

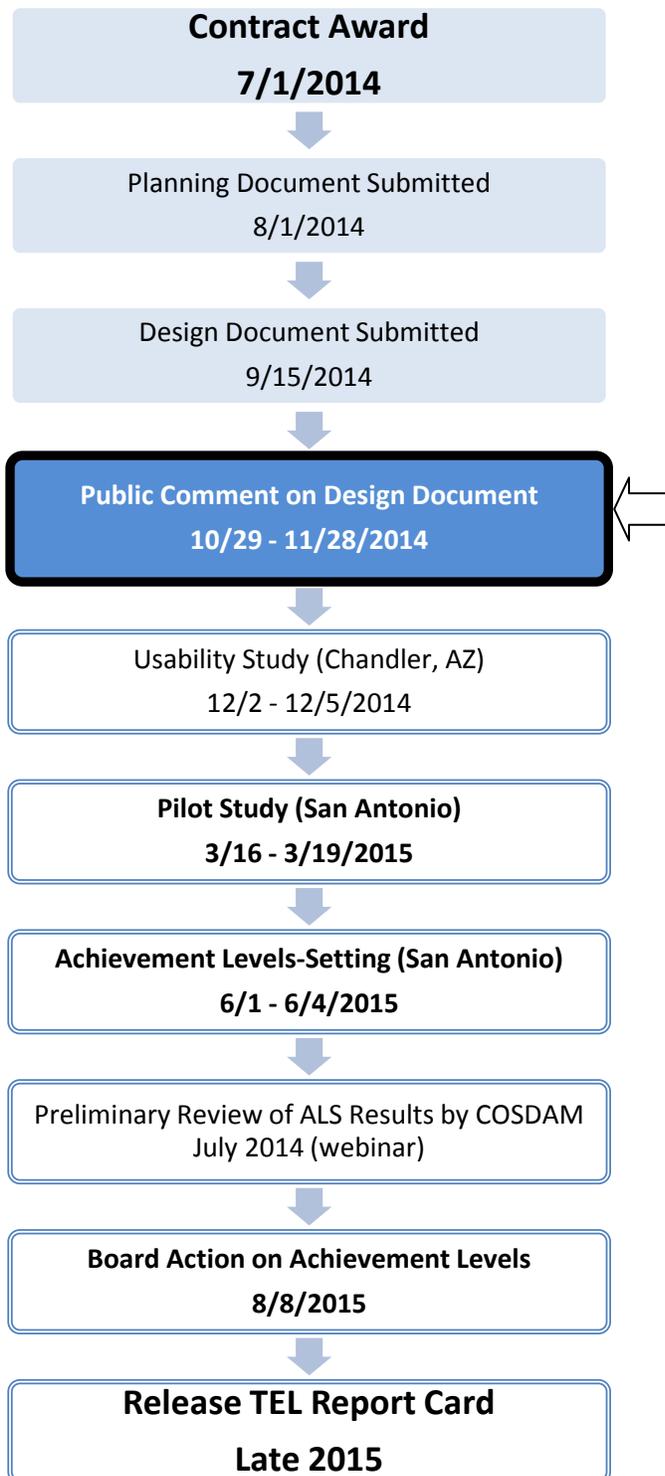
National Assessment Governing Board Committee on Standards, Design and Methodology

**November 21, 2014
10:00 am – 12:30 pm**

AGENDA

10:00 – 10:05 am	Introductions and Review of Agenda <i>Lou Fabrizio, COSDAM Chair</i>	
10:05 – 11:00 am	Project Update and Design Document for Technology and Engineering Literacy (TEL) Achievement Levels Setting <i>Paul Nichols, Pearson Sharyn Rosenberg, Governing Board Staff</i>	Attachment A
11:00 – 11:10 am	Break	
11:10 – 11:55 am	Update on Transition to Technology-Based Assessment <i>Andreas Oranje, Educational Testing Service</i>	Attachment B
11:55 am – 12:10 pm	Upcoming Procurement: Review of Existing Studies on Motivation and Engagement in NAEP <i>Sharyn Rosenberg, Governing Board Staff Lou Fabrizio, COSDAM Chair</i>	Attachment C
12:10 – 12:25 pm	Questions on Information Items <i>Sharyn Rosenberg, Governing Board Staff</i>	See below
12:25 – 12:30 pm	Other Issues and Questions <i>COSDAM Members</i>	
	<p>Information Items:</p> <ul style="list-style-type: none"> • Evaluation of NAEP Achievement Levels Contract Award • Update on Academic Preparedness Research • Origin of English Language Learners (ELL) Inclusion Guidelines 	<p>Attachment D</p> <p>Attachment E</p> <p>Attachment F</p>

**DEVELOPING ACHIEVEMENT LEVELS FOR THE NATIONAL ASSESSMENT OF EDUCATIONAL
PROGRESS TECHNOLOGY AND ENGINEERING LITERACY (TEL) AT GRADE 8**



Purpose: The purpose of this session is to provide an update to the Committee on Standards, Design and Methodology (COSDAM) regarding the development of achievement levels for the 2014 NAEP TEL and to present the plans for implementing the item mapping standard setting methodology. In this session, Paul Nichols, NAEP TEL Achievement Levels-Setting (ALS) Project Director for Pearson, will provide an update on the project and an overview of the Design Document.

**Focus Policy Issue for November 2014
COSDAM briefing on Design**

Document:

How should Pearson collect public comment on the proposed achievement levels?

Legend:

Light shading: Completed

Dark shading: Current status

No shading: To be completed after 11/21/2014

Project Overview: On July 1, 2014, the National Assessment Governing Board awarded a contract to Pearson to develop achievement levels for the National Assessment of Educational Progress (NAEP) Technology and Engineering Literacy (TEL) assessment. The computer-based 2014 NAEP TEL is based on the Board-adopted TEL Framework and consists of both scenario-based tasks and discrete items. The first-ever TEL assessment was administered to a nationally representative sample of more than 22,000 grade 8 students in 2014.

Dr. Paul Nichols is the TEL ALS project director at Pearson. Working with Conference Solutions, EdCount and Measurement Incorporated as subcontractors, Pearson will conduct a usability study, 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 NAEP TEL. Pearson will implement an item mapping methodology using software developed by Measurement Incorporated to collect panelists' ratings and present feedback. Conference Solutions will assist Pearson in planning and delivering meetings. Dr. Lori Nebelsick-Gullett from EdCount will serve as the process facilitator for the pilot and operational ALS meetings; Dr. Johnny Moye will serve as the content facilitator for the pilot and operational ALS meetings; and Dr. Susan Cooper Loomis will serve as a consultant to Pearson.

For standard setting, Pearson will use an item mapping process in which panelists will make criterion-referenced, content-based cut score recommendations. The content-based judgments will be made over three rounds. The process to be implemented for the standard setting meeting follows item mapping procedures used in previous NAEP standard setting studies. In addition, two studies will be completed prior to the pilot study: 1) a study of the functioning of the standard setting software used to collect panelists' item ratings and other judgments, and 2) a dual-computer usability study.

The Governing Board policy on Developing Student Performance Levels for NAEP 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 5 in-person meetings and up to 3 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, The University of North Carolina at Chapel Hill

Dr. Barbara Dodd

Professor of Quantitative Methods, The University of Texas at Austin

Dr. Kristen Huff

Senior Fellow, the University of the State of New York Regents Research Fund

Dr. Matthew Johnson

Associate Professor of Statistics and Education, Teachers College, Columbia University

Dr. Marianne Perie

Director, Center for Educational Testing and Evaluation, University of Kansas

Dr. Mary Pitoniak

Strategic Advisor for Statistical Analysis, Data Analysis, and Psychometric Research, Educational Testing Service (NAEP Design, Analysis, and Reporting Contractor)

November 2014 Update:*Technical Advisory Committee on Standard Setting (TACSS) meeting (August 18-19)*

The TACSS met August 18-19 to discuss the Planning Document and draft Design Document. Among the issues discussed were the project schedule, potential sources of external validity evidence, means to detect panelist' positive response bias, and use of computers during the ALS process:

1. The TACSS was concerned that the original project schedule was overly ambitious and placed at risk the success of the ALS activities. The TACSS recommended that the cut score recommendations be presented to the Governing Board at the August 2015 meeting, instead of the May 2015 meeting as initially proposed. The schedule of ALS activities was adjusted to reflect this extended timeline (the chart at the beginning of this briefing shows the key events for the updated timeline).
2. The TACSS recommended eliminating the consideration of external validity evidence from the ALS process. This recommendation came after Pearson described their attempts to identify sources of relevant external validity evidence and presented the conclusion that these sources of evidence—other measures of technology and engineering literacy and related knowledge and skills—were not available for the pilot study or the ALS meeting. The TACSS, after exploring options for external validity evidence, recommended forgoing the external validity evidence as part of the ALS process. The Governing Board staff, following the recommendation of the TACSS, has eliminated the consideration of external validity evidence from the pilot study and the ALS meeting.
3. During the COSDAM meeting on August 1, there was a discussion about the extent to which standard setting panelists may be apt to overstate their confidence in the process. COSDAM members suggested that a “lemon item” might be incorporated into the panelist evaluation process to measure positive response bias, or that panelist feedback could be compared to previous standard setting activities of more traditional subjects. The TACSS discussed these issues and expressed a concern that the use of “lemon items”

could be confusing and insulting to panelists. They noted that evaluation questions are worded in a way that selecting the highest response option does not always indicate a positive outcome. They endorsed the suggestion to use some of the same evaluation questions from previous NAEP standard settings to allow for appropriate comparisons across subjects.

4. Given the uniqueness of the NAEP TEL assessment, and the dual-computer setup for the panelists' activities (one for viewing the items/tasks and another for interacting with the standard setting software), the TACSS recommended that a separate usability study be conducted prior to the pilot study. The goal of the usability study is to try out the ALS panelist setup and allow for planning and modifications prior to the pilot study. A follow-up webinar of the TACSS was conducted on October 7th to discuss the proposed design for this study, which has been incorporated into the Design Document.

The TACSS also made recommendations on recruitment of standard setting panelists, procedures for the item mapping methodology, and the agenda for the pilot and ALS meetings. The Design Document was revised based on TACSS member feedback and was submitted to the Governing Board staff on September 15th. Additional revisions were made based on the staff review and the October 7th TACSS webinar. The Design Document was distributed for public comment from October 29th to November 28th at (<http://www.pearsonassessments.com/naeptelassessment>).

The TACSS is scheduled to meet in December 2014 and February 2015. During the December meeting, the TACSS will discuss any feedback on the Design Document received from the public comment period and COSDAM; results from the dual-computer usability study; and initial preparations for the March 2015 pilot test, including the distribution of items across ordered-item booklets. During the February meeting, the TACSS will review facilitator training materials and plans for implementing the pilot test.

Design Document and Procedures for Collecting Public Comment on Proposed Levels

The Design Document is intended to provide the foundation for all achievement levels-setting activities. The Design Document for the TEL achievement levels-setting process includes discussion of the methodology, procedures, and documentation of the entire project.

During the November 2014 COSDAM session, TEL ALS Project Director Paul Nichols will provide an overview of the Design Document (see Attachment). In particular, he will seek input from COSDAM on options for collecting public comment on the ALS outcomes.

According to the Governing Board policy on [Developing Student Performance Levels for the National Assessment of Educational Progress](#), public comment will be sought at critical junctures throughout the process, including for the proposed levels:

“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” (p. 6).

For previous achievement levels-setting activities, the proposed levels have been treated as embargoed data and as such have not been widely distributed for public comment in advance of the Report Card release. In prior achievement levels-setting activities for NAEP, public comment has been sought on the achievement levels descriptions and/or a sample of exemplar items.

Public comment on the TEL achievement levels descriptions was already collected in May 2014, as part of the process of finalizing the TEL ALDs that were adopted by the Governing Board on August 2, 2014. Pearson requests input from COSDAM on what additional activities, if any, should be conducted to collect public comment on the proposed TEL ALS achievement levels. One idea that has been proposed is to convene a small group of state testing representatives in conjunction with the National Conference on Student Assessment meeting scheduled to take place in San Diego from June 22-24, 2015. The proposed TEL ALS levels and results could be shared with a pre-selected small group of interested stakeholders for comment, provided that participants sign a confidentiality agreement.

Next Steps for COSDAM:

The March 2015 update to COSDAM will include information on the results of the usability study and a description of plans for the pilot study.

National Assessment Governing Board

Developing Achievement Levels on the National Assessment of Educational Progress for Technology and Engineering Literacy at Grade 8 Design Document

Submitted: September 20, 2014

NAEP TEL ALS Design Document

Submitted to:

National Assessment Governing Board
800 North Capitol Street, NW, Suite 825
Washington, DC 20002-4233

This study was funded by the
National Assessment Governing Board
under Contract ED-NAG-14-C-0001.

Submitted by:
NCS Pearson, Inc.
2510 N. Dodge Street
Iowa City, IA 52245-9945
Phone: 319.354.9200

Table of Contents

Table of Contents.....	2
List of Figures.....	4
List of Appendices.....	4
Executive Summary.....	5
Section 1: Achievement Levels-Setting Methodology.....	7
Selecting a Standard Setting Methodology.....	8
Section 2: Studies of Software Functionality and Computer Usability.....	11
Study 1: Item Mapping Software Functionality.....	12
Study 2: Dual-Computer Usability Study.....	14
Section 3: Achievement Levels Panels.....	20
Identification of Panelist Nominators.....	22
Selection of Panelists.....	28
Section 4: Pilot Study.....	30
Section 5: Briefing Materials.....	31
Section 6: Achievement Levels-Setting Procedure.....	32
Division of Panelists into Subgroups.....	33
Division of Items into Subsets.....	34
Use of NAEP-Like Scales.....	36

Training of Facilitators	36
Use of Computers	36
Procedure for Achievement Levels-Setting.....	41
Completion of Consequences Questionnaire	55
Selection of Exemplar Items.....	56
Procedural Validity Evidence	57
Section 7: Data Analyses and Results Presentation.....	58
Analyses of Data.....	58
Strategies for Results Presentation	61
Section 8: Exemplar Tasks and Items/Responses.....	63
Section 9: Public Comment.....	67
Public Comment on the Design Document	68
Public Comment on the ALS Outcomes	68
References	69
Appendix A: Draft Agenda for the Pilot Study and ALS Meeting	72
Appendix B: Evaluation of the Success of Standard Setting Training	77
Appendix C: Evaluation of the Round One Standard Setting Process	79
Appendix D: Draft Text for the Design Document Public Comment Web Page...82	

List of Figures

- Figure 6.1: The assignment of item sets to panelist subgroups.
- Figure 6.2: A slide used to introduce the idea of a borderline student.
- Figure 6.3: An illustration of an item map (ACT, Inc., 2010).
- Figure 6.4: A slide used to illustrate the ordered item book (OIB).
- Figure 6.5: The rater location chart used as a model for the software developers.

List of Appendices

- Appendix A: Draft Agenda for the Pilot Study and ALS Meeting.
- Appendix B: Evaluation of the Success of Standard Setting Training.
- Appendix C: Evaluation of the Round One Standard Setting Process.

Executive Summary

The National Assessment of Educational Progress (NAEP), known as the “The Nation’s Report Card,” provides information on what students in the United States know and can do in various subject areas. The 2014 NAEP for Technology and Engineering Literacy (TEL) assessment is a new assessment framework and is unique in that it is wholly computerized, consists of both scenario-based tasks and discrete items, and is based on a diffuse curriculum.

The TEL assessment was designed to measure the following three interconnected areas:

- Technology and Society
- Design and Systems
- Information and Communication

In addition, the TEL assessment was designed to measure three ways of thinking and reasoning that are used when solving a problem. The following three ways of thinking and reasoning, called practices, are expected to be demonstrated in in each of the three areas:

- Understanding technological principles
- Developing solutions and achieving goals
- Communicating and collaborating

For more information on the design of the TEL assessment, please visit

<http://nces.ed.gov/nationsreportcard/tel/whatmeasure.aspx>.

The TEL assessment was administered for the first time in 2014 to a nationally representative sample of more than 22,000 grade 8 students. The assessment covered a breadth of content and included 20 scenario-based tasks and 98 discrete items. Because the assessment consisted of more questions than a single student could answer, each student took just a small portion of the assessment. Students responded to assessment questions for 60 minutes.

Achievement levels have become a powerful way to communicate student achievement on an assessment like the NAEP TEL because achievement levels interpret test performance with reference to cut scores that quantitatively define ordered categories of achievement such as basic or proficient (Haertel & Lorie, 2004). An important source of evidence used by policymakers to establish achievement levels is the cut score recommendations that result from an achievement levels-setting (ALS) meeting. Cut scores are the outcome of a facilitated process, called a standard setting meeting, that systematically elicits judgments from experts related to the test content and the knowledge, skills and abilities of the test takers (Hambleton, Pitoniak & Copella, 2012).

The National Assessment Governing Board has issued a contract to Pearson to implement a process to produce a set of cut score recommendations to assist the National Assessment Governing Board in developing achievement levels for the 2014 NAEP TEL. On behalf of the Governing Board, Pearson has developed a Design Document that describes in detail the NAEP TEL achievement levels-setting activities. This document, the Design Document for developing achievement levels on the NAEP TEL at Grade 8, is intended to provide the foundation for all ALS activities. The Design Document for the TEL ALS process includes discussion of the ALS methodology, the collection of public comment and the identification of exemplar items. Once adopted, the Design Document will be used to guide the achievement levels-setting activities to produce a set of cut score recommendations for reporting achievement levels for the 2014 administration of the NAEP TEL.

For standard setting, Pearson has proposed using an item mapping process in which panelists will make criterion-referenced, content-based cut score recommendations. The content-based judgments will be made over three rounds. The process to be implemented for the standard

setting meeting follows item mapping procedures used in previous NAEP standard setting studies. In addition, the following two studies will be completed prior to the pilot study: a study of the functioning of the standard setting software used to collect panelists' item ratings and other judgments and a dual-computer usability study.

To make the process more efficient and strengthen the validity argument for the ALS outcomes, Pearson will use computers during both the pilot study and the ALS meeting. Using computers and specially developed software will reduce the time required for panelists to complete most steps in the standard setting activities. The use of computers strengthens the validity argument for the ALS outcomes because panelists will be able to interact with the discrete and scenario-based items directly, just as students did, and more accurately judge the cognitive demands imposed upon students by the assessment. Panelists can more accurately judge the cognitive demands placed on students by the items if the panelists can attempt to respond to the items in the same context as the students' experience.

