Egan et al., 2012, p. 80). The interrelated system of ALDs are then a strong source of evidence for the validity argument, and there should be strong alignment between the Range,

[3] When Threshold ALDs are needed for standard setting (e.g., Bookmark), they are created by the standard setting panelists at the beginning of the workshop. The Threshold ALDs are not reported (Fitzpatrick & Hickey, 2016).

Target, and Reporting ALDs. If developed correctly, there should be obvious alignment between Range and Target ALDs because Target ALDs are a subset of the Range ALDs. It follows that the alignment between Range and Reporting ALDs is central to the validity argument such that Reporting ALDs capture consistent summarization of Range ALDs within each level.

The use of Range ALDs as a primary source of evidence in a validity argument makes good sense. Evidence-centered design (ECD) has drawn attention to the importance of grounding test development in research-based models of student learning and cognition (Mislevy & Haertel, 2006; Pellegrino, Chudowsky & Glaser, 2001). Additionally, best practice in assessment development demands that evidentiary arguments for validity start with the proposed use and interpretation of test scores (Kane, 2006). Range ALDs offer a vehicle for pulling together these theoretical design elements with empirical understandings of how student learning occurs.

By definition, all types of ALDs specify the way that test developers intend to interpret test scores. By specifying intended interpretations for various levels, Range ALDs articulate the intended progression of knowledge across the test scale. Nonetheless, no form of an ALD is treated as a primary piece of evidence that test designers create and use at the beginning of an assessment program (Egan et al., 2012; Perie, 2008). Instead, ALDs' long association with standard setting means that they are typically developed after items are written but before cut scores are set. In some cases, ALDs (usually Reporting ALDs) are written at the very end of the test development process. ALD development early in the process allows for them to inform test blueprints and content development. When ALDs are included in a foundational document for test development, they promise greater coherence between the test items, intended uses and interpretations of resultant scores, and even classroom practice by reflecting learning progressions.

When developed at the beginning of a testing program, Range ALDs are a rich source of evidence that can be used for item writing and blueprint design. The Range ALDs can articulate the importance of test claims in each grade level. This information can be operationalized in the test blueprint. By using Range ALDs, item writers can see into the intent of the test designers. For example, through Range ALDs test designers can articulate their expectations for how students perform on each test claim (or test strand). Alignment to standards is enhanced because item writers are able to purposefully develop items that are intended to measure the test claims at various areas of the test scale.

If the item writers are successful, then the items can be statistically mapped to specific areas of the test scale. In the best possible case, the item writers would create items for the Proficient range that actually map to the Proficient range of the test scale. When the Reporting ALDs are created, then the knowledge and skills demonstrated by successful performance on the items in the Proficient range are the same knowledge and skills that were articulated in the Range ALDs.

By creating Range ALDs at the beginning of the test development process, we create an artifact that guides test development. At the end of the process, the Range ALDs can be evaluated once data are collected by comparing them to the Reporting ALDs. The intended interpretation (Range ALDs) can be compared to the enacted interpretations (Reporting ALDs) created from items' performance in the field. This means that there should be an ongoing validation of the Range ALDs against the Reporting ALDs. When they differ, this is an opportunity to evaluate whether the expectations of the Range ALDs should be changed to reflect actual student performance or the item writing should be adjusted for better measurement within an achievement level.

This ongoing validation should occur once the item bank has sufficient breadth and depth to support an analysis of items (e.g., 200 items). Past research has shown that the Reporting ALDs will not completely align from test form to test form (Schneider, Egan, Kim, & Brandstrom, 2008); however, these Reporting ALDs were created from non-ECD systems where a content-sampling approach was used. Nonetheless, this suggests that Reporting ALDs should be constructed from an item bank instead of a test form and that Range ALDs should be validated against an item bank.

SECTION III. REVIEW OF LITERATURE

As its own entity, the literature on ALDs is rather thin—ALDs are most often mentioned in the context of standard setting. In contrast, our literature review focuses on the use of ALDs in test development. The interested reader is referred to Egan et al. (2012) for a historical review of ALDs. In brief, the use of ALDs in K-12 assessment finds its roots in the 1990s when the Governing Board adopted ALDs for use with the NAEP. State educational agencies soon followed suit by writing and adopting ALDs for the purposes of standard setting and score reporting. By the early 2000s, almost all state educational agencies had adopted ALDs in one form or another.

CALLS FOR THE USE OF ALDS TO GUIDE TEST DEVELOPMENT

The idea that ALDs should be used for test design and development has been around since the 1990s when the National Assessment Governing Board developed preliminary ALDs that were supposed to guide item writers and test developers (Pellegrino, Jones, & Mitchell, 1999). At the time, researchers concluded that it was not clear that the preliminary ALDs were being used to guide item writers. Researchers noted that NAEP's item pools did not clearly

reflect the preliminary ALDs (Pellegrino et al., 1999; Mills & Jaeger, 1998). This led Pellegrino et al. (1999) to recommend that, “Preliminary achievement-level descriptions should guide the development of assessment items and exercises” (p. 177).

In related literature, researchers called for better alignment between assessment policy (as represented through ALDs) and assessment outcomes (Haertel & Lorie, 2004; Kane, 1994; Linn, 2001; Mills & Jaeger, 1998). These ideas would be articulated by Bejar, Braun, & Tannenbaum (2007) when they called for a prospective design for standard setting using an ECD framework. Bejar and colleagues (2007) proposed that performance standards be written at the beginning of the test development process (following the creation of multi-grade content standards and a competency model). The performance standards would be refined through an iterative process that would inform the development of evidence and task models. The creation of ALDs would flow from this process, and the ALDs would be used to guide test specifications. The Bejar et al. (2007) paper proposed a framework for developing ALDs using ECD, and they proposed how ALDs could be used in the test development process.

USING ALDS TO GUIDE ITEM WRITING

Drawing on the work of Bejar and colleagues (2007), the College Board developed ALDs using an ECD framework when building Advanced Placement assessments in history and science. Using an iterative process, ALDs were written at the claim level; however, claims and evidence were often further refined based on ALD development (Plake, Huff, & Reshetar, 2010; in non-ECD speak, the College Board developed ALDs following the development of content standards). This iterative process meant that the claims and evidence were embedded within the ALDs themselves (Hendrickson, Huff, & Luecht, 2010). Once ALDs were developed, the College Board created task models and templates to align to ALDs so they were ordered along

the achievement continuum (Hendrickson et al., 2010). This means that the ALDs had a direct impact on the directions given to item writers.

In 2012, the Smarter Balanced Assessment Consortium developed ALDs based on their content specifications (CTB/McGraw-Hill, 2013). This resulted in ALDs written to each assessment target. The ALDs were then embedded into the item specifications being used by item writers (see Figure 3). It is not clear from the College Board or Smarter Balanced efforts how well the item writers were able to incorporate the ALDs into their own work nor how much the ALDs impacted the items written.

Achievement Level Descriptors:	
RANGE Achievement Level Descriptor (Range ALD) Target D: Solve problems involving the four operations and identify and explain patterns in arithmetic.	**Level 1** Students should be able to represent and solve one-step problems using addition and subtraction within 100 and multiplication and division within 100.
	Level 2 Students should be able to solve two-step problems using addition and subtraction with numbers larger than 100 and solutions within 1000, assess the reasonableness of an answer, and identify patterns in the addition table.
	Level 3 Students should be able to solve two-step problems using multiplication and division within 100. They should be able to represent the problem using equations with a letter or symbol to represent an unknown quantity. They should also be able to explain patterns in the multiplication table.
	Level 4 Students should be able to use the properties of operations to explain arithmetic patterns (including patterns in the addition and multiplication tables).

FIGURE 3. PORTION OF SMARTER BALANCED GRADE 3 MATHEMATICS CLAIM 1 TARGET D ITEM SPECIFICATIONS (HTTP://WWW.SMARTERBALANCED.ORG/ASSESSMENTS/DEVELOPMENT/)

OUTCOMES OF ALDS AND ITEM WRITING

There is some literature investigating the use of ALDs to guide item writing. Ferrara, Sventina, Skucha, and Davidson (2011) instructed item writers to create mathematics items aligned to ALDs for a summative statewide assessment. Once cut scores were set, Ferrara et al., (2011) compared the items' intended ALD to their actual ALD. Across all grade levels studied, 35 percent of the items were accurately targeted (57 of 161 total items). Schneider, Huff, Egan, Gaines, and Ferrara (2013) asked item writers to assign items to range or target ALDs built from

test items. Across the two grades studied, item writers correctly assigned 33 percent of items to target ALDs and 37 percent of items to range ALDs.

