



Pearson

Developing Achievement Levels on the 2014 National Assessment of Educational Progress in Grade 8 Technology and Engineering Literacy

Process Report

May 2016

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: Pearson
2510 N. Dodge St.
Iowa City, IA 52240

*Developing Achievement Levels on the
2014 National Assessment of Educational
Progress in Grade 8 Technology and
Engineering Literacy:
Process Report*

**Lori Nebelsick-Gullett
edCount, LLC**

**Steve Fitzpatrick
Pearson**

May 2016

National Assessment Governing Board

BOARD MEMBERSHIP

(2015 - 2016)

Terry Mazany, Chair

President and CEO
The Chicago Community Trust
Chicago, Illinois

Lucille Davy, Vice Chair

President and CEO
Transformative Education Solutions, LLC
Pennington, New Jersey

Alberto M. Carvalho

Superintendent
Miami-Dade County Public Schools
Miami, Florida

Honorable Mitchell D. Chester

Commissioner of Elementary and Secondary
Education
Massachusetts Department of Elementary and
Secondary Education
Malden, Massachusetts

Frank K. Fernandes

Principal
Kaimuki Middle School
Honolulu, Hawaii

Honorable Anitere Flores

Senator
Florida State Senate
Miami, Florida

Rebecca Gagnon

School Board Member
Minneapolis Public Schools
Minneapolis, Minnesota

Shannon Garrison

Fourth-Grade Teacher
Solano Avenue Elementary School
Los Angeles, California

Honorable James E. Geringer

Former Governor of Wyoming
Cheyenne, Wyoming

Doris R. Hicks

Principal and CEO
Dr. Martin Luther King, Jr.
Charter School for Science and Technology
New Orleans, Louisiana

Andrew Dean Ho

Professor of Education
Harvard Graduate School of Education
Harvard University
Cambridge, Massachusetts

Carol Jago

Associate Director
California Reading & Literature Project at
UCLA
Oak Park, Illinois

Tonya Matthews

President and CEO
Michigan Science Center
Detroit, Michigan

Tonya Miles

General Public Representative
Mitchellville, Maryland

Honorable Ronnie Musgrove

Former Governor of Mississippi
Madison, Mississippi

Dale Nowlin

Twelfth-Grade Teacher
Columbus North High School
Columbus, Indiana

Joseph M. O'Keefe, S.J.

Professor
Lynch School of Education Boston College
Chestnut Hill, Massachusetts

W. James Popham

Professor Emeritus
University of California, Los Angeles
Wilsonville, Oregon

B. Fielding Rolston

Chairman
Tennessee State Board of Education
Kingsport, Tennessee

Linda P. Rosen

CEO
Change the Equation
Washington, DC

Cary Snider

Associate Research Professor
Portland State University
Portland, Oregon

Honorable Ken Wagner

Commissioner of Elementary and Secondary
Education
Rhode Island Department of Education
Providence, Rhode Island

Chasidy White

Eighth-Grade Teacher
Brookwood Middle School
Vance, Alabama

Joseph L. Willhoft

Assessment Consultant
Tacoma, Washington

Ex-officio Member**Ruth Curran Neild**

Deputy Director for Policy and Research
Institute of Education Sciences (IES)
Delegated Duties of the Director of IES
U.S. Department of Education
Washington, D.C.

Table of Contents

Executive Summary	1
Background	1
Panelist Recruitment and Selection	2
Advanced Materials and Preparation of Panelists	3
Achievement Levels-Setting Procedures	3
Preparation for Applying the Item Mapping Methodology	5
Panel Studies in the TEL ALS Process	8
Operational ALS Study	9
Results of the NAEP TEL ALS Process	11
Governing Board Action	14
Introduction	15
Background on NAEP Achievement Levels-Setting Activities	15
Background on the Current Project	16
Coordination with NAEP Operations	19
Contract Activities Prior to the ALS Meeting	20
Kick-Off Meeting	20
Planning Document	20
Design Document	21
Collection of Public Comment	21
Studies for the TEL ALS Process	22
Dual-Computer Usability Study	22
Introduction and Purpose	22
Participants	24
Method	24
Results	26
General Procedures Applied to the Pilot Studies and the Operational ALS	28
Recruitment and Selection of Panelists	28
Advance Materials and Preparation of Panelists	32
Panelist and Item Pool Division and Assignment of Forms	33
ALS Methodology	36
Completing the Consequences Review	48

Selection of Exemplar Items	51
Process Evaluation Procedure.....	52
Facilitator Guide and Training	53
Pilot Studies	54
Pilot 1	54
Pilot 2	59
The Operational Achievement Levels-Setting Study	68
Panelists	68
Materials	69
Logistics.....	70
Round 1: Understanding the Assessment and Student Achievement	79
Round 2: Using Feedback	80
Round 3: Using Consequences Data.....	81
Post-Round 3 Activities	82
Outcomes of the Achievement Levels-Setting Process.....	84
Process Evaluation.....	84
Summary of Cut Scores	98
Reliability of Cut Scores	102
Exemplar Item Ratings.....	107
Governing Board Action	108
Recommendations for Future Standard Settings.....	108
References	110
Addendum: Governing Board Action on the Achievement Levels	111
Appendices.....	114
Appendix A: NAEP TEL ALS Facilitator Guide	114
Appendix B: Institutions Contacted to Nominate Non-teacher Educators.....	114
Appendix C: Organizations Contacted to Nominate General Public Representatives	114
Appendix D: Sample Panelists Recruitment Letters	114
Appendix E: Briefing Book	114
Appendix F: Lists of Panelists.....	114
Appendix G: NAEP TEL Achievement Level Descriptions	114
Appendix H: NAEP Policy Definitions	114

Appendix I: Item Map	114
Appendix J: ALS Questionnaires and Panelist Responses.....	114
Appendix K: Email Sent to Panelists Regarding Addition of Pilot 2.....	114
Appendix L: Pilot 1 Room Layout.....	114
Appendix M: Pilot 2 and ALS Room Layout.....	114
Appendix N: NAEP TEL ALS Operational Study Agenda	114
Appendix O: ALS PowerPoints	114
Appendix P: Excel Tools	114
Appendix Q: Feedback by Round	114
Appendix R: Summary of Exemplar Ratings.....	114
Appendix S: Exemplar Items Recommended to the Governing Board	114
Appendix T: Final Exemplar Item Set.....	114

List of Tables

Table 1: ALS Panelist Demographics.....	10
Table 2: Pattern of Response to Repeated Evaluation Questions Across Rounds ...	11
Table 3: ALS Panel Median Cut Scores and Standard Deviations by Round	12
Table 4: ALS Panel Recommendations for NAEP TEL after Round 3 (Final Round) .	12
Table 5: ALS Panel Recommendations for NAEP TEL after Round 3: Operational versus Pilot 2	13
Table 6: Number of Panelists in Each Response Category for Questions 4, 5, and 18 on the Final Process Evaluation.....	14
Table 7: Achievement Levels-Setting (ALS) Meetings.....	22
Table 8: Minutes Required for Item Ratings.....	26
Table 9: Percent of Total Time Spent on the Out-of-system Item Review	27
Table 10: Characteristics of Ordered Item Lists.....	36
Table 11: Pilot 1 Panelist Demographics.....	55
Table 12: Summary of Schedule and Activities.....	58
Table 13: Pilot 2 Panelist Demographics.....	60
Table 14: Summary of Pilot 2 Schedule and Activities	63
Table 15: The Pilot 2 Panel Recommendations for NAEP TEL after Round 3.....	64