Section 1: Achievement Levels-Setting Methodology

Performance standards have become a powerful way to communicate student achievement because they interpret test performance quantitatively, with reference to cut scores, by defining ordered categories such as basic or proficient (Haertel and Lorié, 2004). An important source of evidence that policymakers like the National Assessment Governing Board (hereafter referred to as the Governing Board) use to establish performance standards is the cut score recommendations that result from an ALS meeting. Cut score recommendations are the outcome of a facilitated process that systematically elicits judgments from experts related to the test content and the skills of the test takers (Hambleton, Pitoniak, & Copella, 2012).

Selecting a Standard Setting Methodology

The Governing Board faces the following two challenges for establishing cut scores on the NAEP TEL. First, the ALS methodology used for the NAEP TEL must meet all requirements for NAEP ALS as described in the policy framework entitled *Developing Student Performance Levels on the National Assessment of Educational Progress*. In addition, the standard setting methodology must be appropriate for the TEL Framework that requires both discrete and scenario-based items for the assessment. Pearson explored several possible standard setting methodologies for recommending cut scores for the NAEP TEL. An alternative that Pearson examined was the Body of Work methodology (Kingston & Tiemann, 2012). The items in the scenario-based tasks might be viewed as representing a coherent set of work. In addition, the Body of Work methodology has been used with small numbers of discrete items combined with extended constructed responses. But scaling analyses by the Data Analysis and Reporting contractor indicated that both items from the scenario-based tasks and discrete items could be scaled using a unidimensional approach. Pearson felt that the standard setting methodology adopted should be consistent with the scaling approach used for the NAEP TEL.

Pearson proposed an item mapping approach (Lewis, Mitzel, Mercado, & Schulz, 2012). The item mapping approach appeared to satisfy the main considerations when choosing an appropriate standard setting methodology (Hambleton & Pitoniak, 2006): (a) the method is appropriate for the item types and item scaling, (b) the judgments were likely to be completed in a reasonable amount of time, (c) the Governing Board has experience with the item mapping method, and (d) the measurement field appears to view item mapping as supported by current validity evidence.

Second, the Governing Board is facing an increased focus on external validity evidence for standard setting results. The *Final Report on the Evaluation of NAEP* encouraged NAEP to prioritize gathering external validity evidence that supports the uses and interpretations of its achievement levels. The Governing Board in the request for proposals directed that the proposed ALS methodology provide evidence of the external validity of the outcomes. The Governing Board asked that the proposed ALS methodology address the design, implementation, and reporting of research to provide the Governing Board with validity information relevant to evaluating the cut scores recommended.

With respect to the validity of the cut score recommendations, Sireci, Hauger, Wells, Shea, and Zenisky (2009) noted that a number of writers (Cizek, Bunch, & Koons, 2004; Hambleton, 2001; Kane, 1994, 2001) have supported three general categories of validity evidence for standard setting results: procedural, internal, and external. Procedural validity refers to the appropriateness of the standard setting procedures and how well those procedures were implemented. Internal validity refers to the internal consistency of data generated within the standard setting meeting. External validity refers to the relationships between decisions made using the performance level scores and the same kinds of decisions made using a different reference source as the basis.

Despite the important role of validity evidence for the cut score recommendations, gathering this validity evidence continues to be an area of controversy for practitioners and researchers. External validity evidence for cut score recommendations is often not addressed until too late into the process, which in some cases can lead to the need for post hoc adjustments by policymakers (Haertel, 2002, 2008; McClarty, Way, Porter, Beimers, & Miles, 2013).

Pearson proposed using a coherent ALS process that took into consideration various sources of external validity evidence for cut score recommendations. The Technical Advisory Committee on Standard Setting (TACSS) reviewed the proposed approach and the availability of external validity evidence for the standard setting results. The TACSS consists of five outside technical consultants. In addition, a member of the Design, Analysis, and Reporting contractor's staff is included on the TACSS. The role of the TACSS members is to advise Pearson on issues related to standard setting and psychometrics. The TACSS recommended eliminating the consideration of external validity evidence from the ALS process. This recommendation came after Pearson described their attempts to identify sources of relevant external validity evidence and presented the conclusion that these sources of evidence—other measures of technology and engineering literacy and related knowledge and skills—were not available for the pilot study or the ALS meeting. The TACSS, after exploring options for external validity evidence, recommended forgoing the external validity evidence as part of the ALS process. The Governing Board staff, following the recommendation of the TACSS, has eliminated the consideration of external validity evidence from the pilot study and the ALS meeting.

The TEL assessment includes interactive, scenario-based tasks. Three types of scenario-based assessment sets were used: long (30 minutes), medium (20 minutes), and short (10 minutes). To understand the context of the TEL assessment, panelists must experience and interact with the animations, audio, and video components of the items embedded in the scenarios. In addition to interactive, scenario-based tasks, the NAEP TEL assessment includes a set of discrete items.

The standard setting method has to accommodate the complex performance-based scenarios as well as discrete items. The item mapping process will allow Pearson to collect

content-centered judgments across both scenarios and discrete item blocks using the same standard setting format. Panelists must also be able to analyze the cognitive demands of the discrete items and items embedded in scenario-based tasks within the context of interaction with animations, audio, and video components of the assessment. Panelists will use one computer and the software installed on that computer to collect content-centered judgments. Panelists will use a second computer to interact with the items and analyze the cognitive demands imposed by the items. The use of two computers for both the pilot study and the ALS meeting will be referred to as the dual-computer setup.

Section 2: Studies of Software Functionality and Computer Usability

Pearson has proposed using computers during both the pilot study and the ALS meeting. Using computers and specially developed software will reduce the time required for panelists to complete most steps in the standard setting activities. In addition, the use of computers will allow panelists to interact with the discrete and scenario-based items and more accurately judge the cognitive demands imposed by the items. The panelists can more accurately judge the cognitive demands the items placed on the students if the panelists can attempt to answer the items under the same conditions experienced by the students taking the test.

Pearson has proposed two studies that will examine the use of computers in standard setting meetings. These studies will be completed well before the pilot study is convened so that findings can be used to inform planning of the pilot study and ALS meeting. This section describes the following two studies: (1) a study of the functioning of the standard setting software used to collect panelists' item ratings and other judgments and (2) a dual-computer usability study.

Study 1: Item Mapping Software Functionality

The goal of this study is to evaluate the functioning of the standard setting software used to collect panelists' item ratings. This study will address the following questions:

1. Does the software interface function effectively with the panelists?
2. Are the ratings and item review comments collected accurately?
3. Are the ratings and item review comments transferred accurately from the computers used by panelists during standard setting to the computer used for data analysis?
4. Is the data used to create the consequences data and the Ordered Item Booklet (OIB) accurately loaded into the computer?
5. Is the data used to create feedback following each standard setting round transferred accurately from the computer used for data analysis to the computers used by panelists during standard setting?

Materials

This study will use four laptop computers provided by Pearson. The standard setting software will be loaded onto three of these computers. These three computers will be used to collect item ratings from the study participants. The fourth computer will be loaded with Excel and SAS and the Excel and SAS code required to compute rater feedback. This computer will be used to analyze the item ratings collected from the study participants.

In addition, each computer will be loaded with item data and item images for one short scenario-based tasks and a block of discrete items. Pearson proposes to use one scenario and one discrete item block consisting of six discrete items.

A set of unique, predefined sequence of item ratings will be created. Predefined item rating sequences will be used to evaluate the accuracy of data transfer between the computers used by panelists during standard setting to the computer used for data analysis. A different sequence of item ratings will be created for each panelist and for each of the three rounds of item ratings that the panelist will complete. A total of nine item rating sequences will be created.

Participants

Pearson staff will serve the role of panelists for this study. Three Pearson staff members will be recruited and used for the study. In addition, a data analyst will analyze the panelist ratings and create the panelist feedback. All study participants will sign a confidentiality agreement.

Procedure

The study will be conducted at the Pearson Iowa City, Iowa office. The study will be conducted in a single afternoon.

A Pearson panelist will be assigned to each computer. Initially, panelists will be trained on the use of the standard setting software. Next, panelists will complete the first round of standard setting. Panelists will enter their predefined sequence of item ratings for round one. After completing round one, panelists will complete an evaluation.

The panelist's item ratings and evaluation responses will then be transferred to the computer of the data analyst. The data analyst will analyze the item ratings and create cut score statistics. These statistics will be transferred back to the panelists' computers. The panelist feedback will be created and displayed to the panelists on their computers.

Next, panelists will complete the second round of standard setting. Panelists will enter their predefined sequence of item ratings for round two. After completing round two, panelists will complete a second evaluation. The data will again be transferred to the computer of the data analyst and analyzed by the data analyst. The data analyst will analyze the item ratings, transfer the statistics back to the panelists' computers, and feedback will be displayed. The same procedure will be followed for a third round.

Analyses

Pearson will monitor the computers for errors during both data collection and feedback presentation. During data collection, the responses collected will be reviewed and compared to the unique, predefined sequence of responses entered by each panelist during each round. Any exception will be noted. During feedback return, the tables returned to the staff members will be reviewed for accuracy. Any errors will be recorded. Pearson will share errors with the standard setting software provider.

Pearson will prepare a report and share the findings with the software provider and the Governing Board. Pearson will work with the software provider to identify and address any software problems and usability issues.

Study 2: Dual-Computer Usability Study

The innovative characteristics of NAEP TEL that make the assessment unique, such as the complex scenarios and the item interactivity, are characteristics that are also novel to the ALS process. As such, a separate study will be conducted to investigate the way in which the

unique assessment features will function within the ALS meeting and the impact they may have on the ALS panelists.

Many of the TEL items are embedded within descriptive scenarios. In some cases, the items within a scenario are related to each other and must be administered in a specific sequence. While other assessments have included scenario-based items and groups of items, like those related to reading passages, the NAEP TEL item scenarios include interactive functionality that make them unique. Because of the novelty of these item types, the ALS activities for this assessment will be breaking new ground. In order to complete the ALS activities, panelists will need to be able to review the items within their scenarios, including the scenario functionality and the other items within the scenario, in order to fully understand the knowledge and skills required to answer the items correctly. As such, during the ALS meeting, each panelist will use two computers. One computer will be used to present the items to panelists within the scenarios and with full functionality, as the items were administered by NAEP. Items must be reviewed in order within each scenario; it is not possible to skip to a specific item. A second computer will be used to complete the ALS activities, including presentation of the OIB and the collection of panelists' judgments.

Purpose

Given the uniqueness of the NAEP TEL items, and the dual-computer setup for the panelists' activities, the TACSS has recommended that a separate study be conducted prior to the ALS pilot meeting. The goal of the study is to try out the ALS panelist setup and allow for planning and modifications prior to the pilot meeting and the ALS meeting. Specifically, the research questions for this study are:

1. Does the dual-computer setup hinder the ALS process in any way?
 - 1.1. Can panelists navigate the items, including viewing them within their scenarios, and ratings with the dual-computer setup?
 - 1.2. Does the dual-computer setup distract panelists from the ALS activities?
2. What kind of training is needed to train the panelists to navigate the dual-computer setup?
3. Given the dual-computer setup, how long do panelists need to complete the ALS ratings?

Participants

Pearson will recruit three to six teachers from the Phoenix, Arizona, area for participation in this study. Pearson will investigate the feasibility of recruiting teachers who have productively participated in recent ALS activities in Arizona, such as those done for Arizona's Instrument to Measure Standards (AIMS) or the Arizona English Language Learner Assessment (AZELLA). These teachers will be considered for the study because of their existing knowledge of ALS activities. This experience and knowledge will allow for more of the study time to be spent focused on the NAEP TEL items and the NAEP TEL ALS computer setup and reduce the time spent on general ALS training.

In addition to ALS experience, recruiting will also focus, as much as is feasible, on identifying grade eight science teachers for participation in the study, as this is the subject and grade level relevant to the NAEP TEL assessment.

Study participation is expected to take approximately four hours, and teachers will be compensated for their time.

All teachers who participate in the study will be required to sign confidentiality agreements to ensure the security of the NAEP TEL items.

Materials

The setup and materials used for the study will mirror those planned for use within the NAEP TEL ALS pilot and operational meetings. This includes the computers, items, and training materials used within the study.

The sample of items selected for use in the study will allow for the activity to mirror that of the full ALS activity. Specifically, the group of items used for the study will include two short scenarios and one block of four to eight discrete items. This grouping of items aligns with that used to construct operational NAEP TEL forms. On one computer, these items will be presented as they would appear to a student completing the assessment, meaning the items for each scenario will be presented together with the original stimuli and functionality. On the second computer, the items will be presented in order of item difficulty, like they will appear within the OIB during the pilot and operational ALS activities. Within the OIB, the original functionality of the item will not be accessible; instead, a static depiction of the item will be displayed. The second computer will also include the software needed to record ratings.

Method

The study will take place in a Pearson usability lab in Chandler, Arizona, located outside of the Phoenix metropolitan area. The lab is equipped with a one-way mirror and observation room as well as several cameras that permit sessions to be recorded. Conducting the study in this lab will allow Pearson and the Governing Board staff to observe the participants without interfering in their activities while also capturing participants' feedback and interactions with the NAEP TEL materials.

Participants will be asked to review a packet of materials prior to the study dates to help familiarize them with the activities. This packet will include the NAEP TEL Framework, the achievement level descriptions (ALDs), and an overview of the ALS process.

Study activities will be completed with participants individually after school hours in early December 2014. Each participant will take part in a series of activities designed to mirror those of the pilot and operational ALS meetings. Study activities will include:

- Introduction to the study and description of the study purpose.
- Simulated test-taking experience – The participant will complete the sample of NAEP TEL items as they would be presented to students in regard to order and functionality.
- Review of ALDs – The facilitator and the participant will review the ALDs, specifically focusing on those for Proficient.
- Discussion of the borderline Proficient student – Using a similar approach to that which will be used in the ALS pilot and operational meetings, the facilitator will work with the participant to identify the knowledge and skills that define a just barely Proficient student.
- ALS training – The facilitator will describe the purpose of ALS, provide an overview of the ALS methodology, teach the participant about the OIB, and discuss the process of establishing a rating. As much as possible, the facilitator will use the same materials that will be used for training during the ALS pilot and operational meetings. For those participants with previous ALS experience, the training will include discussion of the variability of ALS methods so that the participant does not expect the activities in the study to exactly align to those of prior ALS meetings. The level of detail included and

time spent on this activity will be determined by the participant's previous experience with ALS activities as well as the facilitator's gauge of the participant's understanding.

- ALS rating activity – The participant will be asked to complete the ALS rating activity in the same manner as that which will be used in the NAEP TEL ALS pilot and operational meetings. For the purpose of the study, the participant will be asked to provide only one rating, specifically that related to the Proficient achievement level. To complete this activity, the participant will be asked to proceed through the OIB; reference the operational presentation of the item, including functionality; consider the ALD of the level relative to the knowledge and skills needed to answer each item correctly; and determine the place at which a just barely Proficient student would stop answering the items correctly with a 0.67 probability. This activity will require the participant to navigate between one computer with the OIB and rating software and the other computer, which will display the NAEP TEL items as they appear to students during the test administration. As the participant completes this activity, he or she will be instructed to verbalize any challenges or confusion that he or she experiences during the rating process. The facilitator will ask the participant questions as needed to elicit commentary regarding any difficulties that the participant experiences with the rating activity.
- Debrief discussion – At the conclusion of the ALS-related activities, the facilitator will have a discussion with the participant about his or her experiences. Specifically, the facilitator will ask questions like, “Did you feel comfortable moving between the two computers for the activity?”, “What, if anything, was the most difficult part of the rating activity?”, and “What, if anything, could we do to make the rating activity easier?”

Results

The video and audio recordings of the study will be reviewed to document the time needed by participants to be trained in the ALS activity and to complete the ALS rating. The qualitative data, specifically comments related to encountered difficulties and ways in which the activity could be improved, will be codified by topic.

Pearson will provide a summary of the findings. As needed, video clips from the study can be used to exemplify any challenges expressed by participants. Based on the study findings, Pearson will provide the Governing Board staff with a list of suggested changes, if any, for the pilot and operational ALS meetings.

Section 3: Achievement Levels Panels

A total of 51 panelists will be recruited to participate in the ALS activities – 15 for the pilot, 30 for the main study, and six as replacements. The panelist pool will include 28 classroom teachers currently engaged in TEL instruction at grade eight (55 percent), 15 members of the general public (30 percent), and eight non-teacher educators (15 percent). The objective of the recruitment plan is to recruit broadly representative, well-qualified panelists to participate in the ALS activities. Panels will reflect an overall balance of gender, race/ethnicity, geographic location, and type of TEL experience, as well as type of institutional affiliation. To reduce the burden of working with the large number of items in the TEL, panelists both in the pilot and the main study will be separated into three rating groups. To the extent possible, each of the groups of panelists will be equivalent in terms of the attributes to be represented on the panels. Given the size of each group in the pilot study (five per rating group), complete equivalence across all attributes is not possible.

The pilot and the ALS study panelist pool will include at least one teacher of English Language Learners (ELL) and one teacher of students in Special Education programs; these teachers must also meet the teacher panelist qualifications presented above. Additionally, a group of six extra panelists will be established, two extra panelists for the pilot panel and four extra panelists for the ALS panel, as backups in the event that some panelists have to drop out before the panel meeting. This group will include three teachers, one non-teacher educator, and two representatives of the general public.

In order to ensure a broad level of representation and a pool of outstanding candidates, the panelists for NAEP TEL ALS will be identified through an iterative three-phase process, as follows:

- Phase 1: Identify nominators through specific stakeholder organizations. These organizations are leaders in the TEL field, and they can identify the best qualified candidates to serve on the ALS panels. In Phase 1, we will contact nominators and ask them to nominate qualified educators and non-educators. Nominators will be asked to provide contact information for candidates and to ascertain that candidates are willing to serve on the panel if they are selected.
- Phase 2: Notify nominees and request that they send a resume and complete an online nominee form to provide information regarding their background in technology/engineering literacy, experience with grade eight students, and other qualifications required for the specific type of panelist (teacher, non-teacher educator, or general public).
- Phase 3: Evaluate the background and experience of nominees and select the most qualified panelists. In addition, the panelists will be selected to be broadly representative with

respect to gender, race/ethnicity, geographic location, type of technology/engineering literacy experience, and type of institutional affiliation.

Phases 1–3 will continue until the sampling target of 51 qualified and available panelists is met. (See Selection of Panelists section for detailed criteria.) Communication with the nominators and nominees will be conducted through email and supplemented by telephone calls as needed to optimize the recruitment process.

