National Assessment Governing Board
Committee on Standards, Design and Methodology

Friday, May 19, 2017
10:30 am – 12:45 pm

AGENDA

10:30 – 10:35 am	Welcome and Review of Agenda *Andrew Ho, COSDAM Chair*	
10:35 – 11:20 am	Revision of Board Policy on Achievement Levels *(SV #5)* *Sharyn Rosenberg, Assistant Director for Psychometrics*	Attachment A
11:20 am – 12:20 pm	Discussion and Next Steps for NAEP Linking Studies *(SV #2)* *Sharyn Rosenberg, Assistant Director for Psychometrics* *Bill Tirre, NCES*	Attachment B
12:20 – 12:35 pm	Working Group on Framework Update Processes *(SV #5)* *Andrew Ho, COSDAM Chair*	Attachment C
12:35 – 12:45 pm	Information Items • 2017 Writing Grade 4 Achievement Levels Setting Project Update • Procurement Update • Next Steps for Implementing Strategic Vision	Attachment D Attachment E Attachment F

Revision of Board Policy on Achievement Levels

Background

During the March 2017 board meeting, COSDAM members discussed the need to revise the 1995 board policy on Developing Student Performance Levels for NAEP (attached). The Governing Board's formal response to the November 2016 evaluation of the NAEP achievement levels 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 still 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 the following seven standard setting experts in March/April 2017:

Dr. Gregory Cizek, Professor of Educational Measurement, University of North Carolina at Chapel Hill
Dr. Edward Haertel, Professor of Education, Emeritus, Stanford University
Dr. Kristen Huff, Vice President, Assessment and Research, Curriculum Associates
Dr. Andrew Kolstad, Independent Consultant (Former Senior Technical Advisor, National Center for Education Statistics)
Dr. Susan Loomis, Independent Consultant (Former Assistant Director for Psychometrics, National Assessment Governing Board)
Dr. Marianne Perie, Director, Center for Assessment and Accountability Research and Design
Dr. Mary Pitoniak, Senior Strategic Advisor for Statistical Analysis, Data Analysis, and Psychometric Research, Educational Testing Service

The participants all have some experience with NAEP achievement levels setting in particular but there was wide variation in their specific roles and degree of involvement. Experts were asked to review the policy and then participate in a short phone call (individually) with Dr. Rosenberg. Each participant was asked, "Given that the Board will be undertaking a revision of this policy, what aspects of the policy should be revisited? Are there elements of the policy that are outdated or not in alignment with current best practices in standard setting?" The phone calls lasted between 30-60 minutes. In addition, two experts sent written edits in "track changes" just prior to the phone conversation; comments were further clarified during the phone calls.

Key Takeaways from Expert Conversations

The expert conversations affirmed that the policy contains a lot of information that is still useful and relevant but also identified several areas in need of updates and reconsiderations. The main takeaways from these conversations are presented below (this is not an exhaustive list of the very rich and detailed feedback and suggested edits received, which will be used to inform subsequent phases of the policy revision process):

- 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; "national consensus approach" should be "consultative approach")
- 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

Proposed Next Steps

The expert conversations identified several editorial and substantive considerations, some of which are fairly straightforward (e.g., updating references, elaborating on certain aspects of procedures) and others which could benefit from additional debate and research evidence (e.g., creating and updating the ALDs, collecting and documenting the validation process). To inform the revision of the policy, the following next steps are proposed:

Proposed Activity	Timeline
Initial COSDAM discussion about scope and process of revising policy	May 2017
Initial full Board discussion about potential elements of policy revision (some issues relate to ADC and R&D)	August 2017
Conduct literature review of best practices for creating and updating the ALDs	November 2017
Convene a technical advisory panel to seek expert advice and debate on major substantive issues – both from the evaluation of NAEP achievement levels and the expert conversations	Late 2017/early 2018
Review of draft policy statement by COSDAM and/or full Board	March 2018
Collect public comment on a draft revised policy via the Governing Board website, technical advisory panel reviews, targeted emails to standard setting experts and users of NAEP data and achievement levels, and at the AERA/NCME annual meetings	April 2018
Review of revised policy statement by full Board	May 2018
Adopt revised policy	August 2018

Discussion Questions

1. Does the proposed approach to updating the Board policy on achievement levels seem reasonable?
2. Do COSDAM members have additional suggestions for substantive aspects of the policy that should be revisited?