Table 16: Number and Percentages of Pilot 2 Panelists Who Changed Cut Score Recommendations during Consequences Questionnaire Activity	64
Table 17: Pilot 2 Panel Recommendations for NAEP TEL Cut Scores	65
Table 18: Pilot 2 Panelist Changes to Repeat Survey Questions Across Rounds	66
Table 19: Number of Pilot 2 Panelists in Each Response Category for Questions 4, 5, and 18 on the Final Process Evaluation	67
Table 20: ALS Panelist Demographics	68
Table 21: Pattern of Response to Repeated Evaluation Questions across Rounds..	84
Table 22: Number of Panelists in Each Response Category for Questions 4, 5, and 18 on the Final Process Evaluation	85
Table 23: Average Ratings of Clarity of Instructions and Presentations	86
Table 24: Average Ratings of Usefulness or Helpfulness of Activities and Information.....	88
Table 25: Average Ratings of Amount of Time Allocated for Tasks.....	89
Table 26: Average Ratings of Understanding of Concepts and Feedback.....	90
Table 27: Average Ratings of Understanding the ALDs and Borderline Performance	92
Table 28: Average Ratings of Comfort with Processes and Procedures	93
Table 29: Average Ratings of Reactions to Consequences Data	96
Table 30: Average Ratings of Confidence with Decisions/Outcomes	97
Table 31: ALS Panel Median Cut Scores and Standard Deviations by Round.....	98
Table 32: NAEP TEL ALS Median Cut Scores by OIL and Table Group	100
Table 33: Number and Percent of Panelists Who Changed their Cut Scores between Rounds in the ALS	100
Table 34: Panel Recommendations for NAEP TEL Achievement Levels after Round 3	101
Table 35: Number and Percent of ALS Panelists Who Changed Cut Score Recommendations during Consequences Review.....	101
Table 36: Median Cut Scores and Standard Deviations (S.D.) by Table and Round for the Basic Achievement Level	104
Table 37: Median Cut Scores and Standard Deviations (S.D.) by Table and Round for the Proficient Achievement Level.....	104

Table 38: Median Cut Scores and Standard Deviations (S.D.) by Table and Round for the Advanced Achievement Level	105
Table 39: Medians and Mean Absolute Difference (MAD) of Round 3 Cut Scores by Demographic Variables	106
Table 40: Standard Errors of Median Scores by Level and Round	106
Table 41: NAEP TEL Cut Score Recommendations by ALS and Pilot 2 Panels after Consequences Review	107

List of Figures

Figure 1: The Assignment of Item Sets to Panelists Subgroups	35
Figure 2: Illustration of Borderline Performance	40
Figure 3: Mapping Student Achievement and Item Difficulty	41
Figure 4: Identifying the Bookmark in the Zone	42
Figure 5: Table Level Feedback.....	44
Figure 6: Rater Location Chart – Table Level	45
Figure 7: Consequences Data	47
Figure 8: Consequences Review Chart	50
Figure 9: Composition of Table Groups for Pilot 1	55
Figure 10: Composition of Table Groups for Pilot 2	61
Figure 11: Composition of Table Groups for Operational ALS Study	69
Figure 12: Summary of Schedule and Activities.	71
Figure 13: Segment of Test Form Review Spreadsheet.	72
Figure 14: Segment of the Item Review Worksheet.....	74
Figure 15: Sample Ordered Item List.....	77
Figure 16: Bookmark Recording Form.....	80
Figure 17: Mean Absolute Difference in Panelist Scores by Round	99
Figure 18: Cut scores by OIL and Round	102
Figure 19: Cut Scores by Table Group, Round, and Level	103

Executive Summary

This report describes the process and outcomes of a study designed and implemented by Pearson to develop achievement levels recommendations for the 2014 National Assessment of Educational Progress (NAEP) in Technology and Engineering Literacy (TEL) for grade 8. The achievement levels-setting (ALS) process was conducted by Pearson under contract with the National Assessment Governing Board to produce recommendations to the Governing Board that would serve as the basis for the Board's decision on setting the NAEP TEL achievement levels. The contract was awarded to Pearson in July 2014, and the achievement levels were set by the Governing Board in November 2015.

This document describes all procedural aspects and outcomes of the operational achievement levels-setting study conducted September 28 – October 2, 2015. Additional studies conducted in preparation for the Operational ALS study are also described in this report: a dual-computer usability study and two pilot studies. Technical aspects of the procedures implemented for the NAEP TEL ALS, as well as the rationales and decisions regarding the procedures implemented, are provided in a Technical Report. In addition, National Assessment Governing Board staff produced a statement of the decisions made by the Governing Board for setting the NAEP TEL achievement levels.

An item mapping procedure was designed for the 2014 NAEP TEL achievement levels setting. Several studies were conducted to evaluate and refine procedures for determining the TEL achievement level cut scores and the selection of exemplar items to recommend to the National Assessment Governing Board for use in reporting student performance on the NAEP TEL. Throughout the process, Pearson staff worked with the Technical Advisory Committee on Standard Setting (TACSS) to help assure that procedures were well designed from a psychometric perspective and well designed for implementation with a nationally representative set of panelists from a variety of backgrounds.

In addition to guidance from the TACSS, Pearson staff provided quarterly briefings and updates to the Governing Board's Committee on Standards, Design and Methodology (COSDAM). Throughout the process of designing and implementing achievement levels-setting procedures and preparing reports, COSDAM monitored activities and provided general guidance and direction regarding the conduct of the work and the final recommendations to be delivered.

Background

The entirely computer-based and interactive NAEP TEL was administered for the first time in 2014 to a nationally representative sample of more than 20,000 grade 8 students. The Governing Board developed the TEL Framework from which the innovative assessment was created. The TEL assessment includes three types

of scenario-based assessment tasks: long (30 minutes), medium (20 minutes), and short (10 minutes). These scenarios incorporate animations, audio, and video components as part of the TEL items. In addition to the interactive, scenario-based tasks, the NAEP TEL assessment includes a set of discrete items.

The ALS methodology used for the NAEP TEL was designed to 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 was adapted to be appropriate for a complex assessment comprising both discrete items and performance-based scenarios. The Governing Board specified that the ALS procedure should be fully computerized. Pearson chose an item mapping approach (Lewis, Mitzel, Mercado, & Schulz, 2012) that allowed for the collection of content-centered judgments across both scenarios and discrete item blocks using the same standard setting procedures.

To implement the ALS procedures, each panelist was provided with two laptop computers. The National Center for Education Statistics (NCES) provided computers used to administer the NAEP TEL to students. To gain an understanding of the assessment and the students' experience, panelists used this computer to take a version of the NAEP TEL assessment that paralleled the student experience. In addition, all of the scenarios were available to view individually in a slightly different review interface that still maintained the interactive nature of the items in the scenario, but was different from how the students experienced the assessment. Pearson provided a second computer for each panelist that allowed the panelist to record content-centered judgments, process ratings, and other relevant information, and to receive feedback throughout the standard setting process.

Panelist Recruitment and Selection

Recruitment and selection of panelists occurred in two cycles. The initial cycle focused on identifying and selecting panelists for two events – the pilot and operational studies. However, a second full pilot study was added to the procedures Pearson implemented to help ensure valid ALS results. Consequently, a second recruitment and selection cycle was initiated following Pilot 2 to identify and select panelists for the Operational ALS study. All recruitment and selection efforts aligned with ensuring that panels met the Governing Board policy requirements that the panel should consist of 55% teachers in the subject area at the grade level of the assessment, 15% non-teacher educators (not K-12 teachers) in the subject area, and 30% general public (not educators) with educational training and/or experience in the subject area and direct experience with children in the age range of eighth grade students. The process also focused on the requirement that panelists' demographics should be representative of geographical region, gender, and race/ethnicity. The TACSS recommended that at least one teacher of English

Language Learners (ELL) and one teacher of students in Special Education programs be selected for each panel.

The number of panelists who participated in Pilot 1 and Pilot 2 was one less than the goal for each of these panels due to last-minute changes in panelists' schedules or personal circumstances. Consequently, Pilot 1 and Pilot 2 had 14 and 29 participants, respectively. For the operational ALS, 31 panelists served on the panel.