Identification of Panelist Nominators

Panelist nominators will be obtained through the allied organizations that were involved in the Steering and Planning Committees for the NAEP TEL Framework development, provided feedback on the Framework, or have a strong background in technology or in providing professional development in TEL. These allied organizations will be supplemented by additional organizations to increase representation and to increase the potential pool of candidates for the panels. The following national organizations will be among those involved in recruiting panelists for the teacher group in four NAEP regions (Northeast, South, Midwest, and West):

- International Technology and Engineering Educators Association (ITEEA), Reston, VA
- International Society for Technology in Education (ISTE), Washington, DC
- Partnership for 21st Century Skills (P21), Washington, DC
- Council of Chief State School Officers, Washington, DC
- State Educational Technology Directors Association (SETDA), Glen Burnie, MD
- National Center for Technological Literacy, Museum of Science, Boston, MA
- FIRST, Manchester, NH

In addition, state superintendents, heads of teacher organizations, school board presidents, and principals of public and private schools in the four NAEP regions will be contacted directly to propose qualified nominators for teacher and non-teacher educator panelists. Based on previous experiences in recruiting NAEP ALS panelists for Writing 2011 (National Assessment Governing Board, 2012) and Science 2009 (National Assessment Governing Board, 2010), the target ratio of nominators from public schools to nominators from private schools will be 9:1.

The process of recruiting panelists for the non-teacher educator group will include contacting deans of a representative sample of technology and engineering higher education institutions, as well as leaders of STEM education centers. The goal is to reach a broad representation of technology and engineering fields in the U.S. that offer education and training in TEL areas (e.g., civil, environmental, chemical, biomedical, biotechnological, electrical, mechanical, space and aeronautics, computer science). An attempt will be made to balance the geographic representation of institutions of higher education. A representative sample of institutions that will be contacted to nominate representatives of the non-teacher educators includes the following:

- School of Engineering (chemical, biology, civil and environmental, electrical engineering), Massachusetts Institute of Technology, Cambridge, MA
- College of Engineering (aerospace, civil and environmental, mechanical, biomedical engineering), Georgia Institute of Technology, Atlanta, GA
- College of Engineering (biomedical, chemical, civil and environmental engineering), Carnegie Mellon University, Pittsburgh, PA

- College of Science and Engineering (computer science, aerospace, biomedical, chemical, civil and environmental engineering), University of Minnesota, Minneapolis, MN
- College of Engineers (civil and environmental, electrical, computing, mechanical engineering), University of Wisconsin-Madison, Madison, WI
- College of Engineers (biomedical, chemical, civil, electrical, industrial, mechanical engineering), The University of Iowa, Iowa City, IA
- School of Engineering (computer science, civil and environmental, mechanical, biomedical, electrical engineering), Stanford University, Stanford, CA
- Center for Integrated Computing and STEM Education, University of California, Davis, CA
- Center for STEM Learning, University of Colorado, Boulder, CO
- School of Engineering (biological and health systems, computing, informatics, and decision systems engineering, transport and energy, civil, environmental and sustainable engineering), Arizona State University, Phoenix, AZ
- College of Engineering (aerospace, agricultural and biological engineering, biomedical, civil and environmental, materials and mechanical engineering), University of Florida, Gainesville, FL
- College of Engineering (biosystems and agricultural, materials science, civil and environmental, chemical engineering), Michigan State University, East Lansing, MI
- College of Engineering (aeronautics and astronautics, agricultural and biological, biomedical and chemical, civil and construction, electrical and computer engineering), Purdue University, West Lafayette, IN

- College of Engineering (aerospace and mechanics, chemical and biological, civil, construction and environmental, computer science, mechanical engineering), University of Alabama, Tuscaloosa, AL
- College of Engineering and Technology (aviation technology, civil, industrial and systems, energy, chemical engineering), Ohio University, Athens, OH
- School of Engineering (aerospace, biomedical, chemical, civil and environmental, electrical, materials engineering), University of Texas, Austin, TX
- The Infinity Project, School of Engineering, Southern Methodist University, Dallas, TX
- UTeach Institute, The University of Texas at Austin, Austin, TX

The group of panelists representing the general public will consist of individuals who are educated in and/or work directly in areas relevant to TEL. This will include individuals from a broad range of engineering industries (e.g., civil, environmental, agricultural, chemical, biomedical, electrical, mechanical, space and aeronautics, computer science). The process of recruiting panelists for the general public group will include contacting individuals in human resources or education offices of companies engaged in technology and engineering activities in each state. Nominators from nationwide companies will be asked to nominate qualified nominees from each of the four NAEP regions who represent the diversity required of the panelists. Companies will be identified from the engineering and technology sectors to represent a broad array of occupations requiring training and experience in engineering and technology. Examples of the organizations and companies that will be contacted to nominate representatives of the general public include the following:

- National Academy of Engineering, Washington, DC

- National Society of Black Engineers, Alexandria, VA
- Society of Women Engineers, Chicago, IL
- Veolia Environmental Services, Chicago, IL
- John Deere, Moline, IL
- Archer Daniels Midland, Chicago, IL
- Verizon, New York, NY
- Comcast, Philadelphia, PA
- Bayer Corporation, Pittsburgh, PA
- Fresenius Medical Care, Waltham, MA
- Cardinal Health, Dublin, OH
- Williams, Tulsa, OK
- Medtronic, Minneapolis, MN
- Monsanto, St. Louis, MO
- General Electric, Fairfield, CT
- Kohler Construction, Pinellas Park, FL
- CH2M Hill, Englewood, CO
- Society of Hispanic Professional Engineers, City of Industry, CA
- Apple, Cupertino, CA
- IBM, San Jose, CA
- Google, Mountain View, CA
- Cisco, San Jose, CA
- Energy Solutions, Salt Lake City, UT
- Koch Industries, Wichita, KS

- Microsoft, Redmond, WA
- Intel, Portland, OR
- Lockheed Martin, Fort Worth, TX
- NASA, Houston, TX
- Rockwell-Collins, Cedar Rapids, IA
- Stanley Consultants, Muscatine, IA
- Mid-American Energy, Davenport, IA
- AT&T, Dallas, TX
- Texas Instruments, Dallas, TX
- ExxonMobil, Irving, TX

Based on previous experiences in recruiting NAEP ALS panelists for Writing 2011 (National Assessment Governing Board, 2012) and Science 2009 (National Assessment Governing Board, 2010), the estimate is that 20 percent of the nominators will respond by submitting at least one nominee for consideration. We further estimate that no more than 20 percent of the nominees would meet the qualifications, satisfy the requirements for representation, and agree to serve on the panel. Thus, the estimate is that 1,275 NAEP TEL-related organizations, companies, and other institutions of the types listed above must be identified and asked to provide nominations of panelists. A 20 percent response would yield 255 active nominators and at least 255 nominees. Assuming that 20 percent of those nominees will be eligible, meet the distribution requirements for representation on the panels, and be available/agree to serve as panelists, the yield would be the target of 51 panelists.

We expect higher response rates among educators in states that have technology and engineering literacy in their curriculum and/or have implemented new practices in line with the Next Generation Science Standards (National Research Council, 2013). Each nominator will be asked to make recommendations for up to four well-qualified nominees. The sample will be drawn to provide roughly equal representation of the four NAEP regions: Northeast, South, Midwest, and West.

Selection of Panelists

Nominees will be asked to complete an on-line questionnaire regarding their qualifications and experiences for serving on the panel. Candidates that have the credentials required of panelists will be contacted by phone to collect any missing information, verify the information provided, and confirm the willingness of the candidate to serve on the panel if selected. The goal is to select the most qualified panelists who are knowledgeable about TEL, while maintaining the goal to recruit 55 percent (28) teachers, 15 percent (8) non-teacher educators, and 30 percent (15) members of the general public to compose each of the panels. Panelists nominated in each panelist group must meet the following qualifications:

Teacher panelist:

- At least five years of overall teaching experience, and
- At least two years of experience teaching TEL in grade eight, and
- Judged to be “outstanding” in their professional performance by a nominator

Non-teacher educator panelist:

- Non-teacher educational staff at secondary schools with education and/or experience with TEL, or

- Curriculum director or content specialist at secondary school or state department of education with education and/or experience in TEL, or
 - Postsecondary technology and engineering faculty teaching introductory courses
- General public panelist:
- An expert in a technology and/or engineering company in one of the TEL-related areas (e.g., civil, environmental, agricultural, chemical, biomedical, electrical, mechanical, space and aeronautics, computer science), and
 - Not a former educator, and
 - Familiar with students in grade eight (e.g., as a parent or volunteer)

The credentials of panelists will be evaluated and scored based on the number and importance of the credentials that are presented. Persons having no distinguishing credentials will score low. Persons having extensive credentials, including having been named outstanding teacher/teacher of the year and/or being actively engaged at the national level in professional activities within the TEL subjects, will score very high. The scoring scheme differs for each panelist type (teacher, non-teacher educator, and general public). Persons with the highest scores are given top priority by placing the best qualified candidates at the beginning of the candidate list. The selection process then selects persons to reach the targets listed above, with persons having the highest qualifications being the first selected each time. All panels will be selected to have approximately equal proportions of males and females and equal proportions of persons from each of the four NAEP regions. We will also attempt to draw panels so that 20 percent of the persons self-identify as a minority.

Expenses for travel, meals, and lodging will be paid for all panelists in compliance with federal travel regulations. The goal is to schedule the meetings in a location that is convenient for

travel and is cost effective. In addition to covering the direct expenses for panelists (consistent with federal travel regulations), panelists will be given an honorarium of \$300 each to cover incidental expenses during their stay at the panel meetings. We will acknowledge that the funds available to offer panelists are not commensurate with their contribution. And, we will emphasize that their participation in the NAEP TEL ALS represents an exceptional contribution to technology and engineering education in the United States. Finally, school districts will be reimbursed for the cost of substitute teachers.

Section 4: Pilot Study

A pilot study of the ALS process will be implemented using the exact procedures planned for the operational standard setting session. The only difference planned between the pilot study and the operational ALS session will be the number of panelists.

Panelists in both the pilot study and the ALS meeting will be divided into subgroups, and each subgroup will be assigned a different subset of items with a subset of items common across the panelist subgroups. Panelists are divided into subgroups and items into subsets so that panelists can analyze item cognitive demands and complete standard setting item judgments in a reasonable amount of time. The subsets of items will be discussed in more detail later in the Design Document.

The TACSS, during the meeting on August 18 and 19, 2014, recommended that the panelists be divided into two or three subgroups. The pilot study will have 20 panelists with ten panelists assigned to each item rating group if the panel is divided into two subgroups or 15 panelists with five panelists assigned to each item rating group if the panel is divided into two subgroups. The operational ALS meeting will have 30 panelists with 15 panelists assigned to

each item rating group if the panel is divided into two subgroups or 10 panelists assigned to each item rating group if the panel is divided into two subgroups. The number of subgroups will depend on the feasibility of creating two or three similar subsets of items as described in Section 6: Achievement Levels-Setting Procedure.

The pilot study has two primary goals:

1. Determine whether modifications for training, instructions, materials, timing, and logistics will be needed for the operational ALS meeting
2. Provide an opportunity for facilitators to practice the process before moving to the operational setting

As already noted, the pilot study and the operational ALS study will follow the same methodology and procedure. A complete description of this common methodology and procedure is given in Section 6: Achievement Levels-Setting Procedure.

Section 5: Briefing Materials

This section describes the briefing materials that will be mailed to panelists prior to the pilot study and ALS meeting. Once the panels are selected, panelists will be sent letters inviting them to participate in the ALS process. When a panelist agrees to participate, another letter thanking the panelist for agreeing to participate and providing the panelist with information about the meeting dates and location, travel information, and other relevant information will be sent. The following information will accompany the letter:

- Description of the ALS process and draft agenda;
- Relevant Governing Board and NAEP brochures;
- Confidentiality Agreement;

- Reimbursement form(s);
- Request for Taxpayer I.D. Number and Certification (W-9);
- Travel instructions; and
- TEL Framework for the 2014 NAEP—complete and abridged
- NAEP TEL ALDs
- (http://nagb.gov/content/nagb/assets/documents/publications/frameworks/naep_tel_framework_2014.pdf; and <http://nagb.gov/content/nagb/assets/documents/publications/frameworks/tel-abridged-2014.pdf>).

The letter sent to the panelist will underscore the importance of the TEL Framework and the ALDs to the process and will urge panelists to review those two documents prior to the ALS meeting.

Approximately two weeks prior to the ALS meeting, an email and a letter will be sent to panelists that will provide more detailed information regarding logistics, hotel and city information, transportation to and from the airport, check-in procedures, and so forth. These briefing materials are intended to serve as a foundation for successfully carrying out the ALS process.

Section 6: Achievement Levels-Setting Procedure

This section outlines the ALS meeting, giving detailed information about the nature of the tasks and the procedures to be implemented. This section includes information about the configuration of panels and materials, training of panelists, the collection of panelists' judgments, and the feedback given to panelists.

Division of Panelists into Subgroups

The NAEP ALS Process has used a split panel design since the 1991 process. In past NAEP ALS meetings, the standard setting panel has been split into two rating groups. Because of the amount of time that may be required to review technology-based items, the standard setting panel in the NAEP TEL pilot study and ALS meeting may be split into three rating groups. These rating groups have loosely been referred to as replicate panels, because the panelists are assigned to groups to be as equivalent as possible to one another. The original reason for dividing the panelists into groups was to reduce the burden of the rating process by assigning approximately half of the items to each rating group. In addition, the two sets of panelists provided an approximate means for analyzing differences between the two rating groups, particularly for the first round of ratings. However, the TACSS, during the August 18 and 19, 2014, meeting, noted that the subgroups of panelists were only replicate groups in experience before the second round of the standard setting meeting. After receiving feedback before round two, the panels will no longer be independent. In this Design Document, these rating groups will be referred to as panelist subgroups.

The TACSS recommended that the panelists divided into three subgroups. For the pilot study, there will be 15 panelists with five panelists in each subgroup. For the operational ALS meeting, the 30 panelists will be assigned to one of three subgroups of 10 panelists each.

Each of the three panelist subgroups in the operational ALS will be further divided into two table groups of five panelists each for individual work and to facilitate table discussion. For the pilot study, the panelist subgroups will be the same as the table groups of five panelists each. The demographic attributes used to recruit panelists will be used when assigning panelists to

subgroups and table groups to maximize the equivalence of the subgroups as well as to maximize equivalence across table groups.

Division of Items into Subsets

The NAEP TEL item pool of 20 scenarios and 98 discrete items will be divided into four rating sets: A, B, C, and a common set of items. As was done under previous NAEP ALS procedures, items will be divided into item rating sets to limit the number reviewed by each panelist and minimize possible fatigue. The item sets A, B and C will be constructed to be as equivalent as possible. Items will be assigned to each rating set based on (a) assignment to one of the three subscales, (b) item type, and (c) item difficulty. Items will remain in the organizational units (blocks) used for administration of the assessment. The common item set will be constructed of blocks of items that have been selected for possible release to the public.

Item difficulty will be calculated for dichotomous items using each item's scale value for which a correct response probability of 0.67 was expected. Item difficulty will be calculated for each score point of a polytomous item where the probability of being awarded that score point was 0.67 or higher. The response probability of 0.67 is based on an Item Response Theory (IRT) model.

The common item set will be selected so as to serve the following purposes:

- Serve as examples in group discussions with panelists during the standard setting meeting;
- Provide a potential source of released items for the TEL assessment; and,
- Offer an empirical basis at round one to evaluate how well the groups were functioning as pseudo replications.

Given the multiple purposes that must be served by the common set of items, TACSS recommended that these items be selected so that items a) map reasonably well across the entire score scale, and b) represent the three TEL content areas. Scenario-based tasks selected for the common item set should consist of items that both map across that score scale and represent the three TEL content areas.

As shown in Figure 6.1, each of the three panelist subgroups will be assigned a unique set of items to review and rate. In addition, all three panelist subgroups will review and rate the same common set of items. But this design will only be followed if the item pool is found to support the construction of three item sets that are roughly equivalent in number of discrete item blocks and scenarios, representation across the three subscales, representation across item type, and median and range of item difficulty.

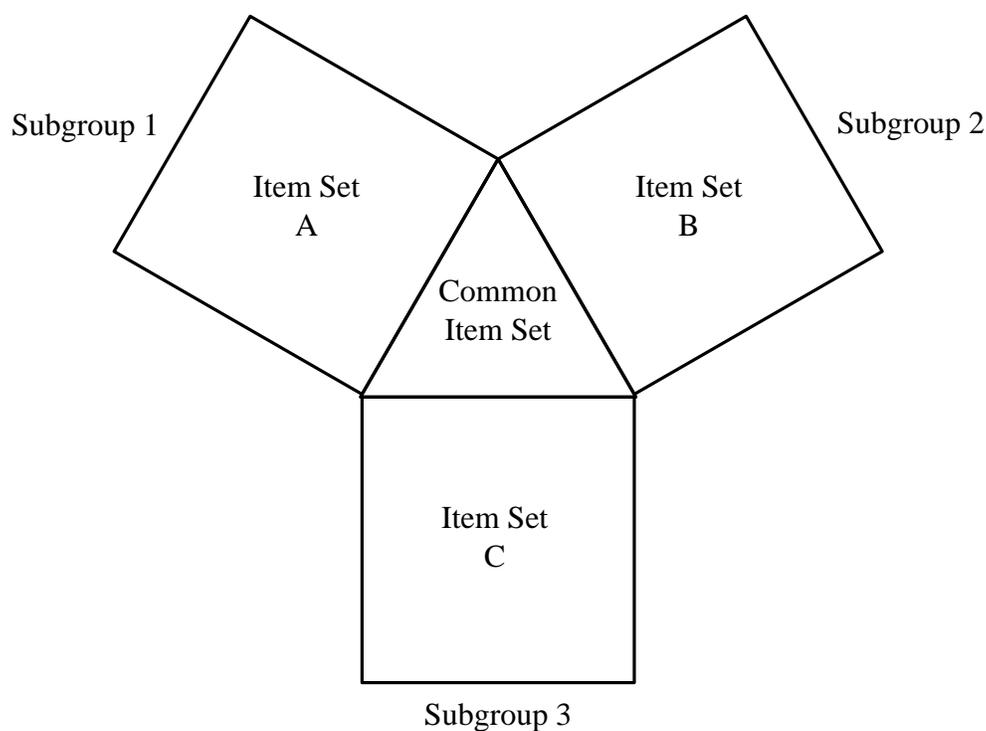


Figure 6.1. The assignment of item sets to panelist subgroups.