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.

Discussion and Next Steps for NAEP Linking Studies

During several previous Governing Board meetings, the Committee on Standards, Design and Methodology (COSDAM) has discussed various studies that were conducted (by both NCES and the Governing Board) to link NAEP to other assessments or data sources. Linking studies involve comparisons between two assessments allowing one to see where a score point on one of the assessments would fall on the scale of the other assessment. In May 2016, Sharyn Rosenberg of the Governing Board staff and William Tirre of NCES used studies conducted during the past ten years to present the primary purposes of NAEP linking studies: to estimate state-level performance on international assessments; to inform the development of a new measure of socio-economic status; to compare state performance standards on a common scale; to compare NAEP achievement levels with external benchmarks; and to estimate the percentage of students academically prepared for college.

One of the recommendations of the recent evaluation of NAEP achievement levels (National Academies of Sciences, Engineering, and Medicine; November 2016) was to conduct additional research on the relationships between the NAEP achievement levels and concurrent or future performance on measures external to NAEP. The Strategic Vision (Inform #2) also includes a goal to increase opportunities for linking NAEP to other assessments and datasets. The purpose of these recommendations is to add interpretability and usefulness to the NAEP achievement levels and scale scores by connecting them to other familiar and meaningful indicators of performance.

In this session, Sharyn Rosenberg and William Tirre will provide a brief overview and examples of how findings from existing linking studies can be used to represent external indicators of performance in terms of NAEP scale scores and achievement levels.

Discussion Questions:

1. Is this a useful way of summarizing findings from existing linking studies?
2. What additional linking studies should be pursued?

Working Group on Framework Update Processes

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

INTEREST IN NEW FRAMEWORK UPDATE PROCESSES

A priori process decisions can potentially support a more continuous, incremental, and systematic model for framework updates, aligning with the Board's Strategic Vision to develop new framework update approaches that address evolving expectations for students and rigorous continued reporting of student achievement trends. A major contribution of a new approach could be to proactively consider how to preserve the student achievement trends reported by NAEP, while ensuring NAEP frameworks remain relevant.

There are several issues to resolve before the Board can determine feasibility of a new approach. For instance, determining what content updates are needed is in the purview of the Assessment Development Committee (ADC), while determining speed of changes and methods for maintaining trend with continuous, incremental changes to content are issues for the Committee on Standards, Design and Methodology (COSDAM) and the National Center for Education Statistics (NCES). Perspectives from both committees will assist the Board in determining what a new framework updating approach might entail.

ISSUES RAISED IN NOVEMBER 2016 AND MARCH 2017 JOINT COMMITTEE DISCUSSIONS

Adopting a new process involves many nuances, such as how items released at each assessment affect the intended incremental progression of framework content updates and how achievement level descriptors will account for these updates. Joint committee discussions have also noted that having a stable measure does not ensure stability in what is being measured—if NAEP continued to assess writing in the traditional paper-pencil format, measurement would be compromised since increasingly students do not write this way. Other issues raised in joint committee and working group discussions include:

- Considering whether NAEP is not detecting changes that are important to capture.
- Addressing new sequencing of content across grades.
- Avoiding the portrayal of a moving target with an assessment that is constantly changing.
- Considering how changes interact with general content drift over time or the accumulation of year-to-year trend inferences over time.
- Leveraging digital platforms for student engagement in NAEP content and the platform.
- Engaging stakeholders in determining needed updates.
- Considering NAEP's leadership for the nation and states.
- Confirming whether NAEP should help spur progress in education, while documenting what students know and can do, since this additional focus could suggest different framework update processes and timelines.
- Shortening lead times.
- Exploring whether context shifts of items alone can represent desired changes.
- Determining how much change is too much and the ideal rate of change.

ADC and COSDAM have jointly discussed how the Board can minimize the risk of having a framework become irrelevant, even though it is inherently difficult to predict certain vulnerabilities far in advance. Change every other year has been acknowledged as extreme, but the current pace of change is likely slower than what is needed.

PROSPECTIVE OUTCOMES FOR POLICY AND PROCESS

Current Board policy prioritizes having NAEP frameworks remain stable for at least 10 years, and does not include processes for updating NAEP frameworks more frequently, while still maintaining trend. Hence, joint- and cross-committee discussions will review and refine policies and procedures for updating frameworks, which may include:

- Criteria to determine whether there is a compelling rationale to pursue content updates.
- Criteria to determine whether a new approach for updating a framework is appropriate.
- A suggestion to pilot the new approach in a particular NAEP subject.