Advanced Materials and Preparation of Panelists

Extensive materials were provided to NAEP ALS panelists—both in advance and during the panel meetings—to ensure that they were well prepared for each task in the process. Panelists for each meeting were sent advanced materials to begin their training and preparation for the ALS process in which they would participate. These materials included information about the NAEP program and the Governing Board; details about travel arrangements and logistics for the meeting site; the NAEP TEL Framework; the NAEP TEL achievement levels descriptions; a draft agenda; and a Briefing Booklet describing each step in the process and its purpose.

Achievement Levels-Setting Procedures

Division of Items into Item Rating Sets

As was done for previous NAEP ALS processes, items were divided into item rating sets to limit the number of items reviewed by each panelist in order to reduce the time required for the process and to reduce the potential for panelist fatigue. The NAEP TEL item pool of 20 scenarios and 97 discrete items was divided into three sets. Each set of the three sets had a group of unique items in addition to items that were common across the three sets. All items from a single scenario were assigned to the same item rating set. The common item set consisted of items that were on one administration form and were a subset of the items selected for possible release to the public. Common items were included in the three sets in order to have items to serve as examples in group discussions with panelists during the standard setting process. The item sets were constructed to be as equivalent as possible in terms of content area, item type, and item difficulty.

Division of Panelists into Subgroups

Panelists were assigned to one of six tables and each of the three item rating sets was assigned to the panelists at two tables. The demographic attributes of the panelists were considered when assigning members to tables to maximize the equivalence across tables; otherwise, the assignments were random. The goal was to have tables as equal as possible with respect to panelist type (i.e., teacher, non-teacher educator, or general public), gender, region, and race/ethnicity.

Item Mapping Methodology

For the NAEP TEL ALS studies, Pearson implemented a criterion-referenced, content-based item mapping method, similar to those previously used in NAEP ALS studies (Math 2005, Grade 12; Science 2009; Judgmental Standard Setting Studies, 2011). Items were arranged along a continuum using IRT-based item difficulty estimates. Standard setting panelists then located the cut points by identifying items that examinees performing at different achievement levels should be able to answer correctly with a specified level of probability.

Given the presentation of items through an electronic format, Pearson replaced the term 'ordered item booklet' (OIB), typically used in item mapping procedures, with an 'ordered item list' (OIL). The order of presentation of items in the OIL was not the order in which students were administered the items; rather, items and item score codes were arranged in order from the easiest to the most difficult, based on student performance on the NAEP TEL assessment in 2014. An item map was used to graphically display TEL items on a pseudo-NAEP student achievement scale depicting the assessment area for each items and the OIL in which it appeared.

The methodology was implemented through an iterative process in three rounds. Panelists were first given extensive instruction and training in order to be prepared for the task of judging performance.

Training and Instruction

The first day of the ALS process was dedicated to giving panelists information and instructions on the fundamentals of the process. Successful standard setting procedures require a common understanding among panelists regarding the purpose for each aspect of the process and a common understanding of the instructions for implementing each aspect of the process.

The orientation process began with a welcoming address by the Pearson project director. This was followed by an overview of NAEP and the role of the Governing Board, historical information about NAEP achievement levels, the Governing Board policy on setting achievement levels, the purpose of setting achievement levels, and their importance to the Governing Board's overarching goals and responsibilities. The Pearson project director then provided an overview of the NAEP TEL ALS process in which the panelists would be engaged over the next several days, including an overview of standard setting and a review of the full agenda.

Panelists were next engaged in taking the NAEP TEL administered as it was for students. This activity provided panelists with first-hand understanding of what students experienced when taking the NAEP TEL. After all panelists completed the assessment, scoring guides and rubrics were provided for panelists to review the requirements for correctly responding to each selected-response item or obtaining a specific score code for constructed-response items. This was their initial training in

the use of the NAEP TEL administration software and computer, the types of assessment tasks, and the scoring rubrics.

Instruction in the NAEP TEL Framework and achievement levels descriptions was the next topic on the agenda for the first day. The step-by-step implementation of the ALS process was under the leadership of a process facilitator and a content facilitator. The process facilitator had expertise in standard setting procedures and in the item mapping procedures implemented for the NAEP TEL ALS process. The content facilitator was a member of the framework development team, which required recognized expertise in the TEL field and provided knowledge of the rationale behind the components of the TEL Framework.

Preparation for Applying the Item Mapping Methodology

In order to make judgments about performance of students relative to the NAEP TEL achievement levels descriptions (ALDs), panelists needed to have a good understanding of the assessment and the knowledge and skills that students need to display in order to respond correctly and fully to the assessment items and tasks. An important part of the item mapping methodology for standard setting is to engage panelists in an exercise to record a brief description of the knowledge and skills required for correctly responding to each item or to each score code for constructed-response items. On the second day of the ALS process, panelists reviewed items and wrote descriptions of the knowledge and skills required for each item in their OIL. Each panelist reviewed an assigned subset of the items in the OIL. All items in an OIL were common to the panelists at the same table and assigned for review by at least two panelists in the table group. After all item reviews were complete, panelists at the same table shared information about reviewed items to ensure that all table members had information on every item in their assigned item list, i.e., in their OIL.

Developing Borderline Achievement Level Descriptions

After completing the review of items and describing the knowledge and skills associated with correct responses, panelists had an understanding of what students needed to know and be able to do on the assessment. The content and process facilitators guided panelists through the process of creating borderline achievement level descriptions (BALDs) for the Basic, Proficient, and Advanced achievement levels. Prior to developing the BALDs the process facilitator provided training and information to ground the panelists in their understanding of what the BALDs represent and their importance to the achievement levels-setting process. While the largest amount of time was allocated to the initial drafting of the BALDs, prior to the first round of using the BALDs to judge performance on items, panelists were instructed that they would have the opportunity to review and modify the BALDs prior to each round of ratings. Further, they were informed that the BALDs were for their use in the ALS process and not part of official documentation and reporting.

Panelists worked iteratively, as individuals, in table groups, and as the whole group to come to an agreement on the descriptions for performance judged to be “just good enough” to meet the criteria of a particular achievement level description. To develop the borderline description for Proficient performance, followed by Basic, and then Advanced, panelists used the NAEP TEL grade 8 ALDs and the NAEP policy definitions, as well as their understanding of the knowledge and skills required to perform well on the NAEP TEL items.

Using the ALS Item Mapping Methodology to Set Cut Scores

Following the instruction and preparation, panelists were next trained in the procedures for setting cut scores. They were instructed about the methods for the mapping of items and students onto the same scale and were provided graphics to illustrate this concept. They were also instructed in the use of 0.67 as the response probability to place items and score codes on the score scale to create the item map. This response probability was described as representing a reasonably high probability (2/3) of a correct response (or of achieving a particular score code) without being overly demanding.

Panelists participated in a practice round, and a questionnaire was administered to ascertain that everyone felt prepared to begin the process of setting cut scores. The ALS process included three rounds during which panelists were encouraged to review the ALDs, BALDs, and Governing Board policy definitions to conceptualize performance at the border of a particular achievement level. Panelists then started with the easiest item in the OIL and worked their way up through more and more difficult items to independently set the Proficient cut score. After locating the Proficient cut score, each panelist next set the Basic cut score, and then the Advanced cut score.

Each round of ratings was followed by feedback. Panelists were told that the group cut score was the median of individual panelist’s cut scores. They were given information about their own cut score location for each achievement level, and they were able to see on a graph where their cut scores were located relative to those of other panelists in their table group (Round 1) and in the whole group (all rounds). Feedback from Round 2 included consequences data showing the percentage of students in the 2014 NAEP TEL that scored within the cut score ranges of each achievement level. Panelists were instructed that the cut scores set in Round 3 would be considered as their final cut scores. These cut scores would be evaluated for recommendation to the Governing Board to use for reporting results of student performance on the 2014 NAEP TEL.