In the context of the ALD literature, a few studies have attempted to ascertain the item features that contribute to item difficulty [see Hambleton & Jirka (2006) for an historical review of the literature]. In theory, these item features could provide a framework for developing ALDs that can better guide item writing. Ferrara et al., (2011) coded items to a framework for cognitive response demands (e.g., reading load, depth of knowledge) and linguistic demands. Of the features studied, Ferrara et al., (2011) found reading load was related to item difficulty. Building on Ferrara et al., (2011), Schneider et al. (2013) had item writers code items for features related to cognitive response demands. Schneider et al., (2013) did not find a relationship between the item features and item difficulty; however, they did conclude that differences in difficulty were likely due to "subtle nuances in content" (p. 112). Kaliski, Huff, and Barry (2011) asked subject matter experts (SMEs) in history to list the features that contributed to the difficulty of history items. The SMEs listed features that were domain independent (e.g., degree of scaffolding, word count). Both the Schneider and Kaliski studies call for a process in which ALDs and items are developed iteratively so that the field can better understand the cognitive and contextual features that correspond to item difficulty.

USING ALDS TO INFORM TEST BLUEPRINTS

Forte, Towles, Greninger, Buchanan, and Deters (2017) studied the relative alignment of achievement levels to test blueprints, looking at whether items were aligned to the test blueprint as well as to the ALDs. They asked how ALDs reflect measurement targets and whether the assessment "system was reasonable and sound" (p. 7). Ostensibly, ALDs capture artifacts of the content standards to which the test is written. Content standards appear as embedded within

ALDs' statements that describe performance within a grade level. Further, ALDs may exclude content standards that do not appear on the test, such as those that cannot be supported by the test delivery system. In this way, ALDs could be used to develop test blueprints (Forte, 2017) in such a way that score interpretation, item development, and test content are synced up and more closely matched with typical instructional patterns.

SUMMARY

This section shows that Range ALDs have the potential to help test developers as they develop items, task templates, and test blueprints. If Range ALDs are developed early in the test development process, they offer potential to guide item writers. Practically, Range ALDs define the construct that is to be measured by the assessment, specifically defining the "processes, strategies, and knowledge structures that are involved in item solving" (Embretson & Gorin, 2001 p. 349) for each level of achievement. Current attempts at using less-specific ALDs to guide item development have been insufficient (Ferrara et al., 2011).

We did not find literature that asked item writers for input regarding ALD usage. Range ALDs for item writing may need to be formulated differently than current ALDs to address the needs of item writers. In addition, it is not clear how much information that item writers can actually use when creating items. More research is needed to understand all of the aspects that items writers consider in order to know if or how Range ALDs may be added.

SECTION IV. CREATING ALDS

The use of ALDs is prescribed for state summative assessments by Critical Element 6.2 in Federal Peer Review, which states: *The State's academic achievement standards are challenging and aligned with the State's academic content standards such that a high school student who scores at the proficient or above level has mastered what students are expected to know and be able to do by the time they graduate from high school in order to succeed in college and the workforce.* The Peer Review guidance does not specify a particular method for creating ALDs; however, it ensures that all states develop ALDs in some form. This section of the paper describes the way that ALDs are created by different assessment programs.

NAEP ALDS

The NAEP ALDs are created in a two-phase process. Preliminary ALDs are created as the assessment frameworks are developed. The preliminary ALDs are intended "to guide item development and initial stages of standard setting" (National Assessment Governing Board, 2013, p. 44). The Governing Board policy further clarifies that, "(t)he preliminary descriptions are *working descriptions* for the panels while doing the ratings. These may be expanded and revised accordingly as these panels conduct the ratings, examine empirical performance data, and work to develop their final recommendations on the levels" (National Assessment Governing Board, 1995, p. 8). Bourque (2009) says that, in practice, the ALDs have been finalized prior to the standard setting since 1998. Figure 4 shows a subset of the preliminary ALDs created for the Technology and Engineering Literacy (TEL) assessment.

	BASIC	*PROFICIENT*	*ADVANCED*
	Students know that:	**Students know that:**	**Students know that:**
Technology and Society	Technology interacts with society, sometimes bringing about changes in a society's economy.	Technology interacts with society, sometimes bringing about changes in a society's economy and culture, and may lead to new needs and wants.	Technology interacts with society, sometimes bringing about changes in a society's economy, politics, and culture, and often leading to the creation of new needs and wants.
Design and Systems	One tool is better than another for a given task as a result of prior improvements.	Tools have been improved over time to further the reach of hands, voices, memory, and the five human senses.	Tools have been improved over time to further the reach of hands, voices, memory, and the five human senses, and to do so more efficiently, accurately, or safely.
Information and Communication Technology	Some information available through electronic means is exaggerated or wrong.	Increases in the ease by which knowledge can be published have heightened the need to check sources for possible distortion, exaggeration, or misrepresentation.	There are various effective ways to check information available through electronic means to identify possible distortion, exaggeration, or misrepresentation.
	Students are able to:	**Students are able to:**	**Students are able to:**
Technology and Society	Identify the impacts of a given technology in a society.	Identify the impacts of a given technology in a society, and predict how it might affect a different society.	Compare the impacts of a given technology on two different societies, noting factors that may make a technology appropriate and sustainable in one society but not in another.
Design and Systems	Design and build a simple model that meets a requirement.	Design and build a simple model that meets a requirement, fix it until it works (iteration), test it, and gather and display data that describe its properties using graphs and tables.	Design and build a simple model that meets a requirement, fix it until it works (iteration), test it, and gather and display data that describe its properties using graphs and tables. Interpret the results to suggest improvements and new designs.
Information and Communication Technology	Select and use appropriate digital and network tools and media resources to collect, organize, and display data.	Select and use appropriate digital and network tools and media resources to collect, organize, analyze, and display supporting data to answer simple questions and test basic hypotheses.	Select and use appropriate digital and network tools and media resources to collect, organize, analyze, and display supporting data to answer complicated questions and test hypotheses, and explain how to go about this to others working on the same project.

FIGURE 4. PRELIMINARY ALDS FOR GRADE 8 TECHNOLOGY AND ENGINEERING LITERACY

A second phase occurs in which a small group of experts (nine panelists developed the TEL ALDs) participate in a short workshop to create summary descriptors from the preliminary descriptors. The panelists will be a combination of committee members who developed the assessment frameworks and people who are new to the process. All panelists have expertise in the content and grade level. The committee's ALDs are vetted through a public review process

and the Committee on Standards, Design, and Methodology (COSDAM). The expert committee and the Governing Board staff finalize the ALDs to account for provided feedback. It is up to the Governing Board to adopt the ALDs. Figure 5 shows the final ALDs for the TEL assessment that were adopted by the Governing Board in 2014. Once the final ALDs are created, they replace the preliminary ALDs in the content framework documents.

Basic:	Eighth grade students performing at the Basic level should be able to use common tools and media to achieve specified goals and identify major impacts. They should demonstrate an understanding that humans can develop solutions by creating and using technologies. They should be able to identify major positive and negative effects that technology can have on the natural and designed world. Students should be able to use systematic engineering design processes to solve a simple problem that responsibly addresses a human need or want. Students should distinguish components in selected technological systems and recognize that technologies require maintenance. They should select common information and communications technology tools and media for specified purposes, tasks, and audiences. Students should be able to find and evaluate sources, organize and display data and other information to address simple research tasks, give appropriate acknowledgement for use of the work of others, and use feedback from team members (assessed virtually).
Proficient:	Eighth grade students performing at the Proficient level should be able to understand the interactions among parts within systems, systematically develop solutions, and contribute to teams (assessed virtually) using common and specialized tools to achieve goals. They should be able to explain how technology and society influence each other by comparing the benefits and limitations of the technologies' impacts. Students should be able to analyze the interactions among components in technological systems and consider how the behavior of a single part affects the whole. They should be able to diagnose the cause of a simple technological problem. They should be able to use a variety of technologies and work with others using systematic engineering design processes in which they iteratively plan, analyze, generate, and communicate solutions. Students should be able to select and use an appropriate range of tools and media for a variety of purposes, tasks, and audiences. They should be able to contribute to work of team collaborators (assessed virtually) and provide constructive feedback. Students should be able to find, evaluate, organize, and display data and information to answer research questions, solve problems, and achieve goals, appropriately citing use of the ideas, words, and images of others.
Advanced:	Eighth grade students performing at the Advanced level should be able to draw upon multiple tools and media to address complex problems and goals and demonstrate their understanding of the potential impacts on society. They should be able to explain the complex relationships between technologies and society and the potential implications of technological decisions on society and the natural world. Given criteria and constraints, students should be able to use systematic engineering design processes to plan, design, and use evidence to evaluate and refine multiple possible solutions to a need or problem and justify their solutions. Students should be able to explain the relationships among components in technological systems, anticipate maintenance issues, identify root causes, and repair faults. They should be able to use a variety of common and specialized information technologies to achieve goals, and to produce and communicate solutions to complex problems. Students should be able to integrate the use of multiple tools and media, evaluate and use data and information, communicate with a range of audiences, and accomplish complex tasks. They should be able to use and explain the ethical and appropriate methods for citing use of multimedia sources and the ideas and work of others. Students should be able to contribute to collaborative tasks on a team (assessed virtually) and organize, monitor, and refine team processes.