PREVIOUS MODELS FOR FRAMEWORK UPDATES

Previously, the Board has pursued framework updates in three ways:

1. ***New Framework with New Trend***
   Research, outreach, content, and policy input show a new framework is warranted to define a new construct, including new content, skills, item types, delivery modes (i.e., digital-based assessment (DBA)), and other modifications. The new construct definition motivates a break in trend reporting from the old assessment's results. Examples:
   - *2011 NAEP Writing*—writing with word processing tools represented a different construct compared with the previous framework's paper-pencil assessment.
   - *2009 NAEP Science*—advancements in science and science curricular standards warranted a different construct with crosscutting content and deeper integration of science practices.
2. ***New Framework with Maintained Trend***
   A new framework is designed to be different from the previous framework. However, empirical investigation reveals that the construct does not differ substantially. Interest in maintaining trend reporting prompts research to try to ensure trend lines can be maintained. Example:
   - *2009 NAEP Reading*—several sub-elements of the previous framework were no longer relevant to the field's conceptualization of reading comprehension, prompting a new framework as in NAEP Writing and NAEP Science. Reauthorization of the Elementary and Secondary Education Act in 2002 required use of NAEP as a monitoring tool for states, prompting interest in maintaining reading trend despite construct changes. Empirical investigation revealed trend could be maintained from 1992.
3. ***Updated Framework/Maintain Trend***
   Making gradual changes to a framework over time may help ensure that trend is maintained. Framework "tweaks" are prompted by important and less dramatic curricular and assessment advances. So these changes are sporadic, rather than ongoing. Examples:
   - *NAEP Mathematics*—over time "tweaks" clarified objectives, shifted content emphases, and refined the process dimension, while the construct definition was unchanged, enabling NAEP to maintain the mathematics trend line for grades 4 and 8 since 1990.
   - *2006 NAEP U.S. History*—clarifications suggested by the NAEP U.S. History test specifications and removal of outdated material were "tweaks" to refresh the framework without disrupting trend.

WORKING GROUP ON FRAMEWORK UPDATE PROCESSES

At the March 2017 joint meeting of ADC and COSDAM, there was unanimous agreement that a working group should be established, to develop a proposal for new approaches to updating frameworks for the entire Board's consideration. The primary goal of the working group is:

> *To explore a systematic process for conducting a series of smaller, more incremental changes to frameworks on a faster schedule in a way that enables maintenance of trends.*

This smaller group would participate in monthly calls to support steady progress in Board deliberations, identifying issues to bring back to the respective committees iteratively for feedback. As noted above, ADC will have primary responsibility for identifying what content should be updated to ensure that assessments remain relevant, while COSDAM will explore the extent to which content changes can be made while maintaining trends.

The following Board members volunteered to participate in the working group: Lucille Davy, Shannon Garrison, Andrew Ho, Dale Nowlin, Linda Rosen, Cary Sneider, Chasidy White, and Joe Willhoft. ADC member Dale Nowlin leads these working group discussions.

The first working group teleconference was held on April 7, 2017 and was focused on reviewing relevant background information and updates and identifying issues to tackle first. Background and updates reviewed:

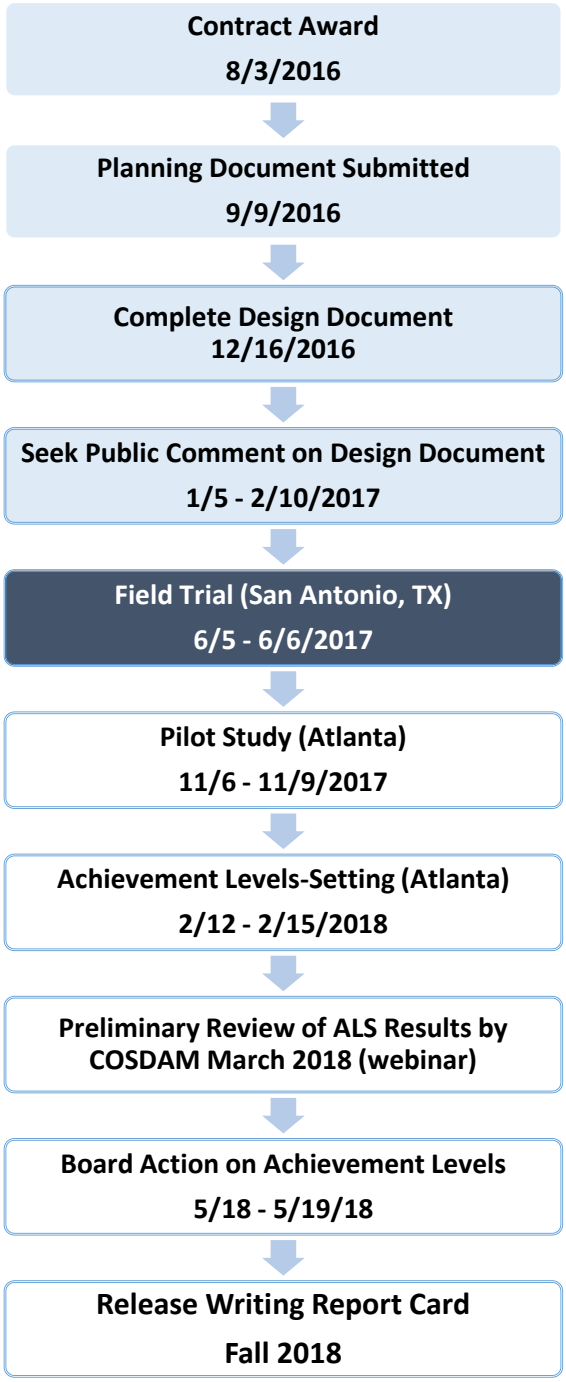
- The Governing Board Framework Development Policy
- Summary chart of framework developments and updates in connection with trend reporting
- Content-related reviews by NCES-convened committees of experts, which includes people who served on the framework development teams convened by the Board
- Current plans and procurements for evaluating framework update needs, i.e., the state math standards review project and upcoming framework update projects
- Input from the April 3, 2017 meeting of Governing Board CCSSO Policy Task Force, sharing how states conduct their own standards revision processes (March 2017 Board meeting discussions requested more information on state processes)

The April 7 working group discussion raised the following issues for discussion and follow-up:

- The current Board Framework Development Policy focuses primarily on processes for developing completely new frameworks, which is appropriate given the Board's early work. Prospective policy revisions should add more guidance for updating frameworks.
- More ongoing processes to indicate when NAEP frameworks do or do not require updates are needed. Establishing a menu of monitoring options may be helpful.
- The 10-year default for framework stability in the policy may be unnecessarily rigid, but drastic reconceptualizations of a subject area prompting a completely new framework are rare. Hence, framework updates are a more likely default than framework replacements.
- Since there could still be policy reasons to pursue a new framework and trend, policy revisions may be needed to clarify when new frameworks should be considered.
- More information on previous processes to determine and implement framework "tweaks," will be helpful in determining new or more formal guidance for framework updating.
- More information is needed to determine opportunities for reducing current lead times between starting a framework change and administering the new assessment.

At the May 2017 Board meeting, working group members will update their respective committees about the latest discussions from the second teleconference on May 11, 2017.

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
Dark shading: Current status
No shading: To be completed after June 2017

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

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

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

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

[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. 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)

May 2017 Update:

Revisions to the Design Document based on Public Comment

Public comments on the Design Document were discussed with the TACSS during the webinar on February 16th and presented to COSDAM at the March 2017 Board meeting. Pearson evaluated each comment received and addressed each one with the Governing Board Contracting Officer's Representative.

Editorial comments were reviewed and revisions to the Design Document were made where appropriate. Several comments were observations only and did not necessitate revisions. Other comments called for more detailed descriptions of processes or procedures. Additional comments had to do with technical considerations and/or clarifications. Generally, where justifications or clarifications were called for, or where process modifications were implemented as a result of consultation with the TACSS, the Design Document was revised accordingly. Otherwise, there were several instances where it was proposed that further elaboration would be better served by including more information in the final report.