Completing the Consequences Review and Final Recommendations

Following the third and final round of setting cut scores, panelists reviewed consequences data. Panelists completed a questionnaire that asked them to rate their understanding of the consequences data from Round 3 and then to evaluate

the percentage of students in each performance category, based on the Round 3 recommended cut scores. Panelists were then asked if they would change the Round 3 cut scores for one or more of the achievement levels. After completing this questionnaire, panelists used an interactive graphics tool that allowed them to adjust cut scores and see the impact or consequences of the cut score adjustments in terms of the percentage of students performing within each achievement level. Panelists were then administered a questionnaire to record any changes they made to each cut score. If a change was recommended, the panelist was asked to provide the rationale for the change.

Selection of Exemplar Items

For the final step in the NAEP ALS process, panelists identified exemplar items for each achievement level from a pool of items designated for release to the public. Exemplar items are a part of the official set of information recommended to the Governing Board as part of the achievement levels-setting process. These items serve to communicate to the public the types of knowledge, skills, and abilities that are required for performance within the Basic, Proficient, and Advanced NAEP achievement levels. Because this is an entirely new, innovative area of assessment for the NAEP program, the role of exemplar items in communicating performance on the NAEP TEL was considered especially important. Student performance on the item must demonstrate the knowledge, skills, and abilities that align with those in the ALD for the level an item represents. Selection of exemplar items was the last set of judgments panelists were asked to make, and they were well prepared at this point to judge which items would appropriately represent the criteria required for performance in each achievement level.

Items designated to be released were categorized into the achievement level at which the item had a response probability of 0.67 at a scale score point within the Basic, Proficient, or Advanced achievement level cut score range. Each panelist reviewed only the items that appeared in his or her OIL. Data about each item were provided for the exemplar selection process: the scale score at which the item had a probability of a correct response of 0.67, the average probability of a correct response across the range of the achievement, the probability of a correct response at the lower boundary of the achievement level, and information about the item content. Panelists were asked to rate each item as "Should be Used," "Might be Used," or "Should not be Used" as an exemplar for the specified achievement level.

Process Evaluation Procedure

Process evaluation questionnaires were administered throughout the ALS process. The questionnaires included both selected-response and open-ended questions that addressed the panelists' understanding and evaluation of instructions, tasks, and materials, as well as their comfort level with particular processes and their

confidence in the results. Questionnaires were completed at the following points of the ALS process:

- End of Day 1
- End of Day 2
- Post Practice Round
- Pre and Post Round 1
- Pre and Post Round 2
- Pre and Post Round 3
- Pre and Post Consequences
- Overall Evaluation at the End of the Process

Most responses were collected on a three- or five-point Likert scale, but several responses were narratives that addressed specific aspects of the process.

Panel Studies in the TEL ALS Process

Dual-Computer Usability Study

A dual-computer usability study was conducted at the recommendation of the TACSS. The structure of the scenarios and the interactive features of the NAEP TEL required that panelists have access to the assessment items in the “live” format in order to make judgments regarding student performance for each NAEP achievement level. The software for administering the NAEP TEL does not allow for items to be extracted from the scenarios nor for panelists to skip around among items within scenarios. A second software system was used to support the achievement levels-setting process and the collection of panelists’ judgments regarding achievement levels cut scores. Previous NAEP ALS studies have required the use of two computers, one for taking the NAEP assessment and the other for the ALS process. NAEP TEL required panelists to use the dual-computer setup throughout the ALS process in order for panelists to review the interactive items as needed. Because this introduced new complexities that required use of the dual-computer setup throughout the panelists’ ALS activities, TACSS recommended that a separate usability study be conducted prior to the ALS pilot study.

Initial Pilot Study

A pilot study to test out all aspects of the procedures planned for implementation in the Operational ALS study was required by the Governing Board. Although fewer panelists were required for the pilot study, the procedures and criteria for recruitment and selection of panelists were the same as for the operational study. Only one pilot study was planned for the NAEP TEL ALS process, and the initial pilot study was conducted March 16-19, 2015, with 14 panelists.

The pilot study revealed the need for changes in several aspects of the procedures planned for the TEL ALS process. The achievement levels-setting software leased by Pearson for this project did not function well in this process, although it had been successfully implemented previously in a large-scale standard setting study. Problems with the functioning of the software, coupled with the large number of items and item score points requiring judgments in the ALS process, led to delays in procedures and the need for more time than scheduled in the agenda to complete some tasks. Because of the magnitude—both in number and importance—of changes required, the TACSS advised that a second pilot study be implemented, and the Governing Board approved a second pilot study. This change required a delay in the schedule for the operational ALS meeting.

Second Pilot Study

A full day was added to the agenda for implementing the process based on the results of the first pilot study. This was done largely to provide more time for panelists to develop descriptions of the knowledge and skills required to respond correctly to items they were to use in judgments for setting cut scores. The second pilot study was implemented June 1-5, 2015, which is when the Operational ALS study had previously been scheduled. Panelists who had been recruited for the Operational ALS study were informed about the change and invited to participate in the second pilot study. A total of 29 panelists participated in the study, and the representativeness of the panel was close to the target percentages.

Pearson staff had developed Excel-based software for panelists to use for all aspects of the ALS process: to review items, to describe the knowledge and skills needed for correct responses, to set cut scores, to review feedback, and to respond to evaluation questionnaires. The software functioned well, and both results and feedback could be produced quickly and accurately.

Overall, the second pilot study was successfully implemented. All panelists (100%) either agreed or strongly agreed with the following statements in the final process questionnaire for Pilot Study 2:

- This ALS process produced achievement levels that are defensible.
- This ALS process produced reasonable achievement levels.
- I would be willing to sign a statement (after reading it of course) recommending the use of the cut scores resulting from this ALS process.
(Yes/No)

Operational ALS Study

The operational ALS panel was convened September 28 – October 2, 2015, in San Antonio, Texas. The same facility was used for both pilot studies and the operational ALS meeting.

Only minor modifications to the process were made following Pilot 2. Slightly more time was provided in the agenda for panelists to describe the knowledge and skills required by items in their OIL, and the number of items assigned to each panelist was reduced. Additional refinements were made to facilitate development of the borderline achievement level descriptions (BALDs) to help ensure adherence to the time scheduled for that activity at each point in the agenda. The software functioned well, and procedures were implemented smoothly and effectively in the operational ALS.

Panelists

A total of 31 individuals served as panelists in the operational ALS. The following table shows the demographic composition of the panel and demonstrates that key factors were generally represented on the panel according to targeted proportions.

Table 1: ALS Panelist Demographics

Demographic Variable	Attributes	Percent
Panelist Type	Teachers	55%
	Non-Teacher Educators	16%
	General Public	29%
Gender	Female	68%
	Male	32%
Race/Ethnicity	White/Non-Hispanic	55%
	Black/Non-Hispanic	26%
	Hispanic	10%
	Asian	3%
	Other	6%
NAEP Region	Midwest	26%
	Northeast	3%
	South	55%
	West	16%
Students with Disabilities	Experienced	87%
	Not experienced	13%
English Language Learners	Experienced	73%
	Not experienced	27%

Panelists' Evaluation of Key Aspects of the ALS Process

Panelists were asked to evaluate various aspects of the ALS process at each round. The following table reports responses (agree and strongly agree) to some of the key questions asked at each round in the process. These responses demonstrate that panelists had a clear understanding of key aspects of the item mapping process and felt comfortable in using the methodology and procedures. It is

especially important to note that panelists' confidence and understanding increased across rounds.