FIGURE 5. FINAL ALDS FOR GRADE 8 TECHNOLOGY AND ENGINEERING LITERACY

It is not clear when the second phase occurs. Guideline 3 of the Governing Board standard setting policy states, in part, "expanded descriptions of the content expected at each level based on the preliminary descriptions provided through the national consensus process" (National Assessment Governing Board, 1995, p. 7). Further explanation of this guideline asserts, "(the ALDs) will reference performance within the three regions created by the cut scores" (National Assessment Governing Board, 1995, p. 8). This text implies that the final ALDs are created after the cut scores are set; however, Bourque (2009) clarifies that, in practice, final ALDs have been adopted before the final cut scores are set since 1998.

There are some practices that are particular to the Governing Board. Per the Governing Board policy, descriptors are not created for the *below Basic* category. In addition, ALDs are written in terms of what students *should know and be able to do*, not what they *can do*. In the case of the TEL, the committee developed the ALDs by achievement level. In other words, one group created the *Basic* ALD, another created the *Proficient* ALD, and a third created the *Advanced* ALD. The composition of the groups changed throughout the workshop so that each panelist worked on each ALD before the end of the workshop (WestEd, 2014).

SMARTER BALANCED ALDS

The Smarter Balanced Assessment Consortium released Policy, Range, and Target ALDs in Spring 2013. These ALDs were created during a multiday workshop involving educators from Smarter Balanced member states. The interested reader may find details of the development of the ALDs in the technical report (https://portal.smarterbalanced.org/library/en/technical-report-initial-achievement-level-descriptors.pdf). Here, we focus on the development of the Range and Target ALDs.

Smarter Balanced invited four educators per grade and content area to develop Range and Target ALDs for Grades 3 – 8. Seventeen educators per content area developed Range and Target ALDs for Grade 11. (A larger group was used for Grade 11 so that higher education faculty and K-12 educators could participate.) The Grade 11 ALDs were created first and provided to the other grade levels. Within each grade level, Range ALDs were created for each assessment target by identifying the knowledge and skills that a Level 3 student should be able to do, followed by the knowledge and skills that a Level-1, -2, or -4 student should be able to do. Once Range ALDs were created, the panelists created the Target ALDs. The ALDs were written by target for all achievement levels. At the end of the workshop, a meta-committee examined the Range and Target ALDs for cross-grade cohesion. See Figure 6 for an example of a Smarter Balanced Range ALD.

Like NAEP ALDs, Smarter Balanced ALDs were released for public review and feedback. The feedback was incorporated by content experts, and the ALDs were adopted by the Smarter Balanced Governing States in 2013 (Smarter Balanced, 2015). Unlike NAEP, Smarter Balanced developed descriptors for Level 1. Also, Smarter Balanced deconstructed the content specifications to create learning progressions for each assessment target.

Number and Operations – Fractions				
RANGE ALD **Target F:** Develop understanding of fractions as numbers.	Level 1 students should be able to identify a fraction as a number and identify a fraction on a number line when the increments are equal to the denominator.	Level 2 students should be able to understand a fraction $1/b$ as the quantity formed by 1 part when a whole is partitioned into b equal parts; recognize simple equivalent fractions; express whole numbers as fractions; and recognize that comparisons are valid only when the two fractions refer to the same whole.	Level 3 students should be able to understand a fraction a/b as the quantity formed by a parts of size $1/b$; represent a fraction on a number line with partitioning; generate simple equivalent fractions and recognize when they are equal to whole numbers; and compare two fractions with the same numerator or the same denominator by reasoning about their size.	Level 4 students should be able to explain why two fractions are equivalent and approximate the location of a fraction on a number line with no partitioning.

FIGURE 6. SMARTER BALANCED RANGE ALD

ADVANCED PLACEMENT ALDS

In 2010, the College Board used ECD to build AP assessments in history and science. During this process, College Board staff worked with experts to develop ALDs as the claims and evidence were being derived. The College Board invited three to six subject experts to first

create the subject-specific ALDs followed by discipline-level ALDs during a two-day workshop[4]. Each subject-specific group developed ALDs for all "enduring understandings" within their subject area. To do this, the group selected 1 to 3 sets of claims and evidence to exemplify the top three of five achievement levels. The subject-specific ALDs were created by extrapolating student performance from the exemplar claims and evidence to all achievement levels (Levels 3, 4, and 5).

The group did not write ALDs for Achievement Levels 1 and 2. [Plake et al. (2010) note that these could be developed later.] The final task of the workshop was for the subject-specific groups to synthesize the subject-specific ALDs for the discipline. Plake et al. (2010) note that the synthesis task largely fell to College Board experts to complete after the workshop.

RANGE ALD DEVELOPMENT

Consideration of the three workshops present opportunities to improve Range ALD creation moving forward. In this section, we examine different aspects of ALD development including group size, group composition, lowest achievement level, and granularity.

Group Size. Except for Smarter Balanced Grade 11, relatively small groups were used to develop the ALDs. Given the difficulty of the task and the creativity required, it is probably not feasible to use large groups to create ALDs. The group size ranged from three to seventeen. For the large groups, they were split into much smaller teams when doing the work (e.g., the Grade 11 Smarter Balanced panels were split into four groups of four or five participants).

Participants. The panelists included K-12 educators, content experts, and college faculty. This provides an appropriate mix of knowledge of students at each grade level and knowledge of

[4] Discipline refers to an overall area, such as History. Subject refers to subtopics within the discipline, such as World History or U.S. History (Plake et al., 2010).

content. When writing Range ALDs, panelists will create detailed statements of achievement along a spectrum of learning and identify those places along an achievement spectrum where they typically see evidence of important progress or mastery. Thus, educators and cognitive scientists can serve an important role in creating Range ALDs because they understand patterns of learning and development. These patterns of student learning are not only important for score interpretation but also for imagining and developing test content that can accurately and meaningfully capture evidence of student achievement.

Lowest Achievement Level. Smarter Balanced and College Board developed ALDs for their lowest achievement levels, but the Governing Board does not develop an ALD for its lowest achievement level. Even when developing Range ALDs, it is important to describe the expected KSAs of students in the lowest performance level. If Range ALDs guide item writing, then we can target our item development to better measure the KSAs of the students in this level. When student-level reports are provided (as is the case with College Board and Smarter Balanced), then it is important to communicate to stakeholders what KSAs that students in the lowest level are able to do. Students in the lowest level are capable of demonstrating KSAs, and test developers should be able to explain the KSAs of students in the lowest level. It is the challenge of the ALD developers to capture what these students may be able to do instead of everything that they cannot do.

Granularity. All three groups purport to use ALDs in item writing; however, the Governing Board does not write their preliminary or subject-matter ALDs at the same grain size as the other two. This is important because the granularity of the ALD should make a difference to item writers. In a follow-up article to the 2010 College Board studies, Hendrickson, Ewing, Kaliski, & Huff (2013) discussed the difficulty of identifying the appropriate grain-size for the

claims and evidence. In a similar vein, it is necessary to address the appropriate granularity for the Range ALDs when using them to guide item writing. For example, *what level of specificity is needed by item writers to develop items targeted to different areas in the achievement continuum?* In addition, *how much specificity can actually be used by an item writer?*

Writing within or across Levels. The Smarter Balanced committees created ALDs by target. In other words, the ALDs for Levels 1, 2, 3, and 4 were created for an assessment target before moving onto the next target. In contrast, the NAEP groups developed the ALDs by level. In this case, the Proficient level would have been written independently and developed for all content strands before moving to the next level. It is not clear that one method is preferable to another. In both cases, it will be important to study the articulation of content to ensure the logical progression of knowledge and skills across the levels.

SECTION V. POTENTIAL USES OF RANGE ALDS

While ALDs can be conceived to be hypotheses about latent proficiency (Hendrickson et al., 2013), they could also be related to learning progressions as descriptions of instructional practice and practitioner consensus (Shepard, Daro, & Stancavage, 2013). Consistent with principled assessment design, realistic and meaningful cognitive models are necessary to link elements of a validity argument, toward coherence (Deane & Song, 2014). In this regard, learning progressions provide a different ─ although equally appropriate ─ lens on student levels of achievement by approaching them from typical curricular and instructional practice. In this section, we examine potential uses for Range ALDs, including their use as learning progressions and for teaching.