Use of NAEP-Like Scales

For this project, different NAEP-like scales will be used to avoid the risk of having the NAEP achievement level cut scores released before intended. As in past ALS studies, the NAEP-like scale will be a linear transformation of the NAEP reporting scale.

Training of Facilitators

The ALS meeting for the NAEP TEL assessment will involve a content facilitator and a process facilitator. The content facilitator is selected for TEL expertise and experience. The process facilitator is selected for expertise and experience conducting ALS meetings.

The content and process facilitators will be trained to implement the process as intended. We will prepare a PowerPoint presentation that facilitators will use during the ALS meeting. In addition, we will prepare facilitator handbooks that will include the tables and graphs, a script for providing instructions, a description of the activities, and an explanation of the feedback.

Facilitators will attend a one-day, web-based training prior to the pilot study. The project director overseeing the ALS activities will lead the training. In addition, the facilitators and the project director will do a walkthrough of the ALS meeting the day before the pilot study and the day before the operational ALS meeting.

Use of Computers

To make the process more efficient and strengthen the validity argument for the ALS outcomes, Pearson will use computers during the pilot study and the ALS meeting. The use of computers and software will reduce the time required for panelists to complete most steps in the

ALS activities. In addition, the use of computers will allow the panelists to interact with the items as a test taker did so that panelists are better able to understand what test takers would have to know or be able to do to correctly answer the item or achieve a score point.

Computers and software will be used in three different ways during the NAEP TEL pilot study and ALS meeting. For the first use of computers, panelists will take a form of the NAEP TEL assessment via computer to gain an understanding of the assessment as the student experienced it. Computers used in the test administration will be used by the panelists to take the test. For the second use of computers, a special software will be used to allow panelists to view and interact with all the NAEP TEL items outside of the testing context, both discrete and scenario-based. Both of these uses of computers and software will be done on the same computer. This computer will use software provided by a NAEP contractor for both computer uses.

For the third use of computers, computers will be used to collect ratings and present at least some feedback for the modified item mapping procedure. Pearson will be using software provided by Measurement Incorporated. This software will include the following functionality:

- Panelists will be able to make notes on items during the item review. A panelist will be allowed to make comments while reviewing all the items assigned to that panelist.

Comments will be stored with the item for that panelist and displayed only for that same panelist in the item map for the item regardless of the location and score point. For example, if the item appears in three locations on the map, the comment will appear at all three locations.

- Pearson administrators will be able to assign different sets of items to different panelists for item review and item rating activities. The system can be configured for two or three sets of items.
- The software will allow panelists' to review the borderline ALDs and make comments. A menu option will be available that allows the panelists to review the ALDs in html format so the panelists can highlight and comment on ALDs. The facilitator will be enable this menu option when they are ready. The comments and markup will be stored and the system will provide the ability to export the comments and markup for analysis and review.
- The software will record panelist ratings for the selection of potential exemplar items. Panelists will be presented with each item and the data for the item. Panelists will be able to make an independent judgment about the appropriateness of each item to serve as an exemplar for the achievement level to which it is assigned. Panelists will rate each item as “Should be Used,” “Might be Used,” or “Should not be Used.” The system will export this information through a file export.
- Pearson administrators will be able to export the data for panelist individually after a standard setting round is complete. Pearson will be able to load the results of analyses, such as descriptive statistic for each table and the panel, back into the system through a file upload to display the mean bookmark locations.
- Pearson administrators will be able to present table level summary statistics after each round. A table report will be available in the system. For each achievement level (Basic, Proficient and Advanced), the report will display the minimum recommended cut score, maximum recommended cut score, mean recommended cut score, median recommended

cut score, range of recommended cut scores and standard deviation of recommended cut scores for the table and will show the cut score recommendations for each panelist at the table. The cut score recommendation for each panelist will be associated with a panelist's unique identifier. The report will be accessed through a menu option. The menu option will only be available after each round has closed. A panelist will only have the ability to view the table report for their table. The facilitator will be able to view the results for any table by selecting the table number from the menu.

- Pearson administrators will be able to present room level summary charts after each round. The table reports will contain two tabs. The first tab will show the results for the panelist's table. The second tab will show the results for all tables and the room. The room level results will present for each achievement level (Basic, Proficient and Advanced), the minimum recommended cut score, maximum recommended cut score, mean recommended cut score, median recommended cut score, range of recommended cut scores and standard deviation of recommended cut scores for the room.
- Pearson administrators will be able to present rater feedback, in the form of a rater location chart, that will display the distribution of cut scores for panelists at end of a standard setting round. The rater location chart will only be available following a round. The specific round will be configurable during the event setup phase. The rater location chart will be a bar chart displaying every location and where each panelist placed their bookmark. If more than one panelist placed their bookmark on a location, the bar height will reflect that count. Each panelist will be labeled at a particular location in the chart when you hover over the bar. The chart may scroll horizontally since it may be wide. Panelists will access the rater location chart from a menu.

- Pearson administrators will be able to present consequences data, also called impact data, to each panelist. The consequences data will use a bar chart to display the percentage of test takers in each achievement level. The rater location chart will only be available following a round. The specific round will be configurable during the event setup phase.
- Pearson administrators will be able to present each panelist with an interactive bar chart that will allow the panelist to adjust recommended cut scores and see the consequences data that would result. The bar chart will be available after the close of round three. The panelist will be able to manipulate the bar chart by dragging the bars or some other draggable indicator. As the panelist moves the cut score bar, the bar chart will show the percentage of students that would be in an achievement level. Panelists will be able to enter a new cut score value for each achievement level based on the consequences data from manipulating the bar chart.

Panelists will work with two computers. Using the first computer, a form of the NAEP TEL will be administered so that panelists can take the assessment and the scenario-based and discrete items will be available so that panelists can interact with the items in their assigned item subset. Using the second computer, feedback will be displayed to panelists and panelists' ratings will be captured and recorded. Our preliminary plan is that the computer used to capture panelists' ratings will be linked to a main computer or server via secure, wireless internet connection, to facilitate recording, aggregation, and analysis of panel data. Alternatively, files will be captured on media and carried to the work room.

Procedure for Achievement Levels-Setting

An agenda for the pilot study and ALS meeting is shown in Appendix A, titled “Draft Agenda for the Pilot Study and ALS Meeting.” For the NAEP TEL assessment, Pearson has proposed using a criterion-referenced, content-based modified item-mapping method. In the item mapping methodology (Lewis, Mitzel, Mercado & Schulz, 2012) items are arranged along a continuum using IRT-based item difficulty estimates. Standard setting panelists then recommend performance level thresholds or cut points by identifying items on that continuum that test takers representing different performance levels should be able to answer correctly with a given level of probability. Pearson proposes modifying the item mapping methodology in the following ways to be consistent with item mapping implementation in previous NAEP ALS studies (ACT, Inc., 2007; ACT, Inc., 2010) and the Judgmental Standard Setting Studies (WestEd, 2011):

- Items will be presented and panelists’ responses will be collected by computer.
- A graphic display of items on a NAEP-like student achievement scale, an item map, will be given to panelists to accompany the OIBs used to place the cut scores. The item map that will be presented will follow a format similar to item maps used in past NAEP ALS meetings. The item map will be presented on the computer as part of the standard setting software. A paper copy will also be given to the panelists.

Reviewing the Assessment

To begin the meeting, the process facilitator will lead introductions. The introductions will conclude with an overview of the meeting agenda. Panelists will then take a form of the NAEP TEL assessment consisting of one or more scenarios and a block of discrete items under

conditions similar to those of actual student testing in 2014. The specific scenario(s) and block of discrete items used for this exercise will be those that are recommended for release. In addition to their use when panelists take the NAEP TEL, these blocks of items will be common to panelists in each of the three item rating groups and they will be used for the selection of exemplar items. Taking the NAEP TEL form will be panelists' first exposure to the items in the NAEP TEL assessment.

Panelists will take the TEL assessment form using the same computers and the same interface that were used by the grade eight students who originally completed the NAEP TEL assessment. The purpose of this activity is to give panelists an opportunity to experience the assessment as the test taker experienced it. Panelists should become familiar with the content and rigor of the assessment.

After completing the test, panelists will be trained in how to use the scoring rubrics for constructed-response items. Following training on the rubrics, panelists will receive scoring guides and time to score their own responses. A scoring key will be provided that contains the item type, points possible, and the item key.

The content facilitator will then provide an orientation to the NAEP TEL Framework. The orientation to the NAEP TEL Framework is intended to provide panelists an understanding of what students should know and be able to do as measured by the NAEP TEL assessment.

Reviewing the Achievement Level Descriptions

Before the item review and round one item ratings, the process facilitator will provide an overview of the purpose of ALS in general and a description of the process to be used in setting the standards. Next, the content facilitator will review the purpose, meaning, and structure of the ALDs and instruct the panelists to read the ALDs, think about the progression in TEL knowledge

and skills represented by the three levels, and underline key terms in each level. The discussion will demonstrate how the policy definitions of Basic, Proficient, and Advanced are operationalized as the ALDs. The panelists should be able to see clearly the connection between “superior performance” and the Advanced ALD. The content facilitator will ask panelists to discuss within their tables the differences in knowledge and skills they see as the levels move from Basic to Proficient. Finally, the content facilitator will engage the entire group in a discussion of the understandings coming from each table group. Each table group will be asked to summarize the discussion at their table so that the entire group can reach agreement on the meaning of the levels.

Reviewing Items

To set cut scores on an assessment, panelists must have a good understanding of the assessment and the knowledge and skills it requires students to demonstrate to earn successively higher scores on the assessment. In preparation for item ratings, panelists will review items in their assigned item rating pool to identify what test takers need to know and be able to do to respond correctly to the selected-response items or to score at each credited rubric score level for the constructed-response items.

The item review is an important activity, but it can be lengthy and tiring for panelists. Panelists will review all the items from the common item subset. In order to reduce the potential for panelists’ fatigue and to reduce the amount of time required for this task, panelists will be assigned only some of the items from the item subset assigned to their panelist subgroup. Each item from an item subset will be assigned to two or three panelists at the same table and from the same panelist subgroup. These two or three panelists will allow for contrasting analyses of item

cognitive demand. Each panelist will have an opportunity to interact with each item in their item subset during table group discussions.

Item review will be completed in two stages, similar to earlier ALS meetings (ACT, Inc., 2007; ACT, Inc., 2010). Initially, panelists will be trained to review just the scenario-based tasks and the discrete constructed-response items. Having panelists review the knowledge and skills of items in the context of the scenario-based task and discrete constructed-response items by themselves, rather than by score point as in the OIBs, will give panelists the opportunity to interact with each scenario-based task and constructed-response item as a whole and not with one score level at a time. For this activity, panelists will be given constructed-response OIBs, which will be available on the computer. In this constructed-response book, scenario-based tasks and constructed-response items will be presented in their entirety rather than by score-point. Copies of scoring rubrics and exemplar responses will be presented on the computer and as paper copies.

Panelists will be instructed that the scenario-based tasks include both selected-response and constructed response items. The panelists will be instructed to work through each scenario and review the constructed-response items as they are encountered. The panelists will be instructed to ignore the selected-response items.

The content facilitator will model the item review task for the first three or four constructed-response items from the common item set and from a single scenario if possible. Next, panelists will work individually to review the items. Panelists will describe the TEL knowledge and skills test takers need to know and be able to do to receive full credit on the item. Panelists will then describe the knowledge and skills needed to earn partial credit for the item by writing descriptions for each of the successively lower credited responses for the item. These notes will be recorded and stored in the standard setting software. The software associates these

notes with the item and the panelist so these notes will be available whenever the panelist views the item or an item score point.

Second, panelists will work independently to analyze the knowledge and skills required by all the items in their rating pool in the context of their OIBs. Panelists will consider items sequentially, beginning with the first, or easiest, item. During this independent review of the OIBs, panelists will be asked to make notes on what students need to know and be able to do to answer each selected-response item correctly. These notes will be recorded and stored in the software. A panelist's notes will be available to the panelist during the item rating process.

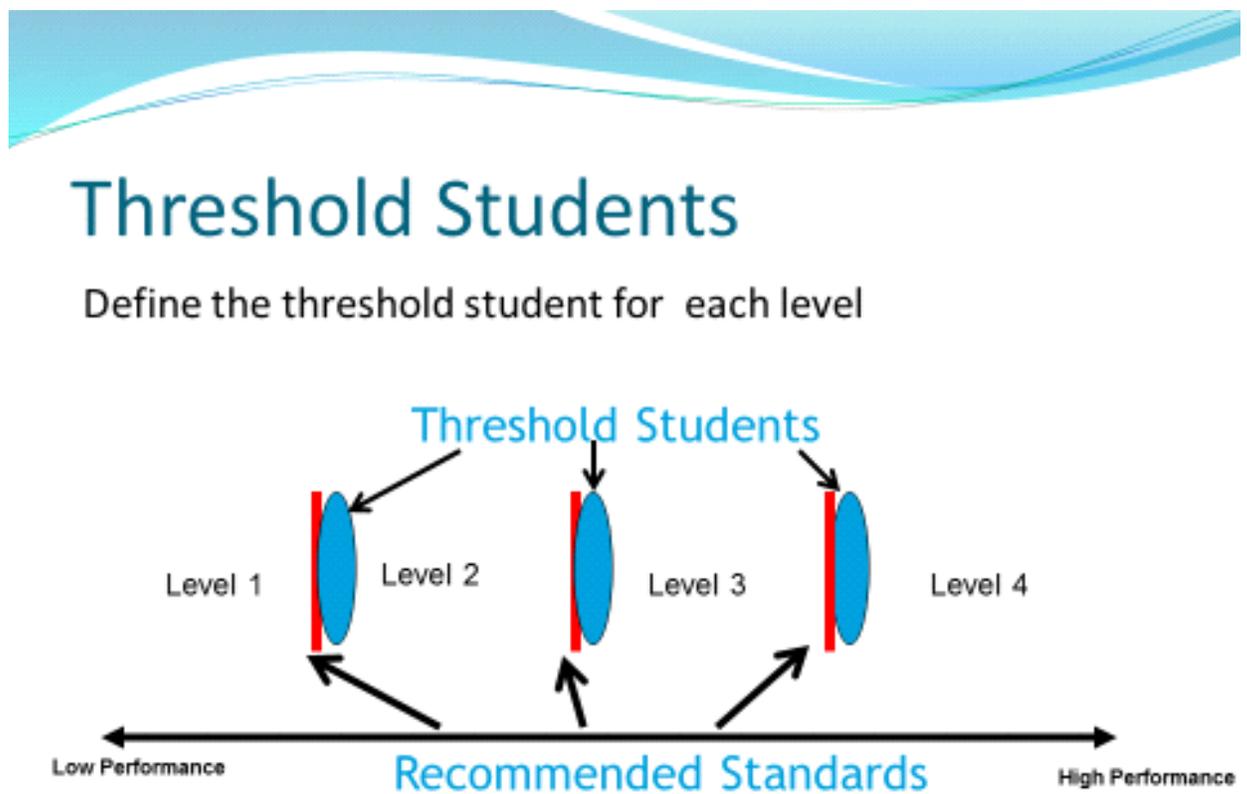
This review will allow panelists to become familiar with the progression of difficulty from one item to the next within their OIBs. Panelists will check off the selected-response and constructed-response items on their item map as they progress through the OIBs. The item check-off process will help panelists see how much more difficult one item is than another and how the increase in difficulty is related to what students need to know and be able to do in the TEL content area.

Next, panelists will review with their partner at their table the TEL knowledge and skills associated with each item or score point in the OIBs. Items will be considered sequentially beginning with the easiest items. Panelists will be asked to share their thoughts with the panelists at their table about the knowledge and skills needed to answer each item correctly or achieve each score point.

Development of Borderline Achievement Level Descriptions

The next step is intended to help panelists focus at the borderline of each achievement level and introduce the borderline performance descriptions. Following the recommendation of

the TACSS made during the August 18 and 19, 2014 meeting, borderline performance descriptions will be developed in both the pilot study and the ALS meeting. The borderline performance descriptions are the descriptions of the ALDs that are represented on the NAEP TEL scale by each cut score. The borderline performance descriptions will be developed after the NAEP TEL Framework has been introduced and the ALDs have been reviewed. Panelists will be told to think of the lower borderline in terms of a student who is just qualified to be in the achievement level. A slide typically used to introduce the idea of a borderline student is shown in Figure 6.2.



8

Figure 6.2 A slide used to introduce the idea of a borderline student.

Initially, panelists will work in table groups. A table representative will be elected or selected for each table, and the table representative will report on the discussions at each table. First, panelists will be asked to consider the NAEP TEL Framework and the ALDs and describe the minimal achievement a student must have to be considered Proficient. Next, panelists will be asked to describe the minimal achievement a student must have to be considered Basic and then Advanced. Table representatives will record descriptions of the borderline performance descriptions for each table using flipcharts. Descriptions may be in a narrative form, bulleted lists, or even key terms or phrases. Descriptions will be collected by the facilitators and key entered by Pearson representatives.

The process facilitator will then work with the panel to develop borderline performance descriptions with the panel. Table representatives will share the table-level discussion with the panel. The process facilitator will then work to create a set of borderline performance descriptions for the entire panel. First, the process facilitator will work with the panel to describe the minimal achievement a student must have to be considered Proficient. Next, the process facilitator will work with the panel to describe the minimal achievement a student must have to be considered Basic and then Advanced. The process facilitator will key enter these descriptions on a computer and project the descriptions on a screen that the panel can view.

Pilot study and ALS meeting panelists will be asked to review the draft borderline performance descriptions prior to each round of collecting content-based judgments.

Placing Round One Bookmarks

Panelists will be trained on using the software, and they will be provided with an item map both on the computer screen and in paper form. As recommended by the TACSS during the

August 18 and 19, 2014 meeting, this item map will also be revised so that item subsets can be identified by panelists. The item map will use colors to identify the different item subsets assigned to the different panelist subgroups. On an item map, items will be ordered from easiest at the bottom to hardest at the top. Constructed-response items will be displayed once for each score point. The score scale at which the item has a 0.67 probability of a correct response will be used to map the items. After each round, panelists receive a new version of the item map with the panel's median scores marked. The updated item maps enable panelists to consider whether a different cut score location from their round one recommendation would represent their (new) understanding of the borderline performance descriptions.