Table 2: Pattern of Response to Repeated Evaluation Questions Across Rounds

Question	Strongly Agree / Agree		
	Round 1	Round 2	Round 3
The instructions on how I was to select my bookmarks were clear.	25 (81%)	31 (100%)	31 (100%)
I had a good understanding of how to use the Borderline Achievement Level Descriptions to select my bookmarks.	23 (74%)	31 (100%)	31 (100%)
I understand the difference between borderline performance and typical performance within an achievement level.	29 (94%)	31 (100%)	31 (100%)
When choosing my bookmarks, I was comfortable taking into account how the Technology and Engineering Literacy principles and practices related to the achievement level.	21 (68%)	29 (94%)	29 (94%)
When choosing my bookmarks, I was comfortable thinking about and using the idea of borderline performance within an achievement level.	22 (71%)	31 (100%)	31 (100%)
The Technology and Engineering Literacy principles and practices required by the items around my bookmarks are appropriate for the borderline of the corresponding achievement level.	19 (61%)	29 (94%)	29 (94%)
I was comfortable using the description of performance at the borderline of Basic when I selected my Round 1/2/3 bookmarks.	24 (77%)	31 (100%)	31 (100%)
I was comfortable using the description of performance at the borderline of Advanced when I selected my Round 1/2/3 bookmarks.	22 (71%)	28 (90%)	29 (94%)
I am confident in my Round 1/2/3 bookmark selections.	13 (42%)	28 (90%)	30 (97%)

Results of the NAEP TEL ALS Process

Cut scores were computed as the median for panelists at a table and across all panelists. The following table shows the cut scores after each of the three rounds. The standard deviation represents the level of variability across panelists' cut scores. The standard deviation generally decreased across rounds as panelists gained greater understanding of the process and confidence in their ability to make

judgments about performance relative to the achievement levels descriptions and their conceptualization of borderline performance.

Table 3: ALS Panel Median Cut Scores and Standard Deviations by Round

Round	Scale Scores and Standard Deviations (S.D.)					
	Basic		Proficient		Advanced	
	Median	S.D.	Median	S.D.	Median	S.D.
1	116	19.4	150	30.0	193	40.2
2	116	12.9	151	17.3	205	27.6
3	116	12.2	151	18.3	209	14.8

Data in Table 3 show that although individual panelists may have changed their cut scores from round to round, the overall cut scores changed not at all for the Basic level and by only one point for the Proficient level. However, at the Advanced level, the cut score increased at each round.

NAEP achievement is typically reported in terms of the percentages of students performing within and at or above each achievement level. Those consequences data were shared with panelists after Round 2 and again after Round 3. Consequences data are reported in Table 4 for Round 3. Panelists were instructed that the Round 3 results were considered final and would be evaluated for recommendation to the Governing Board.

Table 4: ALS Panel Recommendations for NAEP TEL after Round 3 (Final Round)

Level	Cut Scores and Percentages		
	Scale Score	Percent In Level	Percent at or Above
Basic	116	32%	83%
Proficient	151	48%	51%
Advanced	209	3%	3%

After reviewing the Round 3 cut scores, panelists were provided with the consequences data and administered a questionnaire to ascertain whether they would recommend that the Governing Board make further changes to the cut scores to be used for reporting student performance on the NAEP TEL. The consequences data provided after the final round were in an interactive format so that panelists could view the percentage of students within each level associated with changes in cut scores for the level(s). Panelists could recommend changes or no changes to any or all of the cut scores for Round 3. Some panelists

recommended changes to the cut scores, but those individual recommendations did not change the whole group cut scores.

One important reason that the TACSS recommended that a second pilot study be conducted was to produce results that could be compared with the operational ALS results and evaluated for consistency of the results. As can be seen in Table 5, the results for Pilot 2 and the Operational ALS are very similar.

Table 5: ALS Panel Recommendations for NAEP TEL after Round 3: Operational versus Pilot 2

Level	Cut Scores and Percentages					
	Operational ALS Panel Results			Pilot 2 Panel Results		
	Scale Score	Percent In Level	Percent at or Above	Scale Score	Percent In Level	Percent at or Above
Basic	116	32%	83%	119	30%	81%
Proficient	151	48%	51%	151	46%	51%
Advanced	209	3%	3%	204	5%	5%

Exemplar Item Ratings

Statistical criteria were used to categorize items into achievement levels from the set designated for public release. The items having a response probability of 0.67 within the score range of an achievement level were presented to panelists for consideration as exemplars of the performance required of students in each achievement level. Additional data provided to assist panelists in their judgments were the average probability of a correct response for each item within the designated achievement level and the probability of a correct response at the cut score for each achievement level. The goal in the selection of exemplar items was to maximize the number of items recommended to the Governing Board in order to illustrate the content coverage and measurement features of the assessment.

Panelists were asked to review the items in the potential exemplar list and rate each according to whether it “Should be Used,” “Should not be Used,” or “Might be Used.” Items with ratings that met pre-specified criteria for recommendation to the Governing Board as exemplars were vetted by TACSS prior to presentation to the Governing Board for approval. In order to provide the maximum amount of choice among items to be presented as exemplars in the *Nation’s Report Card*, priority was given to items with the highest frequency of panelists’ ratings as “Should be Used” and with the lowest frequency of panelists’ ratings as “Might be Used.” An item rated as “Should not be Used” by more than 20 percent of the panelists was considered only if it was necessary to represent a particular feature of the assessment at a specific level of achievement. Governing Board and NCES staff

then reviewed the recommended items and identified the items that would serve as exemplars in reports. The final set of items selected to serve as exemplars for each achievement level were approved by the Governing Board in November 2015.

Process Evaluation

In each study, panelists were asked three questions that provide key information regarding their evaluation of the results produced in the ALS process. As was the case for Pilot 2, operational ALS panelists were very positive in their evaluations of the process as shown in Table 6, although one ALS panelist had a “neutral” response.

Table 6: Number of Panelists in Each Response Category for Questions 4, 5, and 18 on the Final Process Evaluation

Evaluation Question	Response N (%)				
	Strongly Agree/ Yes	Agree	Neutral	Disagree	Strongly Disagree / No
This ALS process produced achievement levels that are defensible.	20 (65%)	10 (32%)	1 (3%)	0 (0%)	0 (0%)
This ALS process produced reasonable achievement levels.	20 (65%)	10 (32%)	1 (3%)	0 (0%)	0 (0%)
I would be willing to sign a statement (after reading it of course) recommending the use of the cut scores resulting from this ALS process.	29 (94%)				2 (6%)

Governing Board Action

The Governing Board reviewed the results of the ALS process during its meeting of November 19-21, 2015. Prior to that meeting, Pearson presented the results to CODSAM during a webinar meeting on November 3, 2015. CODSAM requested additional analyses from Pearson and Governing Board staff to address questions that arose during the discussion, and a second webinar was held on November 17, 2015. After reviewing that information, the Governing Board accepted the cut scores that resulted from the operational ALS meeting for the Basic (116) and Advanced (209) achievement levels, and adopted a cut score of 158 for the Proficient achievement level as compared to the operational ALS panel recommendation of 151.

Introduction

Performance standards have become a powerful way to communicate student achievement because they provide a frame of reference to interpret test performance quantitatively, with reference to cut scores, by defining ordered categories such as basic and proficient (Haertel and Lorie, 2004). The cut score recommendations that result from an achievement levels-setting (ALS) process are an important source of information that the National Assessment Governing Board and other policy makers use to establish performance standards. 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 required to attain specified achievement levels (Hambleton, Pitoniak, & Copella, 2012).

Background on NAEP Achievement Levels-Setting Activities

Congress authorized the creation of the National Assessment Governing Board in 1988 through Public Law 100-297. The Governing Board was given policy authority over the National Assessment of Educational Progress (NAEP), including development of frameworks for the assessments and identification of “appropriate achievement goals” for each subject and grade in the NAEP. The Governing Board referred to background reports and documents that had helped to shape the 1988 legislation to determine that three achievement levels should be set for NAEP: Basic, Proficient, and Advanced (Loomis & Bourque, 2001, p. 178). Achievement levels are designed to answer the question “How much mastery of content standards is enough to be classified at a given performance level?” The Governing Board developed general policy definitions for each of three achievement levels—Basic, Proficient, and Advanced—that describe performance criteria for students at any grade level and any subject in the NAEP program. These policy definitions are then operationalized for each specific grade and subject in the NAEP program.