RANGE ALDS AND LEARNING PROGRESSIONS

Learning progressions are cognitive models of how students learn—models derived from theoretical and empirical sources to connect processes of curriculum, instruction, and assessment. Called by various names and developed in various subjects, learning progressions have been characterized as descriptions that link instruction and assessment (Shepard et al., 2013; Korbin, Larson, Cromwell & Garza, 2015). Terms used synonymously with or in relation to the concept of learning progressions vary by discipline and subject area, and they include "progress maps, process variables, developmental continua, progressions of developing competence, profile strands, learning trajectories, and learning lines" (Shepard et al., 2013, p.144). In addition, some learning progressions are general while others are more detailed. Their focus on mastery varies from definition at a grade band to definition for a single unit of study. The lack of consensus on the names, granularity, and proper unit of emphasis do not stop experts' hopes that learning progressions could bring educational systems into greater coherence. Recently, experts in curriculum, instruction, and assessment have pointed to learning progressions to drive better alignment and therefore stronger validity arguments for quantitative assessment of student learning (e.g., Pellegrino, 2014).

The process of developing learning progressions could be considered the development of theory-based learning models that can be empirically tested. Corcoran, Mosher and Rogat (2009) identified essential components of learning progressions:

1. Learning targets are the clear end points defined by societal aspirations and careful analysis of the central concepts and themes in a given subject area;
2. Progress variables that identify critical dimensions of understanding and skill that are being developed over time;
3. Levels of achievement that define stages of progress with significant intermediate steps in conceptual and skill development that most children could be expected to move through on a pass toward a specified level of proficiency;

4. Learning performances as evidence of what students know and can do as examples of what each stage of progress looks like, from which specifications for assessment and related activities are developed;
5. Assessments that measure these student understandings of key concepts and practices to track growth or progress over time (pp. 9-10).

Their characterization of learning progressions supports the idea that Range ALDs synthesize features of learning progressions for both test development *and* instructional practice. The development of Range ALDs could be enhanced by using well-developed learning progressions. Levels of achievement in a learning progression focus on the qualitatively different levels of mastery and understanding with greater emphasis on the higher-order thinking and valued principles, techniques, generalizations, and methods of the discipline, rather than discrete facts or specific skills (Kane & Bejar, 2014). By using learning progressions to ground Range ALDs, assessments can be more closely associated with instructional programs and curricula which are also designed to the learning progressions.

While learning progressions can inform Range ALD development, Range ALD development could also inform learning progressions. The process of seeking the meaningful measurement targets for a given achievement level forces a deep examination of the precursor knowledge, skills, abilities, and understandings. Content standards and discussion of minimum proficiency must be considered, forcing ALD developers to contend with alignment between all these pieces of the puzzle. Inconsistencies in the grain size or reasonableness of the progression descriptions can be sussed out and remedied. Thus, learning progressions and Range ALDs could work together to improve overall coherence in the system, drawing "closer attention to the interplay between the statistical and cognitive aspects of assessment than has been customary" (Pellegrino et al., 2001. P. 110).

Once learning progressions and Range ALDs are synced, they become a resource for item writers to develop task models and features that best elicit the student performances most illustrative of mastery.

CHALLENGES

There are many issues to address in order to best develop Range ALDs as learning progressions or vice versa. First, where do we start? Work could focus first on writing the Range ALD and then evaluate it for learning progressions. Alternately, work could start with the learning progressions, and Range ALDs could focus on measurement targets within each progression. Here job task analyses could inform the process. Ultimately, using both lenses may help for initial alignment review as well as confirmatory evaluations.

The challenge of integrating cognitive models into assessment design is multifaceted. First, there are legacy approaches to assessment that reflect various philosophical, societal, and cultural influences. If the decisions that assessment developers make are inconsistent with those that teachers and school professionals make to influence instruction, assessments will likely be out of alignment, and students will see test content that does not relate coherently with their learning as instructed. This disconnect is consequential, as accountability for teachers, schools, and systems depend on accurate data on student achievement and growth (Linn, 2001).

Learning reflects both latent variables but also the opportunities students have to learn. As Shepard and colleagues (2013) described it, “Virtually all researchers studying learning progressions recognize that development is strongly affected by learning opportunities and specific instructional contexts” (p. 151). Both “natural sequences of development and common conventions for the content and delivery of curricula” drive how students learn (Masters & Forster, 1996, as quoted by Shepard et al., 2013, p. 151).

RANGE ALDS AND TEACHING

Range ALDs have potential to influence teaching practice. Whether in the day-to-day event of the classroom, in school- and district-level curriculum coordination, or in the use of test scores for school improvement, Range ALDs provide readily accessible statements of where student performance falls along important learning progressions or in relation to content standards.

Work in the areas of pedagogy and formative assessment points to the need to better understand the work of teaching so that system elements can promote best practice and empower teachers (e.g., Wilson, 2009; Pearson, 2013). Much efficacy is lost in the disconnects between the demands of teachers that do not relate to their classroom practice. For example, formative assessments or formal curricular programs that are imposed on teachers often add to their workload without bringing meaningful improvement to student learning. Efforts like those of Ball, Thames, and Phelps (2008) seek understanding of content knowledge for teaching. Range ALDs must be informed by these data and, subsequently, they can provide bases for professional development.

Range ALDs have the potential to support teaching goals by articulating achievement levels in specific and actionable terms with alignment to the curriculum, instruction, and assessment triangle. Test items developed to learning progressions and aligned to Range ALDs become strong examples for teachers of how evidence can be collected (Shepard et al., 2013). Furthermore, Range ALDs hold promise in the important goal of linking formative assessment and classroom instruction, consistent with Heritage's (2010) enthusiasm for learning progressions. If Range ALDs articulate key elements of the underlying learning progressions in

terms of achievement levels, they become instruments for clarifying what performance looks like at these different levels of mastery.

Range ALDs reflect general consensus about quality learning outcomes affected by teaching. Learning progressions and ALDs should reflect those typical patterns of instruction and effective teaching. Educators appreciate Range ALDs because they show where students are and need to go in relation to the underlying learning progressions. These connections become powerful when teachers can connect their own understandings of teaching and learning with the generalized achievement levels.

For NAEP, Range ALDs could show teachers what it means to be "Proficient" in terms of national expectations, providing a teacher tangible and accessible points of comparison to use for self-evaluation and evaluation of students' learning opportunities in a national context (Shepard et al., 2013). However, if educators are going to use Range ALDs in the ways we suggest here, the ALDs must be validated. With the 1992 NAEP Reading and Math achievement level settings, developers assumed that exemplar items would match the range of anticipated student performance based on verbal descriptions of achievement levels. However, researchers then found inconsistencies between what the ALDs described and actual student performance (Burstein et al., 1996). Linn (1998) showed how this disconnect could arise when empirical evidence of student performance is ignored in test development processes, and how this problem could be eliminated by applying statistical criteria as well as logically matching items to verbal descriptions. Furthermore, Pellegrino and colleagues (2001) demonstrated that, without models of student cognition along with observation and interpretation, incoherence dominates education systems. Range ALDs have the potential to draw on these learnings to build alignment and coherence.

SECTION VI. DISCUSSION

This paper examined the use of an interrelated systems of ALDs for NAEP, with particular attention paid to the use of Range ALDs for test design and development. Specifically, we want to examine how the delineation of ALD type may affect the Governing Board's policies on:

- Using multiple types of ALDs, including ALDs for item writing
- Using "should" versus "can" in ALDs
- Writing descriptors for the lowest achievement level category

ALD FRAMEWORK AND NAEP

The ALD framework creates an interrelated system of ALDs that differentiates ALD type by use. This framework leads to clarity in the intent and purpose of the ALDs being created. The Governing Board already implements different types of ALDs in the test development process; however, the current system of NAEP ALDs (policy, preliminary subject-matter, and final subject-matter) seems to be a one-size-fits-most model where the same ALDs are used for different purposes. For example, the subject-matter ALDs are supposed to guide test development, yet they lack the granularity needed for their intended purposes; instead, the subject-matter ALDs are better suited for reporting to stakeholders the types of knowledge and skills expected in each achievement level.

ALDS AND ITEM WRITING

Principle 2 of the Governing Board's policy on Item Development and Review states that, "The achievement level descriptions for basic, proficient, and advanced performance shall be an important consideration in all phases of NAEP development and review" (National Assessment Governing Board, 2002, p. 4). For over two decades, researchers have made a case for using ALDs in test design and development of NAEP (Pellegrino et al., 1999; Mills & Jaeger,

1998). At this point, it is unclear if the current subject-matter ALDs are actually utilized in this way. Given the coarseness of the current subject-matter ALDs, it is doubtful that they would be helpful to the test development process.