An illustration of an item map used in the 2009 standard setting for NAEP Science grades four, eight, and twelve (ACT, Inc., 2010) is shown in Figure 6.3. The items are separated into content related columns. Each item is represented on the map by a unique identifier consisting of a character followed by a number. The first digit of the unique identifier represents item type (C = constructed response and M = multiple choice). The number following the character represents where that item falls in order of difficulty within type. Extended constructed response items include an underscore “_” followed by the score level. Short constructed response (or dichotomous) items only have one score level so their unique identifier does not include a dash and number. The color of an item unique identifier on the map indicates whether the item is in the Group A pool only (tan), the Group B pool only (green), or in both item pools (yellow).

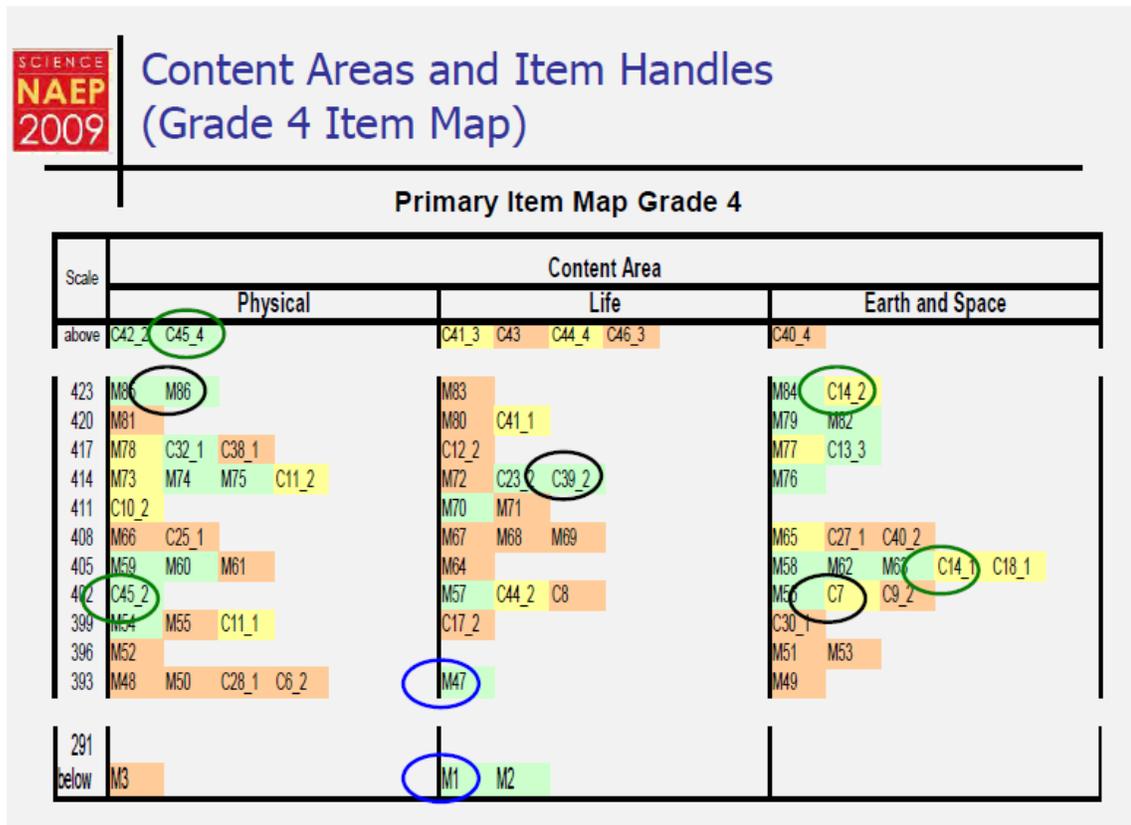


Figure 6.3. An illustration of an item map (ACT, Inc., 2010).

Initially, the process facilitator will introduce the OIB on the computer and explain that items and item score points are arranged in order from the easiest to the most difficult. The process facilitator will explain that the empirical item difficulty of the item, based on the 2014 administration of the NAEP TEL assessment, was used to order the items. The panelists' judgments of item difficulty may not perfectly reflect the order of the items in the OIB. Figure 6.4 shows an illustration that will be used to explain the OIB.

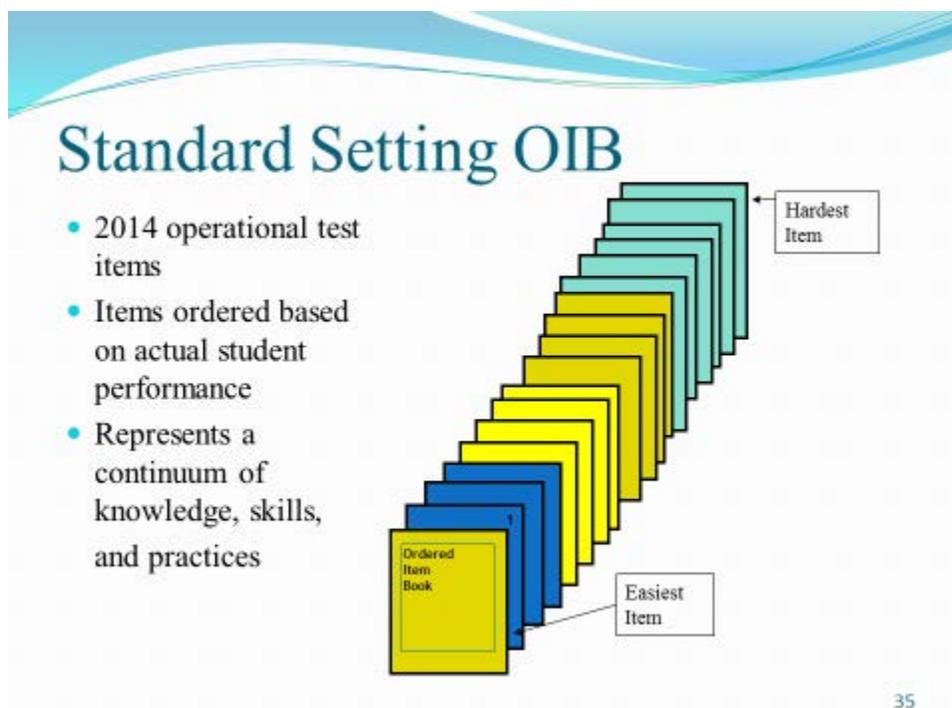


Figure 6.4. A slide used to illustrate the ordered item book (OIB).

Next, the process facilitator will describe the cut score placement task to panelists. The facilitator will explain that each bookmark should represent the place in the OIB such that a minimally competent or threshold student for a given level should correctly answer the items at or before the bookmark. The threshold student should have a 0.67 chance or greater of correctly answering the item at or before the bookmark, and a less than 0.67 chance of correctly answering the items after the bookmark. Panelists will be cautioned that they should continue making judgments past the place in the OIB where they initially place their bookmark because their judgment of item difficulty may not always agree with the order of item difficulty in the OIB.

Item mapping placements will be completed one achievement level at a time starting with Proficient, then Basic, then Advanced. Panelists will be instructed to read the ALDs and the borderline performance descriptions and use their understanding of the borderline performance

descriptions for the given level to place their bookmark for that level. Panelists will be told to place their bookmarks independently and without discussion with their table group.

After all panelists have placed their Proficient bookmarks, they will place their Basic and Advanced bookmarks individually and at their own pace. After placing all bookmarks, panelists will be given an opportunity to adjust their item mapping placements. Panelists will use the software to record their bookmarks.

Placing the bookmark using the software will automatically store each panelist's selected cut score in the database. As a precaution against data loss, panelists will also be asked to document the item identifier by recording the location of their bookmarked items on their paper item map. Note that panelists' cut score recommendations will be computed from the panelists' bookmark placements as described in Section 7: Data Analyses and Results Presentation.

Before beginning the actual item mapping task, panelists will be asked to practice the item mapping task on the computer. To practice, panelists in each rating group will use a subset of 10 items from the set of items assigned to the other item rating group. This will allow panelists to practice the item mapping tasks without exposing panelists to items that will be in their item rating pool to judge for making their cut score recommendations. Before each round, panelists will be asked to complete a computer-based survey on the success of panelist training to undertake the ratings task. A paper-based example of a survey of the success of panelist training to undertake the ratings task is shown in Appendix B. These questions or similar questions will be used in the computer-based survey. Before the panelists begin rating, facilitators will review the responses and offer additional help for any panelist who indicates a lack of preparedness.

Note that as the agenda shown in Appendix A describes, three hours are scheduled for panelists to complete the first round of ratings. Further, the first round of ratings is scheduled at the end of the second day of standard setting so that additional time can be allowed to panelists who need it.

At the end of the second day, panelists will complete an evaluation of the first round of standard setting. Example questions for the evaluation of the first round of standard setting are shown in Appendix C.

Placing Round Two Bookmarks

Following the collection of round one ratings, panelists will receive feedback intended to encourage them to reflect on the knowledge and skills, as described in the ALDs, required by test content near the cut scores from round one. Feedback will be based on the median cut score computed from the cut scores across all panelists. The panel will receive the following two different kinds of feedback:

1. Rater location charts
2. Item maps

Using the computer software, panelists will receive a rater location chart after the first round of item mapping that will display the distribution of cut scores for all panelists. In the software, panelists will be identified using codes to ensure confidentiality. The rater location chart will also display the median cut score for the panel. An example of a rater location chart that served as the model for software develops is shown in Figure 6.5. Following the advice of the TACSS during the meeting on August 18 and 19, 2014, the rater location chart will be color coded so that panelists can identify the three panelist subgroups.

Panelists will be instructed to evaluate their individual cut scores relative to other panelist cut scores and to the median cut score to help determine whether their conceptualizations and understandings of the borderline performance descriptions differed from those of others in the group. If so, the panelists will be asked to consider whether their understanding of the ALDs and borderline performance differs from others in the group, if there was a mistake in recording the bookmark/cut score, or if the panelists want to reconsider the bookmark placement in round two, in light of all the feedback received from round one and the ALDs.

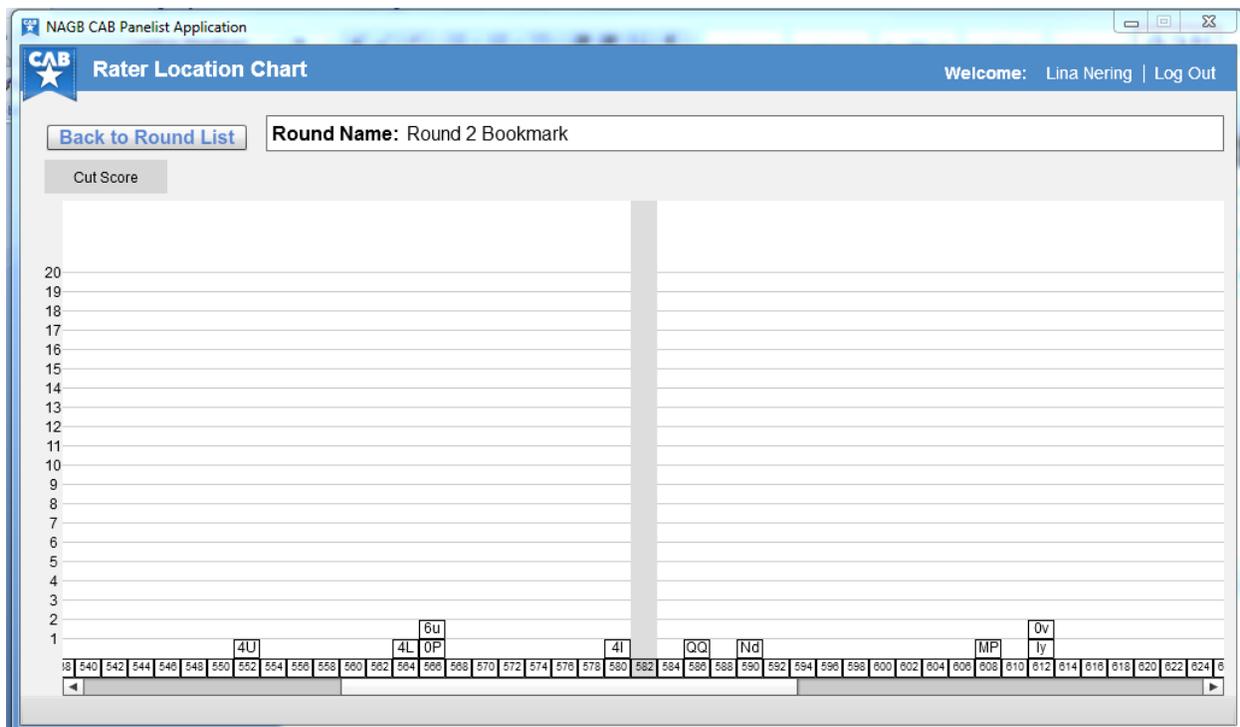


Figure 6.5 The Rater Location Chart Used as a Model for the Software Developers.

In addition to the numerical values of the cut scores, panelists will be given a new version of their item map with the panel cut scores marked on the map. Panelists will be instructed to compare borderline performance descriptions to the knowledge and skills of items around their

individual cut scores and the knowledge and skills of items around the panel median cut score. Panelists will be asked to determine if another location for their bookmark would better represent their understanding of the borderline performance descriptions than their current cut score.

Before starting round two ratings, facilitators will remind panelists of the steps in the item mapping task and the importance of the ALDs. Panelists will then be asked to complete a computer-based survey on panelist training to undertake the round two ratings task.

Panelists will use the software to record their round two bookmarks. Placing the bookmark using the software will automatically store each panelist's selected cut score. The group cut score will be computed by selecting the median cut score for each achievement level.

Placing Round Three Bookmarks

Initial feedback from round two will consist of the median cut scores, the cut score distribution, and the location of panelists' cut scores relative to the median cut scores. Panelists will also be given a new item map.

Next, facilitators will present panelists with the percentage of students within and at or above each achievement level. This feedback has been called consequences data in past NAEP ALS meetings, and it has also been called impact data. These percentages will be based on the grade eight NAEP TEL assessment operational results and the round two median group cut scores. This feedback will be presented in the software, which includes functionality that allows the percentage of students at or above the cut score and the list of items mapped below and above the cut score to change as panelists adjust the placement of the cut score.

The process facilitator will then lead panelists in a discussion of the consequences data and ask them to consider if the consequences data seem reasonable given the knowledge and

skills described in the ALDs, the borderline performance descriptions, and what panelists know about the distribution of knowledge and skills described in the NAEP TEL Framework. The facilitator will remind panelists that consequences data represent the distribution of a nationally representative sample of test takers relative to each achievement level. The panelists should attempt to take a national, rather than a local, perspective. And, they should make their evaluation relative to the ALDs, stating what students *should* know and be able to do.

After presentation of round two feedback, panelists will be given an opportunity to adjust their cut score recommendations. The process facilitator will instruct panelists to carefully consider round two feedback, the ALDs, and discussions with other panelists. Then panelists should independently determine if their round two cut score recommendations should be changed. The facilitator will instruct panelists that if they do decide to change any of their round two recommendations, panelists should review the items in their OIBs and the item map to identify the knowledge and skills required by items near the new cut scores. Changes should be supported by the knowledge and skills required by the items and described in the achievement level descriptions. Panelists will then record their cut score recommendations using the software, as well as on their item maps.

After round three, panelists will be provided their final cut scores, the distribution of cut scores, and a new item map with the final cut scores based on median cut scores from the round three recommendations.

Completion of Consequences Questionnaire

Following round three, panelists will be provided with consequences data based on their round three ratings. The consequences data graphically displays both the percentage of grade eight test takers from the operational administration of the NAEP TEL that score at or above

each achievement level and the percentage within each level. Based on recommendations from the TACSSS at the August 18 and 19, 2014, meeting, Pearson will adapt an existing questionnaire used in past NAEP standard setting meetings to collect the panelists' opinions regarding the consequences of their panel's cut scores. The TACSSS suggested that questionnaires from past NAEP standard setting meetings could serve as benchmark against which to compare results from the pilot study and ALS meeting.

Pearson proposes that for this activity panelists be able to use the software that includes functionality that allows the percentage of students at or above the cut score and the list of items mapped below and above the cut score to change as panelists adjust the placement of the cut score. The questionnaire will ask panelists if they want to make changes to any of the cut scores after considering the final round three consequences data. Panelists will be able to recommend a different cut score to represent each achievement level for any or all three cut scores using the software to explore the implications for consequences data of different cut score recommendations. Panelists will be reminded that the cut score recommendations should represent fidelity between the borderline performance descriptions and the items at the cuts.

Selection of Exemplar Items

Following the final round of ratings, panelists will rate the potential exemplar items as to whether they should be used to illustrate the NAEP TEL assessment. Prior to the ALS meeting, the National Center for Education Statistics will be asked to identify blocks of released items. More information on selection of exemplar items is provided in Section 8: Exemplar Tasks and Items/Responses.

Procedural Validity Evidence

A number of writers (Cizek, Bunch, & Koons, 2004; Hambleton, 2001; Kane, 1994, 2001) have concluded that validity evidence for ALS include procedural and external evidence. Procedural evidence refers to the appropriateness of the standard setting procedures and how well those procedures were implemented.

Evidence for procedural validity may come from a number of sources, including criteria for selecting panelists, the justification for the ALS method, the quality of the implementation of the procedure, and the completeness of the documentation of the process (Sireci, Hauger, Wells, Shea, & Zenisky, 2009). All of those sources for procedural validity evidence have been addressed in this proposal. As another source of evidence of procedural validity for the NAEP TEL assessment, panelists will be asked to complete evaluation forms after each round of standard setting and each major activity of the ALS process. Evaluations will include both selected-response and open-ended questions that address the panelists' understanding of the process and confidence in the results.

The selected-response answers on the evaluations will be collected each day from panelists using the computer. The written responses will be scanned by the facilitators for possible problems as they are collected during each day. Any evaluations completed at the end of a day will be immediately processed by Pearson staff. Summary statistics will be computed for all ratings items, and written responses that signal any problems will be summarized. These analyses will be reviewed at the end of each day, and any sources of confusion will be identified for clarification with individual panelists or the panel.

Section 7: Data Analyses and Results Presentation

In this section, we present plans for data analysis during the pilot study and ALS meeting. In addition, we outline a strategy for examining and displaying results from the pilot study and ALS meeting to the Committee on Standards, Design and Methodology (COSDAM) and the full Governing Board so that they can make decisions on recommended cut scores.

Analyses of Data

We will conduct all analyses using the raw data that panelists provide (e.g., bookmarked pages, responses to readiness surveys at each round, questionnaires after each round on understanding of the process and confidence in the results) and, when appropriate, using NAEP-like scale scores that are a linear transformation of the NAEP reporting scale. We will complete the same analyses for both the pilot and ALS meetings, within and across panelist subgroups, and compare results from the pilot and ALS meetings.