There were several comments that were discussed with the TACSS prior to reaching a decision around how best to address. One of the first had to do with a recommendation to include highly discrepant bodies of work (where scores on a student's responses to two prompts differ markedly) in the ALS process. TACSS discussion focused on the need to have bodies of work reflect similar composition to the overall population while not introducing confusion into the judgment task. Analysis of 2012 grade 4 writing pilot data revealed that there were roughly 3 percent of cases where scores differed by more than 2 rubric points. As this reflected such a small percentage, TACSS recommended excluding such cases from being used for the judgment task. This was clarified within the Design Document and it was noted that the recommendation was a by-product of public comment.

Another comment assumed that ability estimates for individual bodies of work would be conditioned on background variables (as is utilized as a means of producing score distributions from NAEP assessment data). What had been proposed for this ALS approach was to use unconditioned expected a priori (EAP) ability estimates based on existing item response theory parameter estimates and student item-level scores. TACSS discussion focused on the fact that within the judgment task, since background information is unknown to panelists, this would introduce noise and possibly confusion to the task. This was noted to be in line with the decision made for the 2011 writing achievement levels setting at grades 8 and 12. The Design Document was revised to include the rationale being the chosen approach.

One comment questioned the plan to explicitly engage panelists in a separate activity during the Field Trial in order to gauge and assuage any potential concern over grade 4 students' ability to write on tablets. TACSS recommended that there should not be an over-emphasis on concerns about typing ability. As a result, the Design Document was modified to reflect a change whereby this information will be presented in opening comments from the Governing Board only, premised on recent findings out of the 2012 grade 4 writing pilot study.

TACSS also discussed comments suggesting the name of the method be referred to as a "modified" Body of Work approach. This they recommended against. Also, one comment questioned the presentation of mean cut scores in addition to median cut scores for panelist feedback (where the median will be used as the cut score of record). TACSS supported only providing panelists with the median cut scores. This revision was also applied to the Design Document.

In public comment, one reviewer questioned the value of using the interactive tool to understand how different cut scores result in different impact data when no explicit explanation of the broader context is provided. That is, it is important to tie that activity directly to student performance via particular bodies of work and not to simply relate the resulting impact to the performance distribution. This was clarified within the Design Document.

All revisions were reviewed and discussed at the TACSS meeting on April 20th and 21st. No additional changes to the Design Document were made at this time; the Design Document is now considered final.

Update on Preparations for the Field Trial

Pearson is currently in the process of finalizing all materials and tools necessary to conduct the field trial, to include creation of the Moodle interface. The target number of panelists for the field trial is 20 (11 teachers, 3 non-teacher educators, and 6 general public). Initial recruitment was within a 30 mile radius of the San Antonio meeting site and resulted in 69 nominee recommendations (as of 4/26/17) out of roughly 400 nominator contacts across each of the recruitment categories (teacher, non-teacher educators, and general public). Due to the difficulty of recruiting general public panelists (professionals in the field of writing), recruitment for up to 5 individuals in this category was expanded to the Austin area.

Materials (including presentation slides), qualifications of potential panelists, meeting logistics, and the Moodle interface were reviewed and discussed with the TACSS during the recent meeting on April 20th and 21st. The field trial will be conducted from June 5-6, 2017.

The Moodle interface was also demonstrated to interested Governing Board members via a webinar on April 17th. The demonstration included a functional walk-through of the tool features within the context of the upcoming field trial. One question was raised with respect to the judgment task activity in which panelists will have access to grade 4 writing achievement level descriptions as supporting information. The question was about the extent to which the NAEP policy achievement level descriptions will be emphasized, and whether and how they would be available to panelists during the judgment activity. Policy ALDs are incorporated into training materials and will be emphasized throughout the meeting as to their importance relative to each field trial activity. Additionally, the policy ALDs will be provided to panelists as part of pre-meeting materials delivered via Moodle and will be available in printed form for all panelists to refer to during the judgment activity.

During the August 2017 COSDAM meeting, project director Tim O'Neil will describe lessons learned from the field trial, including any potential revisions to procedures planned for the November 2017 pilot study. There will not be a presentation on this project at the May 2017 COSDAM meeting.

Procurement Update

Technical Support in Psychometrics, Assessment Development, and Preparedness for Postsecondary Endeavors

The Governing Board has a need for technical support to implement some of the technical activities included in the Strategic Vision, such as: planning and designing statistical linking studies; researching how NAEP is used by various audiences and the extent to which various uses are intended and appropriate; developing approaches to updating assessment frameworks while maintaining trends; exploring options to reconfigure the Long-Term Trend assessments; learning best practices from other assessment programs; and exploring the use of NAEP as a measure of preparedness for postsecondary education and careers.