If the Governing Board chooses to create Range ALDs to guide the test development process, then there is the need for additional research in this area. In theory, it makes sense that the test design process would benefit from careful consideration of the types of knowledge and skills expected in each area of the achievement scale. This approach should create additional coherence in the assessment system. At the same time, there is not very much research to support this theory. The existing research suggests that item writers have difficulty assigning or writing items to ALDs (Ferrara et al., 2011; Schneider et al., 2013). In addition, current research points to the need for processes in which ALDs, items, and test blueprints are developed iteratively and driven by Range ALDs (Forte, 2017; Schneider et al., 2013; Kaliski et al., 2011).

ALDs and Reporting

It is a bit odd to discuss ALDs in terms of reporting when NAEP does not provide individual student reports, which is typically the primary use for Reporting ALDs. Even so, the current NAEP subject-matter ALDs are probably best utilized for communication with stakeholders. The subject-matter ALDs are aggregate ALDs that describe expected student knowledge and skills at the strand level. Following Guideline 3 of the Governing Board's policy on achievement levels, the subject-matter ALDs are "articulated in terms of what students should know and should be able to do" (1995, p. 8) and they are not written for content below the *Basic* level" (1995, p. 8).

Whenever ALDs are created prior to the determination of cut scores, they are necessarily written as *should* statements (Egan et al., 2012). These ALDs reflect our expectations for student

performance rather than actual student performance. Bourque (2009) points out that when ALDs are developed using evidence, it may not make sense to use *should* statements. Bourque makes this point for ALDs being created with ECD. The same point can be made for Reporting ALDs that are developed after cut scores are set. When the Reporting ALDs are based on items over which students have demonstrated mastery, those ALDs may be written as *can* statements.

Within this area, it is important to reconsider the lowest performance level—Below Basic. The Governing Board has taken the stance that nothing should be written for the below Basic category. This is understandable because there is no lower bound for this achievement level making it nearly impossible to make a statement that reflects what all students can do in this category. In addition, this area of the scale is notoriously unreliable. At the same time, the Governing Board should consider that the performance of a sizable chunk of the population is classified in this category.

Even if the Governing Board chooses to not create Reporting ALDs for the lowest category, they should consider creating Range ALDs for this category. If Range ALDs are viewed as learning progressions, then information on the types of learning that are expected of those in the Below Basic group will be useful to states or districts.

ALDs and The Validity Argument

The interrelated system of ALDs provides an opportunity to create and evaluate key sources of documentation that will be central to the validity argument. Range ALDs serve as a primary source of evidence in a validity argument, grounding test development in the intended uses and interpretations (Kane, 2006) as well as in research-based models of student learning and cognition (Mislevy & Haertel, 2006; Pellegrino et al., 2001). Range ALDs become a vehicle for pulling together theoretical design elements with empirical understandings of how students learn.

While ALDs by their definition specify the way to interpret test scores, Range ALDs go further to articulate the intended progression of detailed knowledge across the test scale with the specificity that can connect to classroom instruction more directly than Policy or Reporting ALDs.

To be succinct, the Range ALDs provide the expectations for student performance while the Reporting ALDs summarize actual student performance. This means that the Reporting ALDs can be used to validate the Range ALDs. The real concern becomes when the knowledge and skills of the Reporting ALDs do not align with the Range ALDs. What does this mean for the Range ALDs?

The very real possibility that Range ALDs do not align with the Reporting ALDs means that the Range ALDs should be monitored against the results of NAEP. Unlike other summative assessments, NAEP assesses many items over the course of an administration. Actual student performance can be compared to the expectations of the Range ALDs. In those areas where the theory and reality do not coincide, it will be important to study why this occurs. Should a modification be made to the Range ALDs, or should adjustments be made to the item writing process? At this point, clear answers do not exist to these questions.

CONCLUSION

In sum, Range ALDs pull together test design, development, and score use and interpretation. They are by their nature the nexus of (a) content standards which articulate educational aspirations, (b) test design which samples content and defines task models, (c) development of items to elicit student performance, and (d) score interpretations. Given this, Range ALDs have the potential to serve as a common denominator across parts and players in educational systems. A teacher can seek Range ALDs to envision how to approach content

standards and anticipate student performance within a given unit or across a school year. A test developer can use the same Range ALDs to build an assessment blueprint for a benchmark assessment and an aligned summative assessment. A group of district-level curriculum designers could use the same Range ALDs to track learning progressions and build formative assessment programs. Finally, decision makers at state and federal levels can look to the process of creating Range ALDs as a way to engage experts and stakeholders.

REFERENCES

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special?. *Journal of Teacher Education, 59*, 389–407.

Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. Lissitz (Ed.), *Assessing and modeling cognitive development in school* (pp. 31–63). Maple Grove, MN: JAM Press.

Bourque, M. L. (2009). *A history of NAEP achievement levels: Issues, implementation, and impact 1989–2009.* Retrieved February 2, 2010, from http://www.nagb.org/who-we-are/20-anniversary/bourque-achievement-levels-formatted.pdf

Burstein, L., Koretz, D., Linn, R., Sugrue, B., Novak, J., Baker, E. L., & Harris, E. L. (1996). Describing performance standards: Validity of the 1992 National Assessment of Educational Progress PLDs as characterizations of mathematics performance. *Educational Assessment, 3*(1), 9–51.

Burt, W. M., & Stapleton, L. M. (2010). Connotative meanings of student performance labels used in standard setting. *Educational Measurement: Issues and Practice, 29*(4), 28-38.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform*. New York: Columbia University, Teachers College: Center on Continuous Instructional Improvement, Consortium for Policy Research in Education.

CTB/McGraw-Hill (2013). *Smarter Balanced Assessment Consortium: technical report initial achievement level descriptors.* Author.

Deane, P., & Song, Y. (2014). A case study in principled assessment design: Designing assessments to measure and support the development of argumentative reading and writing skills. *Psicología Educativa, 20*, 99–108.

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: history, practice, and a proposed framework (pp. 79-106). New York, NY: Routledge.

Embretson, S. E., & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*(4), 343–368.

Ferrara, S., Sventina, D., Skucha, S., & Davidson, A. (2011). Test development with performance standards and achievement growth in mind. *Educational Measurement: Issues and Practice 30*(4), 3–15.

Fitzpatrick, S. & Hickey, M. (2016, May). *Developing achievement levels on the 2014 National Assessment of Educational Progress in grade 8 Technology and Engineering Literacy*. Iowa City, IA: Pearson. Retrieved from: https://www.nagb.gov/content/nagb/assets/documents/publications/achievement/developing-achievement-2014-naep-grade-8-tel-technical-report.pdf

Forte, E. (2017). *Evaluating alignment in large-scale standards-based assessment systems* (White Paper for the Technical Issues in Large Scale Assessment State Collaborative on Assessments and Student Standards of the Council of Chief State School Officers). Retrieved from: http://edcount.com/images/PDFs/ccsso_tilsa_forte_evaluating_alignment_2017

Forte, E., Towles, E., Greninger, E., Buchanan, E., & Deters, L. (2017). *Evaluation of the alignment quality in the Georgia Milestones Assessment System in ELA, mathematics, science, and social studies* (Technical Report for the Georgia Department of Education).

Retrieved from https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/Milestones/Georgia_Milestones_Alignment_Evaluation_Executive_Summary.pdf

Haertel E. H., & Lorie, W. A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives*, *2*, 61-103.

Hambleton, R. J. & Jirka, S. J. (2006). Anchor-based methods for judgmentally estimating item statistics. In S. E. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 399-420). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hansche, L. N. (Ed.). (1998). *Meeting the requirements of Title I: Handbook for the development of performance standards*. Washington, DC: U.S. Department of Education.

Hendrickson, A., Huff, K., & Luecht, R. (2010). Claims, evidence, and achievement-level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education 23*(4). 358-377.

Hendrickson, A., Ewing, M., Kaliski, P., & Huff, K. (2013). Evidence-centered design: Recommendations for implementation and practice. *Journal of Applied Testing Technology, 14*. Retrieved from https://atpu.memberclicks.net/assets/documents/evidence-centered%20design%20jatt%20special%20issue%2013.pdf

Heritage. M. H. (2010). *Formative assessment: Making it happen in the classroom*. Thousand Oaks, CA: Corwin.

Kaliski, P., Huff, K. L, and Barry, C. (2011). *Aligning items and achievement levels: A study comparing expert judgments.* Retrieved from: https://files.eric.ed.gov/fulltext/ED563464.pdf

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425-461.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th Ed.) (pp. 17-64). Westport, CT: Preaeger Publishers.