The Pearson data analyst will complete all analyses using Excel workbooks and SAS code. All data capture procedures and analyses will be programmed, tested, and certified before we convene the pilot study.

The panelist cut score for an achievement level will be computed from the recommendation of a panelist. The panelist cut score for an achievement level will be computed as the theta value that is the midpoint between the theta value of the item bookmarked by the panelist and the next item in the OIB. The rationale for this computation is based on the instructions given to panelists. The panelists are instructed to bookmark the last item for which a borderline student would have a 0.67 probability or greater of answering correctly or achieving the score point. The theta value that would separate two achievement levels, Proficient from Advanced, for example,

would be higher than the theta value for the item bookmarked by the panelist but lower than the theta value for the next item in the OIB not bookmarked by the panelist.

Central Tendency of Cut Scores

We will calculate means and medians of recommended cut scores for each round, each panelist subgroup, and across all panelist subgroups. We also will calculate the percentage of test takers at each achievement level for each round within and across panels. We will use these measures to document panelist recommendations and as the basis for subsequent analyses, which we describe below.

Changes in Cut Scores

Standard setting is a convergence process in which individuals make informed judgments, where some of the information includes insights and comments of other panelists. One indicator of the degree of convergence of panelists' views and the influence of feedback between rounds is the amount of change in cut scores from round to round. We will calculate the number and percentage of cut scores that increased, decreased, and remained unchanged across rounds. In addition, a measure of the variability of bookmark placements within a panel is the mean absolute deviation, which is computed as the average difference between each panelist's cut score and the panel's median cut score. We will provide a plot of the mean absolute deviation of cut scores of individual panelists from the table cut score and group cut score by round.

Standard Error of Cut Scores

The standard error of a cut score is an estimate of the uncertainty in the reported cut score due to various sources of error. The reliability of cut scores emerging from a standard setting process is typically thought of with regard to how consistent the cut scores are across tables, panelist subgroups, and panelist type. This consistency across groups is used as an estimate of reliability. Due to the difficulty surrounding calculating the standard error of a median, we will report two nonparametric standard errors procedures that have been used in past NAEP ALS studies: an empirical-based method and a bootstrap method. The first estimator of the standard error is based on the Maritz-Jarrett procedure (Maritz & Jarrett, 1978). Let $X = X_1, \dots, X_n$ be the sample of size n and $X_{(1)} \leq \dots \leq X_{(n)}$ are order statistics of this sample. The process of computing this statistic involves computing a weighted sum of the order statistics of the sample at hand, with weights based on incomplete beta ratio. The values of the weights are distributed in such a way that the most weight is given to the values from the sample that are involved in computing sample median, and progressively smaller amounts are given to the rest of the order statistics as their values are further away from the center.

The second estimator of the standard deviation of the median is based on the bootstrap technique (Efron & Gong, 1983). In this procedure, repeated samples with replacement are taken from the original distribution of cut scores and the median is calculated for each resample. The standard deviation of these medians is then calculated and used as the standard error estimate. We will sample 1,000 cases to compute the bootstrap statistic.

Reported standard errors for rounds two and three of standard setting should be interpreted with caution. Panelist judgments for these rounds are no longer independent due to the interactive nature of the standard setting process. Panelist judgments for rounds two and three

are influenced by the location of the cut scores for the other panelists, other feedback, and panelists' discussions that are part of the standard setting process.

Analysis of Means

No satisfactory method exists for estimating the significance of the differences between groups on their median cut scores. But if the mean and median cut score are similar, then Pearson recommends that analysis of the effects on means be performed using an ANOVA procedure. Means are amenable to analysis using ANOVA whereas medians are not. We will analyze cut scores for differences across panelist variables such as gender and race/ethnicity and process variables such as table group and replicate group. If the mean and median cut score are dissimilar, then Pearson does not recommend using this analysis.

We will report the F-values and the associated *p*-values for rounds one and three and for each of the panelist and process variables in the analysis. Given the number of comparisons made and the small sample size for some of the groups, we likely will find a few statistically significant results. Differences that are statistically significant will be reported along with the associated differences in medians to help judge the substantive meaning.

Strategies for Results Presentation

In considering the recommended cut scores from the achievement level standard setting process, Governing Board members are likely to be most concerned about two things: Are the recommended cut scores and consequences data reasonable? Was the achievement level process rigorous so that panelists were able to make solid, content-based recommendations informed by

NAEP consequences data? With these two questions in mind, we make the following recommendations.

Strategy for Examining Results with COSDAM

This proposed strategy is based on the idea that COSDAM is the technical subgroup of the Governing Board, includes some members with significant psychometric and technical expertise and that they will be interested in and are responsible for developing detailed and well supported responses to the two questions above on behalf of the full Board. A recommended sequence of steps for the COSDAM presentation includes:

1. Provide a high level overview of the modified item mapping process.
2. Next, present final recommended cut scores with NAEP consequences data.
3. Provide summaries of the feedback data (from above) for each round.
4. Provide context for these recommendations using summary information from other subjects.
5. Provide summaries of panelist responses to all questionnaires, as part of considering procedural validity.

Strategy for Examining Results with the Full Governing Board

This proposed strategy is based on the idea that the Governing Board is composed primarily of policymakers. The proposed strategy assumes the Governing Board will be interested in a well-reasoned but nontechnical argument addressing these two questions: Are the recommended cut scores and consequences data reasonable? Was the achievement level process rigorous so

that panelists were able to make solid, content-based recommendations informed by NAEP consequences data? A recommended sequence of steps for the full Board presentation includes:

1. Provide a high level overview of the modified item mapping process.
2. Next, present final recommended cut scores with NAEP consequences data.
3. Provide context for these recommendations using summary information from the other NAEP subjects.

Information presented to COSDAM can be used to respond to questions and concerns expressed by Board members.

Section 8: Exemplar Tasks and Items/Responses

Exemplar items are a part of the official set of information that Pearson is to recommend to the Governing Board for setting achievement levels. Exemplar items serve to communicate to the public the types of knowledge, skills, and abilities that are required for performance within each of the three NAEP achievement levels. The role of exemplar items in communicating performance on the NAEP TEL is especially important because this is an entirely new, innovative area of assessment for the NAEP program. The items selected to illustrate performance at each achievement level will illustrate the way technology and engineering literacy is assessed by NAEP, as well as illustrating the performance required for each level of achievement. Fidelity with the ALDs is the most important criterion for selection of exemplar items to illustrate the achievement levels. Student performance on the item must demonstrate the knowledge, skills, and abilities that align with those in the ALD for the level it represents.

The items in the block(s) marked for public release by NCES (or recommended for public release) will determine the set from which exemplar items may be selected. The released block(s) will include at least one scenario consisting of a set of items related to the scenario. As previously described, the block(s) marked for release will be common to panelists in each of the two or three panelist subgroups when item rating subsets are developed. This ensures that all panelists are familiar with all items to be considered for exemplar status.

Selection of exemplar items is the last of the judgments that panelists are asked to make in the ALS process. This procedure is implemented after collecting the panelists' final cut scores. By this time in the process, panelists are very familiar with both the achievement level descriptions and the items. They are well prepared to make the judgments regarding items that will serve to represent the achievement levels. They will be instructed in the purpose of the task and the statistical criteria that were used for presenting the items for their consideration. Further, they will be given information regarding the criteria Pearson will consider for selecting items from the panelists' recommendations to be used in reporting student performance relative to the achievement levels. They will be instructed in the use of the statistical information to evaluate the items for judging those that would best serve as exemplars. They will be told, however, that the most important consideration is the relationship between the achievement levels descriptions and the knowledge, skills, and abilities assessed by each item.

All items that have a response probability of 0.67 at a score point within each achievement level range will be presented to panelists to judge their appropriateness as exemplar items. The correct response will be given for multiple choice items; the scoring rubric and a student response scored at the relevant score point will be presented for each item score value. Items will be assigned to the lowest achievement level at which the 0.67 response probability is

attained. The items will be presented to panelists to show the item identification number, the subcontent area assessed by the item, the location of the item in the ordering presented for the rating process, the overall response probability (p -value), the response probability at the cut score of each achievement level, and the scale score at which RP.67 is reached.

Panelists will be asked to review each item and the data for the item, discuss the items with their table mates, and then make an independent judgment about the appropriateness of each item to serve as an exemplar for the level to which it is assigned. Panelists will rate each item as “Should be Used,” “Might be Used,” or “Should not be Used.”

Exemplar items will be used in reporting the results of student performance on the NAEP TEL relative to each of the three levels of achievement (and below Basic) and for each of the three major content components of the NAEP TEL: technology and society, design and systems, and information and communication technology. Given the somewhat limited set from which exemplar items may be selected, the goal is to maximize the number of items recommended to the Governing Board. To the extent possible, given the specific items in the released block(s), the following criteria will be used for the selection of items to recommend to the Governing Board as exemplars for each achievement level:

1. The items in the scenario(s) marked for release should range in difficulty so that they map across the score scales and represent performance at each of the three achievement levels.
2. There should be a mix of items across the subcontent areas, with at least one item from each of the three subcontent areas.
3. There should be a mix of items of each item format type, e.g., multiple choice and constructed response.

4. Items with an average probability of a correct response for students in an achievement level near 0.25 for that achievement level will be given priority for selection to represent the minimal level of performance required for each achievement level. Items with an average probability of a correct response for students in an achievement level near 0.50 will be given priority for selection to represent the mid-range, solid performance within the achievement level.

Following the procedure reported in ACT, Inc. (2010), the average probability of a correct response for students in an achievement level for item i and score level h is calculated as

$$\text{Average probability} = \frac{\sum_{j=C_L}^{C_{L+1}-1} \Pr(U_i \geq h|j) * f_j}{\sum_{j=C_L}^{C_{L+1}-1} f_j} \quad (\text{Equation 1})$$

where j is a scale value, C_L represents the cut score for the achievement level, C_{L+1} is the cut score for the next higher achievement level, and f_j is the number of students scoring at scale value j . For the advanced level, C_{L+1} will be set to the highest possible scale value plus 1. The values in the numerator and denominator of equation X can be calculated as a function of the cumulative expected probability and the cumulative distribution function. The cumulative expected frequency (CEF) at scale point k is defined as

$$CEF(k) = \sum_{j \leq k} \Pr(U_i \geq h|j) * f_j \quad (\text{Equation 2})$$

and the cumulative frequency is the cumulative distribution function of the student estimates,

$$F(k) = \sum_{j \leq k} f_j \text{ (Equation 3).}$$

5. Items with the highest frequency of panelists' ratings as "Should be Used" and lowest frequency of panelists' ratings as "Might be Used" will be given priority. Items rated as "Should not be Used" by 10 percent or more of the panelists will be eliminated from further consideration unless it is necessary to represent a particular feature of the assessment at a specific level of achievement.
6. In order to provide the maximum amount of choice among items to be presented as exemplars in the Nation's Report Card, all items rated as "Should be Used" or "Might be Used" by at least 50 percent of panelists and that meet the other criteria will be considered for recommendation.

Items under consideration for recommendation to the Governing Board as exemplars will be vetted before they are presented to the Governing Board for approval. Items meeting the statistical criteria will first be presented to TACSS to evaluate the results relative to the statistical criteria. If the TACSS recommends adjustments in statistical criteria, those will be implemented to further modify the pool of exemplar items. The final set of items deemed to be appropriate exemplars for each achievement level will be recommended to the Governing Board in August 2015 for approval to use in reporting achievement levels for TEL.

Section 9: Public Comment

Pearson will collect public comment on both the TEL ALS Design Document and on the ALS Outcomes. The collection of public comment on the ALS Outcomes will be done before the

Governing Board has released the achievement level cut scores and so must be done in a manner that maintains the security of the ALS Outcomes.

Public Comment on the Design Document

Pearson will create a website to obtain public comment on the Design Document. The website will provide a means for stakeholders and the public to find information about the Design Document and to leave feedback. A draft of the web page Pearson will use to collect public comment on the Design Document is shown in Appendix D.

All comments entered into the comment box will be saved in a secure location.

Pearson will submit the site to the Governing Board staff for review before the site goes live to the public.

When contacting people and organizations to provide panelist nominations, Pearson will include the website link and information about the opportunity to provide comment on the Design Document.

Public Comment on the ALS Outcomes

The public comment collection for the ALS outcomes must be done in a way that takes into account the restrictions on the release of data prior to the official release of The Nation's Report Card for Technology, Engineering, and Literacy. Pearson is exploring approaches to collecting comments that protect the confidentiality of the ALS outcomes. Pearson is exploring the following possibilities:

1. Organize a meeting of state education staff concurrent with a national conference; or,
2. Present at a NAEP State Coordinators meeting.

Pearson and the Governing Board staff seek guidance from COSDAM on the feasibility of collecting public comment on the ALS outcomes given the restrictions on the release of data prior to the official release of The Nation’s Report Card for TEL.

References

- ACT, Inc. (2007). Developing achievement levels on the 2006 National Assessment of Educational Progress in grade twelve economics: Process report. Iowa City, IA: Author.
- ACT, Inc. (2010). Developing achievement levels on the 2009 National Assessment of Educational Progress in science for grades four, eight, and twelve: Process report. Iowa City, IA: Author.
- Cizek, G. J., Bunch, B. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. [An NCME instructional module]. *Educational Measurement: Issues and Practice*, 23(4), 31–50.
- Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1), 36–48.
- Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice*, 21(1), 16–22. doi:10.1111/j.1745-3992.2002.tb00081.x
- Haertel, E. H. (2008). Standard setting. In K. E. Ryan and L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 139–154). Routledge.
- Haertel, E. H., & Lorie, W. A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives*, 2(2), 61–103.

- Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Hillsdale, NJ: Erlbaum.
- Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (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* (2nd ed., pp. 47–76). New York, NY: Routledge.
- Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433–470). Westport, CT: Praeger.
- Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64, 425–461.
- Kane, M. T. (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*. Mahwah, NJ: Lawrence Erlbaum.
- Kingston, N. M., & Tiemann, G. C. (2012). Setting performance standards on complex assessments: The body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 201–223). New York, NY: Routledge.
- Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The Bookmark Standard Setting Procedure. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (2nd ed., pp. 225–254). Mahwah, NJ: Lawrence Erlbaum.
- Maritz, J. S. & Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, 73, 194–196.

- McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Miles, J. A. (2013). Evidence-based standard setting: Establishing a validity framework for cut scores. *Educational Researcher, 42*, 78–88.
- National Assessment Governing Board (2010). *Developing achievement levels on the 2009 National Assessment of Educational Progress in Science at grades 4, 8, and 12: Technical report*. Washington, D.C.: National Assessment Governing Board.
- National Assessment Governing Board (2012). *Developing achievement levels on the 2011 National Assessment of Educational Progress in grades 8 and 12 writing: Technical report*. Washington, D.C.: National Assessment Governing Board.
- National Research Council (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.
- Sireci, S. G., Hauger, J. B., Wells, C. S., Shea, C. & Zenisky, A. L. (2009). Evaluation of the standard setting on the 2005 grade 12 National Assessment of Educational Progress Mathematics Test. *Applied Measurement in Education, 22*, 339–358.
- WestEd (2011). National Assessment of Educational Progress Grade 12 preparedness research project judgmental standard setting (JSS) studies: Process report San Francisco, CA: Author.

Appendix A: Draft Agenda for the Pilot Study and ALS Meeting

The following pages show a draft agenda for the pilot study and ALS meeting.

Day 1

Registration	8:00
Opening Remarks	8:00–8:30
Welcome and Why You Are Here	
Introductions	
Review of Agenda	
Administrative Tasks	
Panelist and Staff Introductions	
Panelist Recruitment	
Complete NAEP TEL Assessment Form	8:30–9:45
Score NAEP TEL Assessment Form	9:45–10:15
Panelist training on using scoring rubric	
Panelist review of their own responses	
BREAK	10:15–10:30
Overview of the NAEP TEL Framework	10:30–11:00
History	
Purposes	
Definition	
Overview of Achievement Level Setting	11:00–11:15
Purpose	
Item Mapping Methodology	

Review NAEP TEL Achievement Level Descriptions	11:15–noon
LUNCH	noon–1:00
Stage 1 Item Review	1:00–3:00
BREAK	3:00–3:15
Stage 2 Item Review	3:15–4:45
End-of-Day Activities	4:45–5:00
Review of Agenda for Day 2	
Evaluation #1	
Check in Materials	
END OF DAY 1	
<i>DAY 2</i>	
Reconvene	8:00–8:15
Answer Questions	
Review Day 2 Agenda	
Development of Borderline Achievement Level Descriptions	8:15–9:45
Review Achievement Level Descriptions	
Table-Level Discussion	
Group-Level Discussion	
BREAK	9:45–10:00
Overview of Modified Item Mapping Methodology	10:00–11:00
Item Mapping	
Ordered Item Booklet	

Item Map	
Ratings Forms	
Introduction to Dual-Computer Arrangement	11:00–noon
LUNCH	noon–1:00
Practice Round	1:00–1:45
Round One Ratings	1:45–4:45
(Break when Convenient)	
Review Method	
Training Evaluation	
Complete Ratings	
End-of-Day Activities	4:45–5:00
Review Day 3 Schedule	
Evaluation #2	
Check in Materials	
END OF DAY 2	
 <i>DAY 3</i>	
Reconvene	8:00–8:15
Answer Questions	
Review Day 3 Agenda	
Round One Feedback	8:15–9:15
Median Cut Scores	
Rater Location Chart	

Item Maps	
Review Borderline Achievement Level Descriptions	9:15–9:45
Round Two Ratings	9:45–noon
(Break when Convenient)	
Training and Feedback Evaluation Form	
Review Method	
Complete Ratings	
LUNCH	noon–1:00
Round Two Feedback	1:00–2:15
Median Cut Scores	
Rater Location Chart	
Item Maps	
Consequences (Impact) Data	
BREAK	2:15–2:30
Round Three Ratings	2:30–4:15
Review Method	
Training and Feedback Evaluation Form	
Complete Ratings	
End of Day Activities	4:15–4:30
Check In materials	
Review Day 4 Schedule	
Evaluation #3	
END OF DAY 3	

DAY 4

Reconvene	8:00–8:15
Answer questions	
Review Day 4 Agenda	
Round Three Feedback	8:15–9:15
Median Cut Scores	
Item Maps	
Consequences (Impact) Data	
Consequences Questionnaire	9:15–9:45
End of Standard Setting Activities	9:45–10:00
Evaluation #4	
Check in materials	
BREAK	10:00–10:15
Identification of Exemplar Items	10:15–11:45
Complete Exit Survey	11:45–12:15
Check in Materials	
END OF DAY 4	