In addition, the Governing Board's response to the recent evaluation of NAEP achievement levels (National Academies of Sciences, Engineering, and Medicine; November 2016) refers to several follow up activities to be pursued over the next few years. Technical support will be needed to assist the Governing Board with implementing some of the activities such as: providing input to inform Board policy revision on setting achievement levels for NAEP; conducting research on the relationship between NAEP achievement levels and concurrent or future performance on measures external to NAEP; conducting research on how the achievement levels are used by various audiences; preparing validity arguments to support the intended uses and interpretations of the achievement levels; and conducting research and producing guidance on inferences best made with achievement levels and those best made with scale scores.

The technical support will include research studies; technical memos; literature reviews and syntheses of best practices; attendance at Governing Board and other meetings; expert consultant services (including the convening of panels); and other ad hoc and quick turnaround requests.

On April 28th, a Request for Proposals (RFP) was issued on www.fbo.gov: https://www.fbo.gov/spg/ED/NAGB/NAGB/ED-NAG-17-R-0003/listing.html. Proposals are due on June 14th, with an anticipated award date of August 2017. The contract period of performance is 12 months for the base year, with two one year options of 12 months.

Review and Revise Achievement Level Descriptions in Mathematics and Reading

At the March 2017 meeting, COSDAM discussed the need for a future procurement to review and potentially revise the achievement level descriptions in mathematics and reading at all three grades. This was one of the primary recommendations of the recent evaluation of NAEP achievement levels. The Governing Board's acquisition plan has been updated to indicate that a procurement to conduct this work will be awarded during Fiscal Year 2018. The results are intended to be used in the reporting of the 2019 Mathematics and Reading Report Cards.

Next Steps for Implementing Strategic Vision

During the November 2016 board meeting, a Strategic Vision was formally adopted to guide the Board's work over the next several years, with a general goal of increasing the impact of NAEP through increased dissemination and innovation. At the March 2017 board meeting, COSDAM discussed a proposed list of draft activities for which the committee was assigned primary responsibility. COSDAM members noted that the proposed activities seemed reasonable but that it would be helpful to better understand how each activity might be implemented.

The Governing Board staff is working on a plan for documenting milestones and timelines for all Strategic Vision activities using project management software; additional information will be shared with the Board during the August 2017 meeting.

In the meantime, a preliminary list of next steps has been drafted for each of the activities primarily assigned to COSDAM. 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 committee (e.g., framework update processes) or a task force (i.e., exploring new approaches to postsecondary preparedness) also have not been included below. Finally, the table does not yet specify details about timelines.

Strategic Vision Activity	**Current Status**	**Potential Next Steps**	**Desired Outcome**
2a: 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 discussion at May 2017 board meeting 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; planned studies by NAEP Validity Studies (NVS) panel	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)

3e: 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) *3f: 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 appropriate uses of NAEP Procurement for Technical Support contract specifies that the research study topic for year 1 will focus on how NAEP results are used by various stakeholders	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 this document).
5c: Consider new approaches to creating and updating the achievement level descriptors and update the Board policy on achievement levels.	Initial conversations conducted with 7 standard setting experts in March/April 2017 COSDAM discussion at May 2017 board meeting about scope and process of revising the achievement levels policy	Conduct literature review/synthesis of best practices for creating and updating achievement level descriptors (ALDs) Convene expert panel to discuss technical issues and recommendations for achievement levels policy, including specific guidance about ALDs Draft revised policy statement for Board discussion Seek external feedback and public comment Revised policy statement for Board discussion and ultimately adoption	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.

7a: 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 and posted to Governing Board website (February 2017), symposium held in Washington, DC (March 2017), and follow-up event held at American Educational Research Association (AERA) annual conference (April 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 are needed (7b) and/or whether changes to the design and administration of the LTT assessment are needed (7c)
9b: Pending outcomes of stakeholder input (9a), evaluate the technical implications of combining assessments, including the impact on scaling and trends.	This activity is entirely dependent on activity 9a (ADC's determination that it is advisable to combine multiple subject area frameworks from a content perspective) and on the particular subjects that may be combined.	TBD	TBD
10a: 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 activity 2a)	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.