Kane, M. T., & Bejar, I. I. (2014). Cognitive frameworks for assessment, teaching, and learning: A validity perspective. *Psicología Educativa, 20*, 117–123.

Korbin, J. L., Larson, S., Cromwell, A., & Garza, P. (2015). A framework for evaluating learning progressions on features related to their intended uses. *Journal of Educational Research and Practice, 51*(1), 58–73.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, M. (2012). The Bookmark standard setting procedure. In G. Cizek *Setting Performance Standards* (pp. 225-282). New York, NY: Routledge.

Linn, R. L. (1998). Validating inferences from National Assessment of Educational Progress achievement-level reporting. *Applied Measurement in Education, (11)*1, 23–47.

Linn, R. L. (2001, April). *The design and evaluation of educational and accountability systems.* [CSE Technical Report 539]. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Mills, C. N., & Jaeger, R. M. (1998). Creating descriptions of desired student achievement when setting performance standards. In L. N. Hansche, *Meeting the requirements of Title I: Handbook for the development of performance standards*. Washington, DC: U.S. Department of Education.

Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 25*, 6–20.

National Assessment Governing Board (1995). *Developing student performance levels for the National Assessment of Educational Progress. Policy statement.* Retrieved from: https://www.nagb.gov/content/nagb/assets/documents/policies/developing-student-performance.pdf

National Assessment Governing Board (2002). *Item Development and review. Policy statement.* Retrieved from: https://www.nagb.gov/content/nagb/assets/documents/policies/Item%20Development%20and%20Review.pdf

National Assessment Governing Board (2013). *Reading Framework for the 2013 National Assessment of Educational Progress.* Author.

Pearson, P. D. (2013). Research foundations of the Common Core State Standards in English language arts. In S. Neuman and L. Gambrell (Eds.), *Quality reading instruction in the age of Common Core State Standards* (pp. 237–262). Newark, DE: International Reading Association.

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa, 20*. 65–77.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). *Grading the Nation's Report Card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Research Council/National Academy Press.

Perie, M. (2008). A guide to understanding and developing PLDs. *Educational Measurement: Issues and Practice, 27*(4), 15-29.

Plake, B. S., Huff, K. & Reshetar, R. (2010). Evidence-centered assessment design as a foundation for achievement-level descriptor development and for standard setting. *Applied Measurement in Education, 23*(4), 342-357.

Schneider, M. C., Egan, K. L., Kim, D., & Brandstrom, A. (2008, March). *Stability of achievement level descriptors across time and equating methods.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Schneider, M. C., Huff, K. L., Egan, K. L., Gaines, M. L., & Ferrara, S. (2013). Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement-level descriptors, Educational Assessment, 18:2, 99-121

Shepard, L., Daro, P., & Stancavage, F. B. (2013). *The relevance of learning progressions for NAEP* (Research Report for the National Center for Educational Statistics Contract No. ED-04-CO0025/0012). Retrieved from https://files.eric.ed.gov/fulltext/ED545240.pdf

Smarter Balanced. (2015, January). *Smarter Balanced Assessment Consortium: Achievement level setting final report*. Author.

WestEd, (2014). *Finalizing the National Assessment of Educational Progress (NAEP) eighth grade technology and engineering literacy achievement level descriptions. Project documentation.* Author.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of research in science teaching, 46*(6). 716-730.

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

Arthur A. Thacker, Ph.D.
Tonya Longabach, Ph.D.
Human Resources Research Organization (HumRRO)

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

[1] This is an excerpt of Technical Memorandum #1 (HumRRO Report 2017 NO. 089), developed under contract #ED-NAG-17-C-0002, Technical Support in Psychometrics, Assessment Development, and Preparedness for Postsecondary Endeavors.

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[2] 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 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[3] 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.

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

[3] See www.nationsreportcard.gov.

ALDs highlight what examinees need to accomplish to meet performance standards (Hambleton, Pitoniak & Copella, 2012).

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).

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.

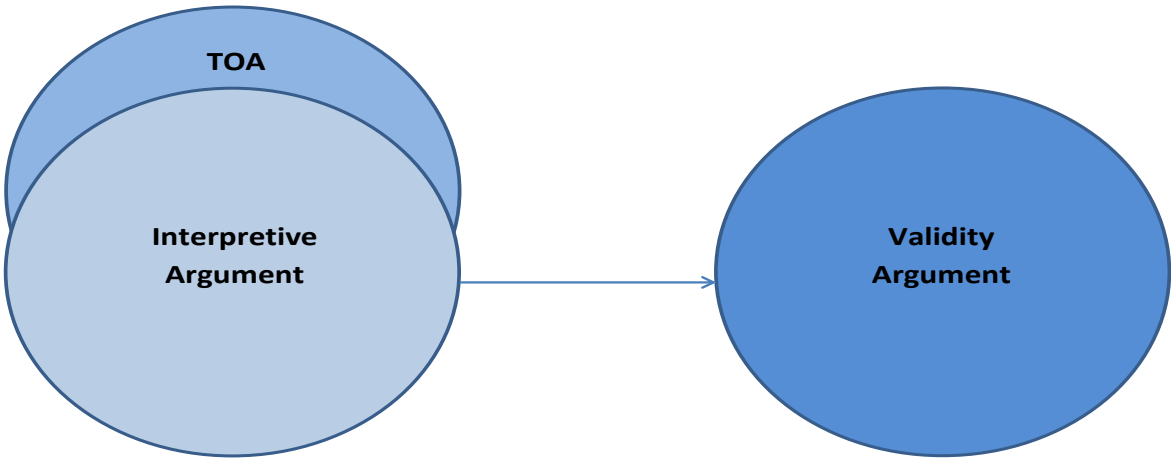


Figure 1. Relationships among theory of action (TOA), interpretive argument, and validity argument.

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 1. 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 does, 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[4] (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 a secondary indicator 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.

[4] 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 are 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 of 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.

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

Purpose: The purpose of this document is to provide an update to the Committee on Standards, Design and Methodology (COSDAM) regarding the development of achievement levels for the 2017 NAEP Grade 4 Writing.

Legend:
Light shading: Completed
No shading: To be completed after 3/1/18

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

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

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

The Governing Board policy on Developing Student Performance Levels for NAEP (https://www.nagb.org/content/nagb/assets/documents/policies/developing-student-performance.pdf) requires appointment of a committee of technical advisors who have expertise in standard setting and psychometrics in general, as well as issues specific to NAEP. These advisors are being convened for 8 in-person meetings and up to 6 webinars to provide advice at every key point in the process. They provide feedback on plans and materials before activities are implemented and review results of the process and analyses. Six external experts in standard setting are serving on the Technical Advisory Committee on Standard Setting (TACSS):

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

Dr. Gregory Cizek [2]
Professor of Educational Measurement, University of North Carolina at Chapel Hill

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

Dr. Steve Ferrara
Independent Consultant

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

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

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

February 2018 Update:

Update on Preparations for the ALS Study

Pearson is currently in the process of finalizing all materials, tools, and logistics necessary to conduct the ALS study. All material and tool revisions were based on lessons learned from the June field trial, the November pilot study, and additional feedback from TACSS. The ALS study will be conducted from February 12-15, 2018 in Atlanta, GA. Thirty-three panelists from around the country were recruited and have committed to participating.

December COSDAM Webinar

On December 14th, COSDAM met briefly by webinar to review and discuss the report from the pilot study. A short presentation highlighted key outcomes from the study, primarily noting that all aspects of the study were conducted as planned. Final round cut score and consequences data from the study were also shared. After the presentation, the floor was open for questions and answers. COSDAM requested that relevant external evidence sources be provided for comparison as well as more contextual information that may lend support to the reasonableness of final results. Both requests will be addressed within delivery of the operational ALS results.

[2] Greg Cizek was appointed to the Governing Board by Secretary Betsy DeVos to serve as one of the three Testing and Measurement experts from October 1, 2017 to September 30, 2021. On October 7th, Dr. Cizek informed Pearson that he was resigning from the TACSS. Given the project timeline and the small number of remaining TACSS meetings, Pearson will not seek a replacement TACSS member.

January TACSS Webinar

On January 18th, the TACSS met briefly by webinar to review and discuss revised materials and procedures for the ALS. Most of the call was dedicated to review of the ALS presentations that had been revised based on lessons learned from the pilot study. Specifically, more context was provided around the use of scoring rubrics, the NAEP reporting scale, and how these are used within the ALS. Modifications to the remaining ALS materials, to include the agenda and facilitator guide, were presented and TACSS members given the chance to address any concerns with each. Overall TACSS felt all revisions were reasonable.