Appendix B: Evaluation of the Success of Standard Setting Training

1. The instructions on what I am to do during this round is . . .

- a. Not at all clear
- b.
- c. Somewhat clear
- d.
- e. Absolutely clear

2. My understanding of the tasks I am to accomplish during this round is . . .

- a. Not at all adequate
- b.
- c. Somewhat adequate
- d.
- e. Absolutely adequate

3. My understanding of the Basic ALD is . . .

- a. Not at all adequate
- b.
- c. Somewhat adequate
- d.
- e. Absolutely adequate

4. My understanding of the Proficient ALD is . . .

a. Not at all adequate

b.

c. Somewhat adequate

d.

e. Absolutely adequate

5. My understanding of the Advanced ALD is . . .

a. Not at all adequate

b.

c. Somewhat adequate

d.

e. Absolutely adequate

Appendix C: Evaluation of the Round One Standard Setting Process

1. The instructions on what I was to do during each round were . . .

- a. Not at all clear
- b.
- c. Somewhat clear
- d.
- e. Absolutely clear

2. My understanding of the tasks I was to accomplish during each round was . . .

- a. Not at all adequate
- b.
- c. Somewhat adequate
- d.
- e. Absolutely adequate

3. At the time I provided the round one bookmark placements, my understanding of the Basic ALD was . . .

- a. Not at all adequate
- b.
- c. Somewhat adequate
- d.
- e. Absolutely adequate

4. At the time I provided the round one bookmark placements, my understanding of the Proficient ALD was . . .

- a. Not at all adequate
- b.
- c. Somewhat adequate
- d.
- e. Absolutely adequate

5. At the time I provided the round one bookmark placements, my understanding of the Advanced ALD was . . .

- a. Not at all adequate
- b.
- c. Somewhat adequate
- d.
- e. Absolutely adequate

6. The most accurate description of my level of confidence in the cut score recommendations I provided was . . .

- a. Not at all confident
- b.
- c. Somewhat confident
- d.
- e. Absolutely confident

7. The amount of time I had to complete the tasks I was to accomplish during each round was . . .

- a. Far too short
- b.
- c. About right
- d.
- e. Far too long

Appendix D: Draft Text for the Design Document Public Comment Web Page

NAEP TEL Achievement Levels-Setting Design Document Review

REQUEST FOR PUBLIC COMMENT: DESIGN DOCUMENT FOR SETTING ACHIEVEMENT LEVELS ON NAEP TEL ASSESSMENT

The [National Assessment Governing Board](#) is soliciting public comment for guidance in finalizing the Design Document for the National Assessment of Educational Progress (NAEP) in Technology and Engineering Literacy (TEL). The [NAEP TEL assessment](#) was administered for the first time in 2014 to a nationally representative sample of over 22,000 grade 8 students.

Achievement levels have become a powerful way to communicate student achievement on an assessment like the NAEP TEL because achievement levels interpret test performance with reference to cut scores that quantitatively define ordered categories of achievement such as Basic, Proficient, and Advanced. An important source of evidence used by policymakers to establish achievement levels is the cut score recommendations. Cut scores are the outcome of a facilitated process, called an achievement levels-setting or standard setting meeting, eliciting judgments from experts related to the test content and the knowledge, skills and abilities of the test takers.

The Design Document for developing achievement levels on the NAEP TEL at grade 8 is intended to provide the foundation for all achievement levels-setting activities. The Design Document for the TEL achievement levels-setting process includes discussion of the methodology, procedures, and documentation of the entire project.

Under [P.L. 107-279](#), the Governing Board is authorized to set policy for NAEP. The legislation specifies that the Governing Board is to develop appropriate student achievement levels for each subject and grade tested, as provided in section 303(e). Such levels are determined by identifying the knowledge that can be measured and verified objectively using widely accepted professional assessment standards; and developing achievement levels that are consistent with relevant widely accepted professional assessment standards and based on the appropriate level of subject matter knowledge for grade levels to be assessed.

To finalize plans for the TEL achievement levels-setting activities, and in preparation for reporting the results of the new TEL assessment at grade 8, the National Assessment Governing Board has issued a contract to [Pearson](#). [Pearson](#) will implement a process to produce a set of cut score recommendations to assist the National Assessment Governing Board in developing achievement levels for the 2014 NAEP TEL at grade 8.

On behalf of the Governing Board, [Pearson](#) has developed a Design Document that describes in detail the NAEP TEL achievement levels-setting activities. The Governing Board now seeks comment on the Design Document to provide recommendations to make improvements. All responses will be taken into consideration before finalizing the Design Document. The Design Document will be used to guide the achievement levels-setting activities to produce a set of cut score recommendations for reporting achievement levels for the 2014 administration of the NAEP TEL at grade 8.

Voluntary participation by all interested parties is urged. Comments can be provided by using the comments box below. Comments may also be provided via mail, to be received no later than November 28, 2014, at the following address:

NAEP TEL Design Document
National Assessment Governing Board
800 North Capitol Street N.W., Suite 825
Washington DC 20002

It is anticipated that the achievement levels recommendations will be presented for approval at the Governing Board meeting on August 6-August 8, 2015.

RESOURCE MATERIALS FOR PUBLIC COMMENT:

Materials that may be helpful in reviewing the TEL Design Document and providing comment include the following:

Technology and Engineering Literacy Framework: The NAEP TEL Framework was used for developing the assessment. An abridged version of the TEL Framework can be found [here](#), along with an [online, interactive version](#) and a [PDF](#) of the full framework.

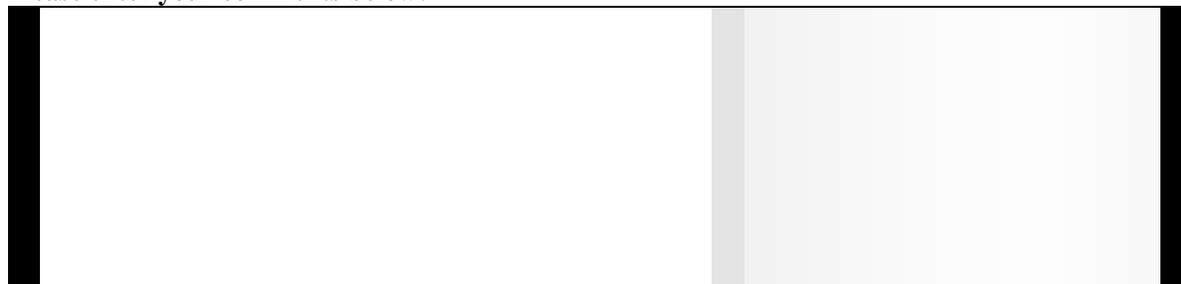
Overview Video: An overview video of the TEL assessment is available [here](#).

Policy on Developing Performance Levels for the National Assessment of Educational Progress: The Governing Board adopted a policy that describes the principles and guidelines to be used when setting achievement levels for NAEP. The policy can be found [here](#).

Focus Issues: While all comments are appreciated, project staff and content experts working on finalizing the Design Document are particularly interested in addressing the following issues:

- Does the Design Document adequately describe the achievement levels-setting activities?
- Does the achievement levels-setting methodology reflect the best practices?
- Does the proposed design reflect the principles and guidelines described in the policy?

Please enter your comments below:



For further information contact:

[Paul Nichols](#)

Project Director for NAEP TEL ALS, Pearson

paul.nichols@pearson.com



Update on White Paper on Transition to Technology-Based Assessments

To help plan NAEP’s transition from its current paper-based assessments to technology-based assessments (TBA), the National Center for Education Statistics (NCES) has commissioned a white paper that describes the overall approach being taken to accomplish this transition and its rationale, followed by subject specific considerations for mathematics, reading, and science. The first draft of the paper was completed in September by a cross-functional team of cognitive scientists, subject matter experts, and psychometricians. It is under review by a committee comprised of three members each of the three standing technical committees that currently provide technical guidance to NCES and its contractors. Those committees are the Design and Analysis Committee (DAC), the NAEP Validation Studies Panel (NVS), and the Quality Assurance Technical Panel (QATP). Each of these committees view NAEP with a different lens, ensuring that a wide range of perspectives is offered to solving the most important technical issues facing the program.

There are many reasons why this transition must begin now for NAEP’s core subject areas. Perhaps the most important reason is that assessment and learning in schools across the country have already started this transition. In order for NAEP to remain relevant and meaningful in the broader educational landscape, the program must begin now to convert to technology-based assessments that reflect how students are being prepared for post-secondary work and academic experiences. That being said, of particular concern to the “Nation’s Report Card” with its decades of valuable performance trends is the ability to maintain trend lines well into the future. As such, the program is planning a multistep process that will carefully and thoughtfully implement this important transition in a manner that is most likely to protect this valuable aspect. Whether or not trends can be maintained across paper-based and technology-based modes of administration is clearly an empirical question. All due care is being taken, however, to increase the likelihood that this important objective is achieved, and that NAEP will maintain its reputation as the gold standard of educational assessments. The white paper including the review process is part of the due care.

In the presentation to the committee, first an overview of the white paper will be provided, starting with design principles relative to different kinds of threats to trend and set against previous trend study designs and findings. Subsequently, the design of the current trend studies will be described for each of the subjects (mathematics, reading, and science), followed by subject specific considerations related to the development of content for technology-based assessments.

Upcoming Procurement: Review of Existing Studies on Motivation and Engagement in NAEP

During the August 2013 COSDAM meeting, Governing Board Executive Director Cornelia Orr reported on the desk side briefings that she had given to policy leaders and organizations about the results of the Governing Board's academic preparedness research. Ms. Orr reported that one of the questions she received was about whether grade 12 students are motivated to try hard on NAEP. Ms. Orr noted that it is important to be aware of the tendency to question whether grade 12 results represent students' best efforts. Some people have a hard time believing that 12th-graders try hard on a test that does not count. On the other hand, TIMSS and PISA are at the secondary level and also do not count.

There is some evidence that grade 12 students do take the test seriously, such as completion rates and completion of open-ended questions in particular. During the March 2014 COSDAM meeting, Samantha Burg of the National Center for Education Statistics (NCES) presented some encouraging data on grade 12 school and student participation rates and item response rates (from 1992 to 2013) and comparisons to grades 4 and 8. A Focus on NAEP report, *Grade 12 Participation and Engagement in NAEP*, is scheduled to be released by NCES next month.

On the other hand, if an ERIC search was performed on the terms "NAEP" and "motivation," the search would likely yield studies that conclude students are not very motivated. Previous COSDAM discussions have noted that the secondary research on NAEP and motivation that has been done and is often quoted has not been critiqued. One idea that has been discussed during previous COSDAM meetings is that a literature review and critique of existing studies could be performed as part of the efforts on preparedness research.

Responsively, we are planning a procurement to conduct a review and summary of existing research on motivation and engagement in NAEP, with the following goals:

- To critically evaluate the claims that have been made;
- To summarize the extent to which results are consistent across studies; and
- To recommend future research that should be performed.

This brief COSDAM session will include the following questions:

1. *Should the procurement include grades 4 and 8, in addition to grade 12?*
2. *What considerations should we take into account when planning for this procurement?*

Evaluation of NAEP Achievement Levels

Objective To receive a brief informational update on the current status of the independent evaluation of NAEP achievement levels that is being performed by the National Center for Education Evaluation and Regional Assistance (NCEE), part of the Institute for Education Sciences (IES). Ongoing updates will be provided at each COSDAM meeting.

Background

The NAEP legislation states:

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

In providing further detail, the aforementioned subsection (f) outlines:

(1) REVIEW-

- A. IN GENERAL- The Secretary shall provide for continuing review of any assessment authorized under this section, and student achievement levels, by one or more professional assessment evaluation organizations.
- B. ISSUES ADDRESSED- Such continuing review shall address--
 - (i) whether any authorized assessment is properly administered, produces high quality data that are valid and reliable, is consistent with relevant widely accepted professional assessment standards, and produces data on student achievement that are not otherwise available to the State (other than data comparing participating States to each other and the Nation);
 - (ii) whether student achievement levels are reasonable, valid, reliable, and informative to the public;-
 - (iii) whether any authorized assessment is being administered as a random sample and is reporting the trends in academic achievement in a valid and reliable manner in the subject areas being assessed;
 - (iv) whether any of the test questions are biased, as described in section 302(e)(4); and

- (v) whether the appropriate authorized assessments are measuring, consistent with this section, reading ability and mathematical knowledge.

(2) REPORT- The Secretary shall report to the Committee on Education and the Workforce of the House of Representatives and the Committee on Health, Education, Labor, and Pensions of the Senate, the President, and the Nation on the findings and recommendations of such reviews.

(3) USE OF FINDINGS AND RECOMMENDATIONS- The Commissioner for Education Statistics and the National Assessment Governing Board shall consider the findings and recommendations of such reviews in designing the competition to select the organization, or organizations, through which the Commissioner for Education Statistics carries out the National Assessment.

Responsively, a procurement was planned to administer an independent evaluation of NAEP achievement levels. The last update COSDAM reviewed on this topic was in August 2014.

Evaluation of NAEP Achievement Levels Contract Award

The National Center for Education Evaluation and Regional Assistance (NCEE), part of the Institute for Education Sciences (IES), will administer the Evaluation of the NAEP Achievement Levels. On September 29, 2014, NCEE awarded a contract to The National Academy of Sciences to perform this work.

Objectives for the evaluation include the following:

- Determine how "reasonable, valid, reliable and informative to the public" will be operationalized in this study.
- Identify the kinds of objective data and research findings that will be examined.
- Review and analyze extant information related to the study's purpose.
- Gather other objective information from relevant experts and stakeholders, without creating burden for the public through new, large-scale data collection.
- Organize, summarize, and present the findings from the evaluation in a written report, including a summary that is accessible for nontechnical audiences, discussing the strengths/ weaknesses and gaps in knowledge in relation to the evaluation criteria.
- Provide, prior to release of the study report, for an independent external review of that report for comprehensiveness, objectivity, and freedom from bias.

- If the optional tasks are authorized by ED, plan and conduct dissemination events to communicate the conclusions of the final report to different audiences of stakeholders.

Design:

This study will focus on the achievement levels used in reporting NAEP results for the reading and mathematics assessments in grades 4, 8, and 12. Specifically, the study will review developments over the past decade in the ways achievement levels for NAEP are set and used and will evaluate whether the resulting achievement levels are "reasonable, valid, reliable, and informative to the public." The study will rely on an independent committee of experts with a broad range of expertise related to assessment, statistics, social science, and education policy. The project will receive oversight from the Board on Testing and Assessment (BOTA) and the Committee on National Statistics (CNSTAT) of the National Research Council.

Cost/Duration: \$1,256,345 over 24 months (September 2014 to September 2016), with options

Current Status: Members of the interdisciplinary review committee are expected to be selected by early 2015, and the committee is expected to meet over the course of 2015. The report from the evaluation is expected to be released in 2016 and will be announced on <http://ies.ed.gov/ncee/>.

NAEP Academic Preparedness Research

Phase 1 Research

The first phase of the Governing Board's research on academic preparedness is now complete; results from more than 30 studies are available at: <http://www.nagb.org/what-we-do/preparedness-research.html>. During the August 2013 meeting, the Board voted on a motion to use the phase 1 research on academic preparedness for college in the reporting of the 2013 grade 12 national results for reading and mathematics, released on May 14, 2014. The motion, validity argument, and phase 1 final report are now available on the aforementioned website.

Phase 2 Research

The second phase of the Governing Board's research on academic preparedness currently consists of the following studies that are planned or underway:

Study name	Sample	November 2014 Update
Statistical linking of NAEP and ACT	National; FL, IL, MI, TN	Page 96
Statistical linking of NAEP and SAT	MA	Page 97
Longitudinal statistical relationships: Grade 12 NAEP	FL, IL, MA, MI, TN	Page 98
Statistical linking of NAEP and Explore	KY, NC, TN	Page 99
Longitudinal statistical relationships: Grade 8 NAEP	NC, TN	Page 100
Content Alignment Studies of the 2013 National Assessment of Educational Progress for Grade 8 Reading and Mathematics with ACT Explore Assessments of These Subjects		Pages 101-102
Evaluating Reading and Mathematics Frameworks and Item Pools as Measures of Academic Preparedness for College and Job Training		Pages 103-105
College Course Content Analysis		Page 106

Brief overviews and informational updates are provided for each study.

National and State Statistical Linking Studies with the ACT

The Governing Board is planning to partner with ACT, Inc. to conduct a statistical linking study at the national level between NAEP and the ACT in Reading and Mathematics. Through a procedure that protects student confidentiality, the ACT records of 12th grade NAEP test takers in 2013 will be matched, and through this match, the linking will be performed. A similar study at the national level was performed with the SAT in 2009. There will not be a national statistical linking study performed for NAEP and the SAT in 2013.

In addition, the state-level studies, begun in 2009 with Florida, will be expanded with 2013 NAEP. Again using a procedure that protects student confidentiality, ACT scores of NAEP 12th grade test takers in the state samples in partner states will be linked to NAEP scores. We are working with four states to be partners in these studies at grade 12: Florida, Illinois, Michigan, and Tennessee. In three of these states (IL, MI, TN), the ACT is administered to all students state-wide, regardless of students' intentions for postsecondary activities.

Research Questions for National and State Statistical Linking Studies with the ACT:

1. What are the correlations between the grade 12 NAEP and ACT student score distributions in Reading and Math?
2. What scores on the grade 12 NAEP Reading and Math scales correspond to the ACT college readiness benchmarks? (concordance and/or projection)
3. What are the average grade 12 NAEP Reading and Math scores and interquartile ranges (IQR) for students below, at, and at or above the ACT college readiness benchmarks?
4. Do the results differ by race/ethnicity or gender?

November 2014 Update: The data sharing agreements with MI and TN have been finalized, and data analyses are underway. Data sharing agreements with ACT, FL and IL are still being worked out.

State Statistical Linking Study with the SAT

In 2009, the Governing Board partnered with the College Board to conduct a statistical linking study at the national level between NAEP and the SAT in Reading and Mathematics. Through a procedure that protects student confidentiality, the SAT records of 12th grade NAEP test takers in 2009 were matched, and through this match, the linking was performed. There will not be a national statistical linking study performed for NAEP and the SAT in 2013.