Whereas many different methodologies have been researched and evaluated for collecting panelists’ judgments regarding performance relative to NAEP achievement levels, these can be categorized as item rating procedures, such as modified Angoff; item mapping procedures, such as Mapmark and Bookmark, and holistic procedures such as booklet classification and Body of Work. Prior to 2005, the Governing Board required the use of a modified Angoff procedure. For the Technology and Engineering Literacy assessment, the Governing Board did not specify a method, but required the use of a psychometrically sound methodology with a research base. This flexibility is necessary based on the diversity in assessment frameworks across the subjects assessed by NAEP, as well as the changes in technology associated with the assessments.

Background on the Current Project

The Governing Board awarded a contract to Pearson in July 2014 to design and implement a procedure to produce achievement levels for reporting results of the National Assessment of Educational Progress in Technology and Engineering Literacy (TEL) for a nationally representative sample of students in grade 8. This contract was initially awarded for an 18-month period, but it was extended to 22 months to provide more time for refining the operational ALS procedures implemented for this innovative assessment.

An item mapping procedure was designed for the 2014 NAEP TEL. Pearson staff designed and implemented several studies to test and refine the ALS procedure, including the procedures used to establish the TEL achievement level cut scores and those used to select exemplar items. (The exemplar items were selected as recommendations to the Governing Board for use in reporting student performance on the NAEP TEL assessment.) Throughout the process, Pearson staff worked with the Technical Advisory Committee on Standard Setting (TACSS) to help ensure that the procedures were psychometrically sound and could be implemented with a nationally representative set of panelists from a variety of backgrounds.

In addition to guidance from the TACSS, Pearson staff provided quarterly briefings and updates to the Governing Board's Committee on Standards, Design and Methodology (COSDAM). Throughout the process of designing and implementing achievement levels-setting procedures and preparing reports, COSDAM monitored activities and provided general guidance and direction regarding the conduct of the work and the final recommendations delivered to the Governing Board.

The computer-based NAEP TEL assessment was administered for the first time in 2014 to a nationally representative sample of more than 20,000 grade 8 students. The Governing Board-adopted TEL Framework formed the basis from which the assessment was developed. The TEL assessment included three types of scenario-based assessment tasks: long (30 minutes), medium (20 minutes), and short (10 minutes). These scenarios incorporated animations, audio, and video components as part of the TEL items. In addition to the interactive, scenario-based tasks, the NAEP TEL assessment included a set of discrete items. Discrete items are independent, stand-alone items not tied to a scenario.

The ALS methodology used for the NAEP TEL had to 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 had to be appropriate for a complex NAEP TEL assessment comprising both scenario-based tasks and discrete items. Pearson chose an item mapping approach (Lewis, Mitzel, Mercado, & Schulz, 2012) that allowed for the collection of content-centered judgments across both scenarios and discrete item blocks using the same standard setting procedures. Pseudo-NAEP

scales were used for all components of the study (Pilots and Operational) to avoid the risk of early release of cut score information. As in past ALS studies, the pseudo-NAEP scale was a linear transformation of the NAEP reporting scale.

To implement the ALS procedures, each panelist was provided with two laptop computers. Westat, a contractor for the National Center for Education Statistics (NCES), provided computers used to administer the NAEP TEL to students. To gain an understanding of the assessment and the students' experience, panelists used this computer to take a version of the NAEP TEL assessment that paralleled the student experience. Once the test was started, the test taker had to proceed through the scenarios and items linearly, and could not return to a previous scenario. The software used for the actual test administration, as experienced by students, was referred to as the "in system" software. In addition, all of the scenarios were available to view individually in a slightly different review interface that still maintained the interactive nature of the items in the scenario, but that differed from the assessment the students experienced. This software was referred to as the "out of system" software on the NAEP computer. The out-of-system software also included nine discrete items that were selected because they had an interactive component that Pearson judged to be difficult to understand when represented as PDF screenshots.

Panelists used out-of-system software on the NAEP computer to view the TEL scenarios and interactive items outside of the test environment throughout the standard setting process, providing panelists the opportunity to analyze and evaluate the complexity and cognitive demands imposed by the items and scenarios. Pearson provided a second computer for each panelist that allowed the panelist to record content-centered judgments, process ratings, and other relevant information and to receive feedback throughout the standard setting process. In the early stages of the project design, Pearson conducted a dual-computer usability study, described below, to determine the feasibility of using two computers and to provide logistical guidance for their use with panelists.

Project Staff

Dr. Steve Fitzpatrick served as the project director for the TEL ALS project. Members of his leadership team included Ross Holstein as the program manager responsible for logistics, and Morgan Hickey the Senior Research Associate from Pearson who provided analytic and technical support. Jennifer Eichel, Conference Solutions, LLC, planned and organized the meeting location and logistics. Dr. Susan Loomis served as a consultant to Pearson throughout the project. Two facilitators worked with the panelists throughout all pilot and ALS studies. Dr. Lori Nebelsick-Gullett served as the process facilitator, ensuring the ALS process was implemented with fidelity, and Dr. Johnny Moyer served as the content facilitator and onsite

content expert. For more information on the facilitators' roles and responsibilities, see Appendix A for the NAEP TEL ALS Facilitator Guide.

This team represented a change from the initial Pearson team that designed and initiated this standard setting work. The staffing shifts were due to early changes in personnel; nevertheless, team members worked effectively together throughout the transition.

Technical Advice

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 serve on a Technical Advisory Committee for Standard Setting (TACSS) that is convened for several in-person meetings and webinars to provide advice at every key point in the process. They provided feedback on plans and materials before activities were implemented and reviewed results of the process and analyses. The discussions with the TACSS were summarized for each meeting and recommendations were noted. The minutes of TACSS meetings are included as an appendix in the Technical Report to provide additional details of the technical considerations and deliberations regarding the procedures implemented for this ALS project and the analyses to be reported.

Plans for the various studies and all results were presented to the Governing Board's Committee on Standards, Design and Methodology (COSDAM) during each quarterly Board meeting (from August 2014 to November 2015) and in two conference calls held during November 2015. In addition to the members of the TACSS, Dr. Sharyn Rosenberg, the Governing Board's Assistant Director for Psychometrics and Contracting Officer's Representative (COR) for this project, provided technical advice to Pearson throughout the project, participated in all TACSS meetings, and attended all panel meetings. Dr. Andrew Kolstad, a former NCES representative to the TACSS, served as a consultant to the Governing Board. Dr. Mary Crovo, Deputy Executive Director for the Governing Board, and Michelle Blair, Senior Research Associate, provided input during TACSS meetings. Dr. Amy Yamashiro and Dr. Bill Tirre served as NCES liaisons.

The names of the experts in standard setting who served on the TACSS are shown below.

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 Huff¹

Vice President of Strategy & Execution, ACT

Dr. Matthew Johnson

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

Dr. Marianne Perie²

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)

Coordination with NAEP Operations

Some of the materials, data, and equipment used to conduct this project were provided through NCES by NAEP Alliance member companies. The provisions obtained from each are listed below.

Educational Testing Service (ETS)

- Item metadata
- Statistical item analysis
- Item response theory item parameters
- List of items requiring special treatments
- Item images
- Scored student-level data file
- Student level background data file
- Scale score means and percentiles for major reporting variables

¹ Kristen Huff was appointed to the TACSS as the representative of a state testing program, during her previous position as Senior Fellow at the State of New York Regents Research Fund. In June 2015, she joined ACT.

² Marianne Perie resigned from the TACSS after the February 2015 meeting.