Next Steps

Given the proximity of the March 2018 COSDAM session to the end of the ALS study, a separate webinar will be planned for mid-March to provide a briefing on outcomes. At that webinar, Writing ALS Project Director Tim O'Neil will present results from the ALS study.

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*	COSDAM discussions at May and August 2017 board meetings to examine how existing findings may be used to add meaning to scale scores and achievement levels, and what additional studies to take on 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 will be provided in the March 2018 COSDAM materials	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 Technical Support contract specifies that the research study topic for year 1 will focus on how NAEP results are used by various stakeholders. Initial ideas for approaching this research study were shared with COSDAM during the November 2017 meeting. The study is currently underway.	Use research to draft short document of intended and appropriate uses for Board discussion (November 2018) 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*	Panel of standard setting experts convened in January 2018 to discuss technical issues and recommendations for achievement levels policy Literature review on considerations for creating and updating achievement level descriptors (ALDs) Technical Memo on developing a validity argument for the NAEP achievement levels (February 2018) The efforts described above will be discussed at this (March 2018) COSDAM meeting to inform policy revision	COSDAM discussion of draft policy statement and supporting materials (May 2018) Revised policy statement for full Board discussion (August 2018) Seek external feedback and public comment (September 2018) Board action on revised policy statement (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
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)	Ongoing board discussion about options for the future of LTT and what additional information may be needed	Determine whether changes to the NAEP LTT schedule, design and administration are needed (led by Executive Committee and NCES)

Strategic Vision Activity	Current Status and Recent Work	Planned Next Steps	Desired Outcome
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 (November 2017) Full Board presentation and discussion on efficiencies in what and how to measure student knowledge and skills (March 2018)	TBD, pending March 2018 discussion	Determine whether new assessment schedule should reflect the concept of "combined assessments" (led by Executive Committee)
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 Decide whether Board should make stronger statement and/or set "benchmarks" rather than current approach of "plausible estimates" Decide whether additional research should be conducted with more recent administrations of NAEP and other tests	Statements about using NAEP as an indicator of academic preparedness for college continue to be defensible and to have appropriate validity evidence

Summary of Ongoing NAEP Linking Studies (SV #2)

The Governing Board's Strategic Vision includes a goal to, "Increase opportunities to connect NAEP to administrative data and state, national, and international student assessments." Both the Governing Board and the National Center for Education Statistics (NCES) are conducting a variety of studies that link NAEP to other assessments and data sources. Linking studies can provide useful information relevant to educational policy by establishing new relationships between NAEP results and other assessments, contextual information, and/or outcome variables that NAEP does not routinely collect.

COSDAM has had several conversations about NAEP linking studies, most recently at the August 2017 Board meeting. The purpose of this informational update is to provide an overview of the linking studies that are currently underway.

As part of its research program on using NAEP as an indicator of academic preparedness for college, the Governing Board (in partnership with NCES and their contractors ETS and Westat) conducted several NAEP linking studies with NAEP Reading and Mathematics assessments at grades 8 and 12. Three efforts related to this work remain underway: a national linking study of 2013 NAEP Reading and Mathematics assessments at grade 12 and ACT Reading and Mathematics scores for students in the NAEP sample; longitudinal studies in Massachusetts, Minnesota, and Tennessee for students who took the 2013 NAEP Reading and Mathematics assessments at grade 12; and longitudinal studies in North Carolina and Tennessee for students who took the 2013 NAEP Reading and Mathematics assessments at grade 8.

NCES is conducting several research studies that link 2013, 2015, 2017, and 2018 NAEP assessments to other NCES data collections (i.e., High School Longitudinal Study in Mathematics; Early Childhood Longitudinal Study – Kindergarten in Reading); international assessments (Trends in International Mathematics and Science Study; International Computer and Information Literacy Study); and selected state assessments.

A brief overview of each study is provided, including: the purpose of the assessment or survey that NAEP is being linked to; what information is expected to be gained from the linkage; whether the linkage is concurrent (i.e., relating NAEP to another outcome that takes place within the same time frame) or predictive (i.e., relating NAEP to a future outcome); which audiences are likely to value the linkage; the current status of the study; and the expected completion date.

NAEP Linking Studies Led by the Governing Board

ACT Reading and Mathematics Scores Linked to 2013 NAEP Grade 12 Reading and Mathematics

Purpose of the survey/assessment that NAEP is being linked to: The ACT assessment is a college admissions test used by colleges and universities to determine the level of knowledge and skills in applicant pools, including Reading, English, Mathematics, and Science tests. The 2013 ACT has *College Readiness Standards* that connect reading or mathematics knowledge and skills and probabilities of a college course grade of "C" or higher (0.75) or "B" or higher (0.50) with particular score ranges on the ACT assessment. In partnership with ACT and using a procedure that protects student confidentiality, the ACT records were matched to the national sample of students who took the 2013 NAEP Reading and Mathematics assessments at grade 12.

Information to be gained from the linkage: The purpose of this study is to identify the NAEP Reading and Mathematics scores that are equivalent to the ACT Reading and Math College Readiness Benchmarks, as a way to inform interpretations of NAEP results, along with other information from other studies. The national linking study will specifically be looking at: the correlations between the NAEP and ACT scores in reading and mathematics; the reading and mathematics NAEP scores that correspond to the ACT benchmarks; descriptive statistics for NAEP reading and mathematics scores for students below, and at or above the ACT benchmarks; the reading and mathematics ACT scores that correspond to the NAEP Proficient cut scores; and whether there are differences by gender or race/ethnicity.

Linkage concurrent or predictive: Concurrent

Which audiences will value this linkage: Researchers and policy makers interested in the relationship between grade 12 NAEP and other established indicators of academic preparedness for entry-level, general education coursework.

Current status: The matching process is now complete and data analysis is currently underway.

Expected completion: A draft report is expected to be shared with COSDAM during the August 2018 Board meeting.

Longitudinal Statistical Relationships Linked to 2013 NAEP Grade 12 Reading and Mathematics

Purpose of the data that NAEP is being linked to: Postsecondary activities of 2013 NAEP 12th grade test takers will be followed for up to six years using the state longitudinal databases in Massachusetts, Michigan, and Tennessee (via data sharing agreements with these states, and using a procedure that protects student confidentiality).

Information to be gained from the linkage: These studies will examine the relationships between 2013 NAEP Grade 12 Reading and Mathematics scores and: scores on placement tests; placement into remedial versus credit-bearing courses; first-year grade point average; persistence (remaining in college after each year); and graduation within 6 years.

Linkage concurrent or predictive: Predictive

Which audiences will value this linkage: Researchers and policy makers interested in the predictive validity of grade 12 NAEP with respect to placement, performance, and graduation from college.

Current status: Some longitudinal data have been received and analyses are currently underway.

Expected completion: Initial results linking NAEP to ACT and SAT scores from these partner states were shared with COSDAM during the August 2016 Board meeting. The next phase of results is expected to be available in time for the August 2018 COSDAM meeting.

Longitudinal Statistical Relationships Linked to 2013 NAEP Grade 8 Reading and Mathematics

Purpose of the data that NAEP is being linked to: Secondary and postsecondary activities of 2013 NAEP 8th grade test takers will be followed for up to five years using the state longitudinal databases in North Carolina and Tennessee (via data sharing agreements with these states and using a procedure that protects student confidentiality).

Information to be gained from the linkage: These studies will examine the relationships between 2013 NAEP Grade 8 Reading and Mathematics scores and: ACT scores in grades 11 and 12; placement into remedial versus credit-bearing courses during the first year of college; and first-year college grade point average.

Linkage concurrent or predictive: Predictive

Which audiences will value this linkage: Researchers and policy makers interested in the predictive validity of grade 8 NAEP with respect to college placement and performance.

Current status: Longitudinal data for the grade 8 statistical relationship studies have not yet been received. It is necessary to establish a new data sharing agreement for the next phase of research with North Carolina.

Expected completion: Initial results linking NAEP to ACT EXPLORE scores from these two partner states[1] were shared with COSDAM during the August 2015 Board meeting. The timing for the next phase of results is not yet known.

[1] Kentucky also participated in the NAEP and EXPLORE linking studies but opted not to participate in the longitudinal component of the research.

NAEP Linking Studies Led by NCES

High School Longitudinal Study (HSLS) Linked to 2013 NAEP Grade 12 Math

Purpose of the survey/assessment that NAEP is being linked to: The HSLS is a longitudinal study that follows a nationally representative sample of students who entered grade 9 in fall 2009. When these students were in grade 12 during the 2012-13 school year, a subset of them (an overlap sample of 3,500) took the 2013 NAEP Grade 12 Mathematics assessment. Grade 12 NAEP mathematics scale scores were linked with Algebra assessments at grades 9 and 11 and other important background characteristics. In addition to administering mathematics assessments in grades 9 and 11, HSLS also collects information from students' high school transcripts and will follow them in postsecondary education and after postsecondary education.