We have partnered with Massachusetts to conduct a state-level linking study for 2013 NAEP and the SAT. Again using a procedure that protects student confidentiality, SAT scores of NAEP 12th grade test takers in Massachusetts will be linked to NAEP scores.

Research Questions for National and State Statistical Linking Studies with the SAT:

1. What are the correlations between the grade 12 NAEP and SAT student score distributions in Reading and Math?
2. What scores on the grade 12 NAEP Reading and Math scales correspond to the SAT college readiness benchmarks? (concordance and/or projection)
3. What are the average grade 12 NAEP Reading and Math scores and interquartile ranges (IQR) for students below, at, and at or above the SAT college readiness benchmarks?
4. Do the results differ by race/ethnicity or gender?

November 2014 Update: The data sharing agreement with MA has been finalized; the data are in the process of being prepared.

Longitudinal Statistical Relationships: Grade 12 NAEP

In addition to the linking of ACT scores to NAEP 12th grade test scores in partner states, the postsecondary activities of NAEP 12th grade test takers will be followed for up to six years using the state longitudinal databases in Florida, Illinois, Massachusetts, Michigan, and Tennessee. These studies will examine the relationship between 12th grade NAEP scores and scores on placement tests, placement into remedial versus credit-bearing courses, GPA, and persistence.

Research Questions for Longitudinal Statistical Relationships, Grade 12 NAEP:

1. What is the relationship between grade 12 NAEP Reading and Math scores and grade 8 state test scores?
2. What are the average grade 12 NAEP Reading and Math scores and interquartile ranges (IQR) for students with placement in remedial and non-remedial courses?
3. What are the average grade 12 NAEP Reading and Math scores (and the IQR) for students with a first-year GPA of B- or above?
4. What are the average grade 12 NAEP Reading and Math scores (and the IQR) for students who remain in college after each year?
5. What are the average grade 12 NAEP Reading and Math scores (and the IQR) for students who graduate from college within 6 years?

November 2014 Update: The data sharing agreements have been finalized for MA, MI, and TN. Data sharing agreements with FL and IL are still being worked out.

State Statistical Linking Studies with ACT Explore

In 2013, linking studies between 8th grade NAEP in Reading and Mathematics and Explore, a test developed by ACT, Inc. that is linked to performance on the ACT, are planned with partners in three states: Kentucky, North Carolina, and Tennessee. In all three of these states, Explore was administered to all students state-wide who were in grade 8 during the 2012-13 school year.

Research Questions for State Statistical Linking Studies with ACT Explore:

1. What are the correlations between the grade 8 NAEP and Explore scores in Reading and Math?
2. What scores on the grade 8 NAEP Reading and Math scales correspond to the Explore college readiness benchmarks (concordance and/or projection)?
3. What are the average grade 8 NAEP Reading and Math scores and the interquartile ranges (IQR) for students below, at, and at or above the Explore college readiness benchmarks?

November 2014 Update: The data sharing agreements are complete; analyses are currently underway.

Longitudinal Statistical Relationships: Grade 8 NAEP

In 2013, the Governing Board will also expand the state-level studies by partnering with two states at grade 8. Again using a procedure that protects student confidentiality, secondary and postsecondary data for NAEP 8th grade test takers in the state samples in partner states will be linked to NAEP scores. These studies will examine the relationship between 8th grade NAEP scores and scores on state tests, future ACT scores, placement into remedial versus credit-bearing courses, and first-year college GPA.

Two states will be partners in these studies at grade 8: North Carolina and Tennessee.

Research Questions for Longitudinal Statistical Relationships, Grade 8 NAEP:

1. What is the relationship between NAEP Reading and Math scores at grade 8 and state test scores at grade 4?
2. What are the average NAEP Reading and Math scores and the interquartile ranges (IQR) at grade 8 for students below the ACT benchmarks at grade 11/12? At or above the ACT benchmarks?
3. What are the average NAEP Reading and Math scores and the interquartile ranges (IQR) at grade 8 for students who are placed in remedial and non-remedial courses in college?
4. What are the average NAEP Reading and Math scores (and the IQR) at grade 8 for students who obtain a first-year college GPA of B- or above?
5. What is the relationship between grade 8 NAEP Reading and Math scores and grade 12 NAEP Reading and Math scores? (contingent on feasibility of sampling the same students in TN and NC)

November 2014 Update: The data sharing agreements are complete; analyses are currently underway (to address the first research question). Additional data will be transmitted when they become available over the next several years.

Content Alignment Studies of the 2013 National Assessment of Educational Progress for Grade 8 Reading and Mathematics with ACT Explore Assessments of These Subjects

Project Overview

In September 2014, the Governing Board awarded a contract to NORC at the University of Chicago, along with its subcontractor, the Wisconsin Center for Education Products and Services (WCEPS), to conduct content alignment studies with the ACT Explore assessments in reading and mathematics and the 2013 National Assessment of Educational Progress (NAEP) reading and mathematics assessments at grade 8. The purpose of this research is to evaluate the extent to which 8th grade NAEP is aligned in content and complexity with ACT Explore. For each subject area, the studies will compare the two assessments (NAEP and ACT Explore) to the NAEP framework, and also to the ACT College and Career Readiness framework. Using the content alignment methodology designed by Dr. Norman Webb for the Preparedness Research Program commissioned by the Governing Board, these studies will measure and describe the degree of alignment between the grade 8 NAEP math and reading assessments and ACT Explore assessments in these same subjects. The results of these NAEP-Explore content comparisons will also inform interpretations from statistical linking studies of 2013 results of NAEP and Explore in grade 8 reading and mathematics.

To support the provision of ACT proprietary Explore materials, the Governing Board also issued a sole source contract to ACT, Inc. The Governing Board will work with ACT to receive information and materials that will be used in the content alignment studies, and will consult with ACT assessment staff to support the work. Four ACT Explore content experts will attend the five-day Content Alignment Institute.

Project Team

Dr. Rolf Blank (NORC), the Project Director, has extensive experience in leading content analysis studies involving content standards and student assessment instruments. Dr. Norman Webb (WCEPS), the study Technical Coordinator, developed the WebbAlign process and the WATv2 online tool which have been used in many prior alignment studies at state and national levels. Along with Dr. Blank and Dr. Webb, several content experts and consultants will comprise the study team developing the study design, analysis procedures, and data analysis and report preparation.

Project Plan

Led by Dr. Blank and Dr. Webb, work has begun in preparation for a Content Alignment Institute to be held in February of 2015. The five-day Institute will be held in the Washington D.C. area at NORC's facilities. The Institute will gather a group of 32 panelists, consisting of 8th grade math and reading teachers that are knowledgeable about their subject content, and have experience with instruction and high-quality assessments of student learning for the target

subject and grade level. These panelists will be trained on the first day to ensure that they understand the WebbAlign concepts and how to use the WATv2 tool that will be used for coding the alignment analysis data. Panelists will work in 8-person subject teams to conduct analysis and coding of assessment frameworks and items. The teams will follow planned procedures for analysis, data review, and adjudication that have been developed by the NORC-WCEPS staff. Each panelist will enter their data into the online data tool which will then be used for data review and analysis. With completion of the Institute, the NORC-WCEPS staff will conduct analysis of the data and prepare draft study reports for math and reading. The study team will provide a final report by July 2015.

Milestones

There are several major milestones over the course of this project. To briefly summarize, the milestones include preparatory work, data collection, and alignment analysis and reporting. The following table includes the major milestones for completing this work:

<i>Milestone</i>	<i>Date</i>
Kickoff Meeting	9/29/14
Submit Planning Document	10/31/14
Conduct Framework Analyses	10/16/14-11/4/14
Recruit Panelists for Content Alignment Institute	11/1/14 – 12/19/14
Convene Design Review and Strengthening Meeting	11/12/14
Conduct Content Alignment Institute	2/9/15-2/13/15
Conduct Data Analysis	2/9/15-2/27/15
Prepare Draft Reports	3/8/15-4/15/15
Prepare Final Reports	5/21/15-7/1/15
Present final reports to COSDAM	7/15/15-8/7/15

COSDAM will receive ongoing updates as the study progresses.

EVALUATING READING AND MATHEMATICS FRAMEWORKS AND ITEM POOLS AS MEASURES OF ACADEMIC PREPAREDNESS FOR COLLEGE AND JOB TRAINING

Project Status Update Contract ED-NAG-13C-0001

The National Assessment Governing Board contracted with the Human Resources Research Organization (HumRRO) in June 2013 to conduct three tasks related to research on 12th grade preparedness:

1. **Evaluation of the Alignment of Grade 8 and Grade 12 NAEP to an Established Measure of Job Preparedness:** In its June 2009 report, *Making New Links: 12th Grade and Beyond*, the Technical Panel on 12th Grade Preparedness Research recommended that content alignment studies be conducted to examine the structure and content of various assessments relative to NAEP. The purpose of such content alignment would be to determine whether the scores on NAEP and the other assessments convey similar meaning in terms of the knowledge and skills of examinees. In fact, the panel specifically recommended that content alignment studies be conducted between NAEP and WorkKeys to determine the correspondence between the content domain assessed by NAEP and that of WorkKeys. If the alignment is relatively high, or even moderately high in some cases, then statistical relations between NAEP and WorkKeys may allow for the interpretation of NAEP results in terms of how WorkKeys would typically be interpreted. Using WorkKeys as a measure of job training preparedness allows the comparison of findings from this research to findings from previous content alignment studies with WorkKeys.

HumRRO extended prior analysis of the relation of NAEP to WorkKeys by including the NAEP grade 8 assessments and by expanding the method for assessing content alignment. ACT provided operational WorkKeys items in support of the study. The study method followed the Governing Board content alignment design document for preparedness research studies, with some modifications. The two-pronged approach included alignment of: (a) WorkKeys to the NAEP frameworks, and (b) NAEP items to the framework from which WorkKeys was developed. The draft report is currently under review.
2. **O*NET Linkage Study:** This study a) identified relevant linkages between the National Assessment of Educational Progress (NAEP) and training performance requirements for selected occupations, and b) compared the levels of knowledge, skills, and abilities (KSAs) required for the relevant NAEP content to the levels of KSAs required for the relevant job training content. For this study, tasks (i.e., performance requirements) for each occupation were extracted from O*NET. The O*NET, or Occupational Information Network, is the U.S.

Department of Labor's occupational information database. The final executive summary from this study was included in the July 2014 Governing Board briefing materials.

3. **Technical Advisory Panel (TAP) Symposium:** HumRRO assembled a technical advisory panel (TAP) of five experts in educational measurement and five experts in industrial-organizational (I-O) psychology to review extant research and to generate ideas for commissioned papers on preparedness. The TAP met in Washington D.C. in late October 2013. This brainstorming session included presentations by Governing Board and HumRRO staff describing findings from previous studies and descriptions of other studies currently underway, followed by an open discussion of issues and possible additional areas of investigation. Each panelist was asked to use this information to propose a paper that he/she could develop. TAP members submitted nine proposals from which Governing Board staff commissioned five papers. Panelists developed three of these papers and presented them in a TAP Symposium on August 20, 2014:

- *Using 8th and 12th Grade NAEP to Measure Student Readiness for Careers*, Barbara Plake, University of Nebraska-Lincoln
- *Grit: A Useful Concept in College and Career Preparedness?* Ann Marie Ryan, Michigan State University
- *Relating NAEP to Commercial Off-the-Shelf Measures*, Nancy Tippins, Corporate Executive Board – Valtera Corporation

A proceedings document summarizing the commissioned papers and discussion is under review. A list of TAP members is included on the next page.

In addition, HumRRO will produce a comprehensive project report at the conclusion of the contract in December 2014.

Work completed as of November 2014:

Evaluation of Alignment of Grade 8 and 12 NAEP to an Established Measure of Job Preparedness: Analyses of workshop ratings were completed and a draft report is under review.

O*NET Linkage: This task was completed in April 2014; see May 2014 Governing Board status update for details.

TAP Symposium: Governing Board staff reviewed proposals submitted by TAP panelists and commissioned five (5) papers to be completed by the panelists. Three papers were produced and

the TAP Symposium to discuss these papers was held on August 20, 2014. Symposium proceedings are under review.

Technical Advisory Panel (TAP) Members

John Campbell

Professor of Psychology
University of Minnesota
(Member, NAGB Technical Panel on 12th
Grade Preparedness Research, 2007-2008)

Michael Campion

Herman C. Krannert
Professor of Management
Purdue University

Gregory Cizek

Professor of Educational Measurement
and Evaluation
University of North Carolina at Chapel Hill

Brian Gong

Executive Director of Center for Assessment
National Center for the Improvement of
Educational Assessment, Inc.

Ronald Hambleton

Distinguished University Professor,
Educational
Policy, Research, & Administration
Executive Director, Center for Educational
Assessment
University of Massachusetts at Amherst

Suzanne Lane

Professor, Research Methodology
University of Pittsburgh School of
Education

Barbara Plake

University Distinguished Professor,
Emeritus
University of Nebraska-Lincoln

Ann Marie Ryan

Professor of Psychology
Michigan State University

Nancy Tippins

Senior Vice President
CEB Valtera

COLLEGE COURSE CONTENT ANALYSIS

Project Status Update Contract ED-NAG- 12C-0003

The College Course Content Analysis (CCCA) study is one of a series of studies contributing to the National Assessment of Educational Progress (NAEP) Program of 12th Grade Preparedness Research conducted by the National Assessment Governing Board (NAGB). The purpose of the CCCA study is to identify a comprehensive list of the reading and mathematics knowledge, skills, and abilities (KSAs) that are pre-requisite to entry-level college mathematics courses and courses that require college level reading based on information from a representative sample of U.S. colleges. The Educational Policy Improvement Center (EPIC) is the contractor working for the Board to conduct this study.

Another goal of the CCCA study is to extend the work of the two previous preparedness studies—the Judgmental Standards Setting (JSS)¹ study, implemented in 2011 and the Job Training Program Curriculum (JTPC) study, implemented in 2012. The CCCA study is designed so the results can be compared to the JSS and JTPC studies, reporting on how this new information confirms or extends interpretations of those earlier studies. The design of the CCCA study is based on the JTPC study but with modifications based on the lessons learned.

November 2014 Update: The project is now complete (see May 2014 COSDAM materials for Executive Summary). The final report is now available on the Governing Board’s website at: http://www.nagb.org/content/nagb/assets/documents/what-we-do/preparedness-research/judgmental-standard-setting-studies/College_Course_Content_Analysis.pdf.

¹ National Assessment Governing Board. (2010). *Work Statement for Judgmental Standard Setting Workshops for the 2009 Grade 12 Reading and Mathematics National Assessment of Educational Progress to Reference Academic Preparedness for College Course Placement*. (Higher Education Solicitation number ED-R-10-0005).

OVERVIEW OF REFERENCED ASSESSMENTS

For additional background information, the following list presents a brief description of the assessments referenced in the phase two academic preparedness research studies. In each case, only the mathematics and reading portions of the assessments are the targets for analysis, although analyses with the composite scores may be conducted.

- ACT – 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. 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.
- ACT Explore – ACT Explore assesses academic progress of eighth and ninth grade students. It is a component of the ACT College and Career Readiness System and includes assessments of English, Mathematics, Reading, and Science. ACT Explore has *College Readiness Standards* that connect reading and mathematics knowledge and skills and probabilities of a college course grade of “C” or higher (0.75) or “B” or higher (0.50) by the time students graduate high school with particular score ranges on the Explore assessment.
- SAT – The SAT reasoning test is a college admissions test produced by the College Board. It is used by colleges and universities to evaluate the knowledge and skills of applicant pools in critical reading, mathematics, and writing. The SAT has calculated preparedness benchmarks are defined as the SAT scores corresponding to a 0.65 probability of earning a first-year college grade-point average of 2.67 (B-) or better.

Origin of English Language Learners Inclusion Guidelines

At the August 2014 Governing Board meeting, the Committee on Standards, Design, and Methodology (COSDAM) and the Reporting and Dissemination (R&D) Committee convened a joint meeting to discuss and ultimately make changes to the Board's policy on NAEP Testing and Reporting on Students with Disabilities (SDs) and English Language Learners (ELLs). At the end of the meeting, R&D Chair Andrés Alonso called for the joint committee to reconsider the requirement that ELL students should be included in NAEP if they have been in U.S. schools for at least one year, following federal guidelines laid out for states by the No Child Left Behind Act. COSDAM Chair Lou Fabrizio suggested a U.S. Department of Education representative be invited to address the joint committee about the origin of this requirement.

As a precursor to further discussion, Board staff contacted the U.S. Education Department to gather specific information on the one-year requirement. Officials sent Board staff a Federal Register notice that contained regulations that were part of Title I—Improving the Academic Achievement of the Disadvantaged, updated and finalized by the Department's Office of Elementary and Secondary Education in 2006. These regulations, amended to federal No Child Left Behind legislation, implemented various changes regarding state and local educational requirements in regard to academic achievement and school accountability for limited English proficient (LEP) students. The notice includes details of updated legislation, and a compilation of some of the 50 or so public comments gathered during the “notice of proposed rulemaking” period and the Department's responses to those comments.

The relevant passage in the Department's regulations stipulates that a state would be able to exempt only “recently arrived LEP students” from one administration of the state's reading/language arts assessment. This category is defined as “a student with limited proficiency in English who has attended schools in the United States for less than twelve months” (not including Puerto Rico as Spanish is considered the language of instruction there).

Many who submitted public comment before the regulations were finalized recommended the definition of a “recently arrived” LEP student to mean a LEP student who has attended schools in the United States for a period of time ranging from 12 months to five years or to tie the definition to a student's English language proficiency. In response, the Department defended the 12-month rule in part stating, “We believed it was important to have a time limit to ensure that the one-time exemption is used only for LEP students who have recently arrived in schools in the United States, not for those students who have lived in the United States for a number of years and attended United States schools but who still possess limited proficiency in English.

The final regulations also require that recently arrived LEP students take the mathematics assessment. In response to public comments, the Department stated: “The Secretary believes that English language proficiency is not a prerequisite to participating in State mathematics

assessments to the same extent as it is to participating in State reading/language arts assessments. Research provides evidence on accommodations that can be used with LEP students in mathematics and have been shown not to compromise the validity of the test and skills being measured when appropriately implemented.”

This is an information item on the COSDAM and R&D agendas for the November 2014 meeting. A joint meeting with both committees featuring a presentation and in-depth discussion on this issue and the implications for NAEP will be scheduled for the March 2015 meeting. To review the Federal Register notice with the full regulations and summary of public comments gathered, visit: <http://www2.ed.gov/legislation/FedRegister/finrule/2006-3/091306a.pdf>.