- Lookup table for percent at or above each scale score point (1-300) for major reporting variables
- Item/Person distribution maps for grade 8 for 2014 TEL, 2013 Math, and 2011 Science

Pearson³

- PDF copies of student responses for constructed-response items
- Scoring rubrics and key annotations for constructed-response items

Westat

- NAEP computers used for taking the test and viewing items

Fulcrum IT

- Test administration and item viewing software on the NAEP computers

Contract Activities Prior to the ALS Meeting

Kick-Off Meeting

A kick-off meeting was conducted July 11, 2014, in Washington, DC. This meeting provided the opportunity for staff from Pearson and the Governing Board to meet in person. The purpose of this meeting was to review and discuss specific aspects of the contract, especially budgeting and reporting procedures required for government contracts. In addition, Pearson staff had the opportunity to ask questions and discuss some specific details of the contract work. Members of the Pearson team and Governing Board staff participated in the meeting.

Planning Document

A Planning Document was developed that described the procedures recommended for setting the achievement levels and the timelines associated with those procedures. The document was organized around the six tasks described in the Governing Board's solicitation.

- Task 1 was to attend the kick-off meeting.
- Task 2 was the Planning Document.
- Task 3 was the production of the Design Document that guided the ALS work.

³ Pearson has a scoring contract with NCES, which is separate and independent of the contract with the Governing Board to conduct the achievement levels setting work.

- Task 4 was to conduct a pilot study to implement the exact procedure, as described in the Design Document.
- Task 5 was to conduct the ALS process in order to generate recommendations to assist the Governing Board in establishing achievement levels on the NAEP 2014 Technology and Engineering Literacy assessment.
- Task 6 was to develop all final reports.

After Task 4 was conducted, a decision was made to significantly modify implementation of the planned procedures. As a result, Task 5 was converted to a second pilot study and Task 7 was added for conducting the ALS meeting.

Design Document

A Design Document was developed to describe in detail the NAEP TEL achievement levels-setting activities to be implemented. This document was intended to provide the foundation for all ALS activities. The Design Document for the TEL ALS process included discussion of the ALS methodology, the collection of public comment, and the identification of exemplar items, as well as the information to be provided in the final reports. The procedures described in the document were developed in consultation with the TACSS, COSDAM, and Governing Board staff. Once adopted, the Design Document was 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.

Collection of Public Comment

Pearson created a website to obtain public comment on the Design Document. The website provided a means for stakeholders and the public to find information about the Design Document and to leave feedback. Pearson submitted the site to the Governing Board staff for review before the site went live to the public.

The organizations and stakeholder groups that were identified to serve as nominators of panelists for the ALS process were sent information and asked to provide comments on the Design Document. Pearson included the website link and information about the opportunity to provide comment on the Design Document. While these groups and organizations appeared to be the key stakeholders for the TEL assessment and achievement levels, none chose to provide comments on the Design Document. The public comment period took place from October 29 – November 28, 2014.

The Governing Board also required that Pearson plan collection of public comment on the ALS outcomes. Pearson proposed collecting public comment for the ALS outcomes in conjunction with the National Conference on Student Assessment (NCSA) in June of 2015, and both the TACSS and the Governing Board accepted this recommendation. However, given the addition of a second pilot study conducted during the June timeframe originally established for the operational ALS

event, public comment could not be collected at that time. No feasible alternatives for collecting public comment on the ALS outcomes were identified through Pearson’s discussions with TACSS and the Governing Board representatives. The Governing Board staff and COSDAM agreed to delete this requirement for public comment.

Studies for the TEL ALS Process

Preliminary usability and pilot studies were conducted to determine how best to have the panelists engage with the items and standard setting materials due to the unique scenario-based and interactive nature of the test items. Table 7 lists the studies that led up to the operational ALS meeting in September 2015. The primary purpose of each meeting is identified in the table. Each of these studies is described in the sections that follow.

Table 7: Achievement Levels-Setting (ALS) Meetings

Meeting	Primary Purpose	Dates	Venue
Dual-Computer Usability Study	To test the logistics involved in using two laptop computers	December 2-4, 2014	Chandler, AZ
Initial Pilot Study	To implement the process designed for the operational meeting and evaluate the need for change(s)	March 16-19, 2015	San Antonio, TX
Second Pilot Study	To test implementation of modifications based on initial pilot study findings	June 1-5, 2015	San Antonio, TX
Operational ALS Meeting	To implement achievement levels-setting procedures to develop recommendations for consideration of the Governing Board	September 28 – October 2, 2015	San Antonio, TX

Dual-Computer Usability Study

Introduction and Purpose

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. For example, many of the TEL items within a scenario are related to each other and must be administered in a specific sequence. While other NAEP assessments have included groupings of items, such as those related to reading passages, the NAEP TEL item scenarios include interactive functionality that makes them unique. The innovative features of the NAEP TEL required innovative

strategies for the NAEP TEL ALS process. Panelists need to have an understanding of the knowledge and skills required to answer the items correctly in order to complete the ALS activities. This makes it necessary to have access to the items within the context of their scenarios, including the scenario functionality and the relationships of items within the scenario. As previously noted, each panelist used two computers for the pilot and operational studies. One computer was used to present the items to panelists within the scenarios with full functionality, as the items were administered in the NAEP assessment. Items were presented as in the student test: in order within each scenario and without the ability to skip to a specific item. A second computer was used to complete the ALS activities, including presentation of a list of TEL items ordered by a measure of item difficulty and the tools needed to collect panelists' judgments and present feedback data.

Studies on panelists' use of a dual-computer environment have been conducted in previous NAEP ALS procedures (e.g., Writing, 2011). These studies supported the use of a dual-computer format when one computer is used solely for panelists to take the particular NAEP assessment and to review the items as presented to students in the administration of the assessment. The items for NAEP TEL introduce an added layer of complexity because panelists need to evaluate items within the context of the scenarios as well as discrete items with an interactive component. For NAEP TEL ALS, panelists needed continual access to the items to review them as they were presented to students in the online environment. An interface was developed that allowed panelists to review the scenarios and select discrete items throughout the ALS process in an environment that maintained the interactive nature of the items. Given the uniqueness of the NAEP TEL items and the added complexities introduced that required use of the dual-computer setup throughout the panelists' ALS activities, TACSS recommended that a separate usability study be conducted prior to the ALS pilot study.

The plans and procedures detailed in the project Design Document for the pilot and operational ALS studies were subject to change based on information collected prior to each of those events. The goal of the usability study was to use the ALS dual-computer setup as described and to use the information obtained for planning and modifications prior to the pilot ALS meetings. The study focused on how panelists interacted with the computer setup rather than focusing on the quality of their judgments or evaluating the overall ALS procedures. The research questions for this study were:

1. Does the dual-computer setup hinder the ALS process in any way?
 - Can panelists navigate the items, including viewing them within their scenarios, and provide ratings using the dual-computer setup?
 - Does the dual-computer setup distract panelists from the ALS activities?

2. What kind of training is needed for 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 recruited five grade 8 science teachers from the Phoenix, Arizona, area for participation in this study. All teachers recruited for the study had previously participated in recent ALS activities in Arizona, such as those done for Arizona's statewide achievement test, the Instrument to Measure Standards (AIMS), or the Arizona English Language Learner Assessment (AZELLA), and, therefore, knew what was entailed in ALS activities. This experience and knowledge was expected to allow more time to focus on the NAEP TEL items and the NAEP TEL ALS computer setup, as well as to reduce the time spent on general ALS training.

Teachers required approximately four hours to complete the study activities. They were compensated for their time. All teachers who participated in the study were required to sign confidentiality agreements to ensure the security of the NAEP TEL items.

Method

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

Pearson selected the sample of items for the study to represent that of the full TEL ALS assessment in terms of scenario-based and discrete items. The group of items used for the study consisted of short scenarios and six discrete items. The test form began with a scenario, followed by the block of discrete items, and concluded with a second scenario. This grouping of items is similar to that used to construct operational NAEP TEL forms.