Information to be gained from the linkage: 1) Validity of NAEP grade 12 Math in predicting college entrance, college grades, and initial post-graduate employment; 2) improved understanding of the role of motivation and course taking in math performance; 3) improved measures of socio-economic status (SES). The following validity studies are underway:

1. Imputing 12th-Grade NAEP Mathematics Scores for the Full HSLS Sample
2. Mathematics Motivation and its Relationship with Mathematics Performance: Evidence from the NAEP-HSLS overlap sample
3. Examining the ability of Grade 12 NAEP mathematics to predict college acceptance/enrollment
4. Examining STEM Course taking in high school in the prediction of Grade 12 NAEP Mathematics scores.
5. Investigating SES Using the NAEP-HSLS Overlap Sample

Linkage concurrent or predictive: Predictive

Which audiences will value this linkage: Researchers and policy makers interested in the predictive validity of NAEP with respect to college entrance, college grades, and employment outcome variables.

Current status: NAEP Mathematics scores have been related to college admission and high school transcript data, and college transcript data are being prepared for analysis.

Expected delivery of reports: Studies are expected to be completed during 2018, with the exception of study 4 (to be determined).

The Early Childhood Longitudinal Study – Kindergarten (ECLS-K) Linked to 2015 NAEP Grade 4 Reading

Purpose of the survey/assessment that NAEP is being linked to: The ECLS-K is a longitudinal study that follows a nationally representative sample of students who entered kindergarten during the 2010-11 school year through grade 5. When these students were in grade 4 during the 2014-15 school year, a subset of them took the 2015 NAEP Grade 4 Reading assessment. Broad in its scope and coverage of child development, early learning, and school progress, the ECLS-K provides descriptive information on children's status at entry to school, their transition into school, and their progression through the elementary grades. The ECLS-K reading assessment includes items measuring prerequisite skills for comprehension such as alphabetic principles and decoding.

Information to be gained from the linkage: 1) to compare new NAEP socio-economic status (SES) variables with SES variables collected from parents and students in the ECLS-K sample; 2) to identify precursors of low and high achievement in 4th grade reading as measured by NAEP.

Study 1 dealt only with SES-related data and was not a linking study per se.

Study 2 is investigating students' development patterns and skill profiles in reading using the ECLS-K data (both scale scores and item response data); and identifying early indicators and factors associated with student performance on the grade 4 NAEP. In particular, the study focuses on the variables collected from teachers and parents related to reading instruction and reading activities at home to examine the relationships between these variables on students' reading growth trajectories and skill profiles.

Linkage concurrent or predictive: Predictive

Which audiences will value this linkage: The longitudinal nature of the ECLS-K data enables researchers to study how a wide range of family, school, community, and individual factors are associated with academic performance over time as measured by cognitive assessments designed for ECLS-K, and also with NAEP scores in grade 4. This information will be useful to researchers in reading education and education policy makers.

Current status: Projected NAEP scores on grade 4 Reading can be computed for 600 students, which is only half of what was planned. This reduced sample size will limit only the analyses we might have made by race/ethnicity or by National School Lunch Program. If meaningful relationships can be found for Reading with a sample of 600, then we can pursue similar analyses with Mathematics, which also has a sample of 600 students.

Expected delivery of report: July – September, 2018

Trends in International Mathematics and Science Study (TIMSS) Linked to 2015 NAEP Grade 8 Mathematics and Science

Purpose of the survey/assessment that NAEP is being linked to: TIMSS enables participating countries to make evidence-based decisions for improving educational policy. Consumers use TIMSS results to: measure the effectiveness of their educational systems in a global context, identify gaps in learning resources and opportunities, pinpoint areas of weakness and stimulate curriculum reform, and measure the effect of new educational initiatives.

Information to be gained from the linkage: This linking study is a project of the NAEP Validity Studies Panel. Randomly equivalent groups of students were administered NAEP or TIMSS. Because no students were administered both assessments, correlations between the assessments cannot be computed. However, the 2011 NAEP-TIMSS linking study showed that the assessments were highly intercorrelated, and that statistical moderation using the mean and standard deviation equating method yielded the same results as the statistical projection and calibration methods, which are possible only when students took both assessments. NAEP results for the nation and the states can be expressed on the TIMSS scale, making it possible to compare U.S. jurisdictions to international educational systems. In addition, NAEP achievement levels can be compared to TIMSS performance levels. More specifically,

1. For Science grades 4 and 8 we will:
    a. Link 2015 NAEP to 2015 TIMSS in the U.S. national samples and compare NAEP standards to TIMSS standards
    b. Predict state TIMSS performance from state NAEP performance

2. For Mathematics grades 4 and 8 we will:
    a. Link 2015 NAEP to 2015 TIMSS in the U.S. national samples and compare NAEP standards to TIMSS standards
    b. Predict state and TUDA TIMSS performance from state NAEP performance
    c. Estimate the international TIMSS-equivalents of local state standards (including PARCC, SBAC and ACT)

Linkage concurrent or predictive: Concurrent

Which audiences will value this linkage: Stakeholders include U.S. Department of Education, state departments of education, researchers, and policy analysts in universities and think tanks.

Current status: The final report is being revised based on reviews by the NAEP Validity Research Panel in January, 2018.

Expected delivery of report: July - September, 2018

State Assessments Linked to 2017 NAEP Grades 4 and 8 Reading and Mathematics

Purpose of the survey/assessment that NAEP is being linked to: State assessments such as the Partnership for Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC) are used for accountability and thus represent the state's learning goals.

Information to be gained from the linkage: The NVS study comparing NAEP and a sample of state assessments (PARCC, SBAC, and two non-consortia states) is intended to inform the on-going development and reporting of NAEP by providing information on the similarities and differences between the NAEP mathematics, reading and writing assessments at grades 4 and 8 and the current generation of states' mathematics and ELA learning goals, as reflected in states' accountability assessments. The study includes an item comparison component and a statistical component.

The inspiration for the statistical component is to inform the item comparison analyses with regard to the relative difficulty (location on the NAEP scale) of the cognitively complex items on the college and career readiness aligned assessments. By extension, this provides a measure of how NAEP items and state assessment items with the same level of cognitive complexity perform relative to one another. To this end, performance data from students who participated in both their state assessment and NAEP will be used to link and then jointly scale each separate state assessment with the corresponding NAEP assessment.[2]

Linkage concurrent or predictive: Concurrent

Which audiences will value this linkage: Stakeholders include U.S. Department of Education, state departments of education, and various researchers and policy analysts in universities and think tanks.

Expected delivery of report: October – December, 2018

[2] However, SBAC has not agreed to participate in the statistical sub-study.

The International Computer and Information Literacy Study (ICILS) Linked to 2018 NAEP Grade 8 Technology and Engineering Literacy (TEL)

Purpose of the survey/assessment that NAEP is being linked to: The International Computer and Information Literacy study (ICILS) is the first international comparative study that assesses the extent to which students know about, understand, and are able to use information and communication technology (ICT).

The main purpose of ICILS is to determine how well students are prepared for study, work and life in the information age, and how their performance compares with students in other participating countries. In total, 21 countries will participate in the 2018 cycle of ICILS.

The assessment measures Computer and Information Literacy levels and computer use of 8th grade students. This includes students' ability to use computers to investigate, create, and communicate. It also collects information about the contexts in which students develop computer and information literacy within and outside of school.

NAEP TEL is a computer-based assessment using interactive scenario-based tasks in addition to traditional discrete items to measure whether students are able to apply technology and engineering skills to real-life situations. The content of TEL appears to be related to the content of ICILS, but distinct in other respects.

Information to be gained from the linkage: If the contents of the two assessments are sufficiently similar, it might be reasonable to use statistical moderation to obtain projected TEL scores for international education systems. Thus, the U.S. would be able to compare their students' performance on TEL to students globally.

Linkage concurrent or predictive: Concurrent

Which audiences will value this linkage: Policy makers and education officials who are interested in both technology and engineering literacy and the more specific domain of information and communication technology (ICT) will find this study interesting. The content analysis study that precedes the actual linking study will provide knowledge of similarities and differences among the assessments, which will be useful in interpreting the results.

Current status: Data collection will be concluded shortly. TEL is being administered in January-March of 2018, and ICILS in February-April of 2018.

Expected delivery of report: Results of both assessments are expected to be ready for release in 2019; and the linking study could be reported as soon as six months later.