On one computer, these items were presented as they would appear to a student completing the assessment, meaning the items for each scenario were presented together with the original stimuli and functionality. On the second computer, the items were presented in order of item difficulty, as they would be arranged in the ordered item booklet (OIB) planned for the pilot and operational ALS activities. Within the OIB, the original functionality of the item was not accessible; instead, a static depiction of the item was displayed. The second computer also included the software needed to record ratings.

Pearson conducted the study in Pearson's usability lab located in Chandler, Arizona. The lab is equipped with a one-way mirror and observation room as well as several cameras that permitted sessions to be recorded. Conducting the study in this lab

allowed the participants' activities to be recorded without interference while also capturing participants' feedback and interactions with the NAEP TEL materials.

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

Study activities were completed individually with participants during a four-hour session after school hours on December 2, 3 and 4, 2014. Each participant took part in the following activities designed according to plans for the pilot and operational ALS meetings:

- Introduction to the study and description of the study purpose.
- Simulated test-taking experience during which the participant completed the sample of NAEP TEL items, as presented to students with regard to order and functionality.
- Review of Achievement Level Descriptions (ALDs). The facilitator reviewed the ALDs with the participant, specifically focusing on those for Proficient.
- Discussion of the borderline Proficient student. Using a similar approach to that planned for the ALS pilot and operational meetings, the facilitator worked with the participant to identify the knowledge and skills that define a just barely Proficient student.
- Training on the ALS procedure. The facilitator described the purpose of ALS, provided an overview of the ALS methodology, taught the participant about the OIB, and discussed the process of providing the necessary ratings. As much as possible, the facilitator used the same materials planned for training during the ALS pilot and operational meetings. The level of detail included and time spent on this activity was considerably less than the requirements for the pilot and operational meetings.
- Completion of ALS rating activity. The participants completed the ALS rating activity following the procedures planned for the ALS pilot and operational meetings. For the purpose of the study, the participant provided only one rating, specifically that related to the Proficient achievement level. To complete this activity, the participant:
 - Proceeded through the OIB, referencing the operational presentation of the items throughout, including functionality;
 - Considered the descriptions of the Proficient achievement level relative to the knowledge and skills needed to answer each item correctly; and

- Determined the place at which a just barely Proficient student would stop answering the items correctly with a 0.67 probability.

This rating activity sometimes required the participant to navigate between the computer having the OIB and rating software and the other computer displaying the NAEP TEL items as they appear to students during the test administration. Participants were instructed to think aloud and verbalize any challenges or confusion experienced while they were involved in the rating process. The facilitator asked the participants questions as needed to elicit commentary regarding any difficulties the participants experienced with the rating activity.

At the conclusion of the ALS-related activities, the facilitator led a discussion with each participant about their experiences in the study. The facilitator asked the following questions:

- Did you feel comfortable moving between the two computers for the activity?
- What, if anything, was the most difficult part of the rating activity?
- What, if anything, could we do to make the rating activity easier?
- How might the current set-up be used during discussions with other panelists about reasons for bookmarking the current page?

Results

Pearson reviewed the video and audio recordings of the study to document the time needed by participants to train for the ALS activity and to complete the ALS rating. Time was recorded to the nearest whole minute. The comments received were related to difficulties encountered by panelists and ways in which the activity could be improved.

The item review and rating rounds each used the two scenario-based tasks and the one block of discrete items for a total of 20 items. All participants used the full 30 minutes allotted for the item review. No participant was able to review all 20 items in that time. The median number of items reviewed was 14 with a low of 10 and a high of 16. The mean, minimum, and maximum time required across the five participants to complete the item rating rounds is shown in Table 8.

Table 8: Minutes Required for Item Ratings

Round	Mean	Minimum	Maximum
1	29	26	33
2	10	7	13

A participant received “credit” for working in the out-of-system item review for any minute in which the participant used that software for at least part of the minute.

The mean, minimum, and maximum percent of total time working on an activity that was spent on the out-of-system item review is shown in Table 9.

Table 9: Percent of Total Time Spent on the Out-of-system Item Review

Activity	Mean %	Min %	Max %
Item Review	10	0	17
Round 1	5	0	15
Round 2	8	0	23

The facilitator reviewed the recordings and noted the comments expressed by the participants during the debriefing session. The following summarizes their responses to each question.

1. Did you feel comfortable moving between the two computers for the activity?
 - All five participants said that they were comfortable moving between the two computers.
 - Two participants said the wheels on the chair made moving between the two computers easier.
2. What, if anything, was the most difficult part of the rating activity?
 - The three judges who completed the activities mentioned that it was difficult to remember the number of points a polytomous item was worth and what score point was being considered. Total score points available for an item and the score point being considered for the item were added on the second day of the study.
 - Two additional panelists mentioned that mastering the tabs and terminology was difficult.
3. What, if anything, could we do to make the rating activity easier?
 - One participant mentioned creating a split screen.
 - Another participant suggested that the instructions include a step-by-step approach to navigating between the tabs.
 - Three participants said nothing could be done.
4. How might the current set-up be used during discussions with other panelists about reasons for bookmarking the current page? (Participants had varied responses to this question and no common theme emerged.)

- A participant said a projector and screen would be helpful.
- Another participant suggested that enough room is allowed between panelists at a table so that everyone at the table could gather at one computer.
- Another participant said that all tablemates could find the same item in the item map.

Results of the usability study were used to make changes to the software and planned ALS processes prior to the first pilot study.

General Procedures Applied to the Pilot Studies and the Operational ALS

The sections below provide a description of each component of the standard setting process Pearson applied across all three studies. Variations from the general descriptions will be addressed under each study.

Recruitment and Selection of Panelists

Recruitment and selection of panelists occurred in two cycles. The initial cycle focused on identifying and selecting panelists for two events—the pilot and operational studies. However, as described later in this document, a second full pilot study was added to the procedures Pearson implemented to help ensure valid ALS results. Consequently, a second recruitment and selection cycle was initiated, following Pilot 2 to identify and select panelists for the Operational ALS study. All recruitment and selection efforts met the Governing Board policy requirements for the composition of the panel: 55% teachers in the subject area at the grade level of the assessment; 15% non-teacher educators (not K-12 teachers) in the subject area; and 30% general public (not educators) with educational training and/or experience in the subject area and direct experience with children in the age range of eighth grade students. The process also focused on the requirement that panelists' demographics should be representative of geographical region, gender, and race/ethnicity.

The objective of the recruitment plan was to meet the Governing Board requirements through the recruitment of broadly representative, well-qualified panelists to participate in the ALS activities. Recruitment targeted the creation of panels to reflect an overall balance of gender, race/ethnicity, geographic location, and type of TEL-relevant experience, as well as type of institutional affiliation. The processes implemented during the first cycle of recruitment provided panelists for Pilot 1 and the additional Pilot 2. At the recommendation of the TACSS, recruitment also targeted having at least one teacher of English Language Learners (ELL) and one teacher of students in Special Education programs as candidates for each

event. The teachers sought for these positions also had to meet the teacher panelist qualifications presented above. Additionally, a group of five extra panelists was established, two extra panelists for the Pilot 1 and three extra panelists for Pilot 2, as backups in the event that some panelists dropped out before the panel meeting. This backup panel group was targeted to include two 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, Pearson identified the panelists through an iterative multi-phase process focused on identifying and contacting qualified nominators; collecting and reviewing nominees; notifying nominees and collecting nominee information; and selecting and recruiting the sample of nominees to serve as panelists.

Identifying and Contacting Nominators

Panelist nominators were 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 had a strong background in technology or in providing professional development in TEL. These allied organizations were supplemented by additional organizations to increase representation and to increase the potential pool of candidates for the panels. The groups contacted to nominate panelists are described below. In addition, NCES posted a notice for NAEP state coordinators on its website inviting them to nominate people to serve as panelists and providing instructions for how to do so.

The following national organizations were contacted in the process of 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. (A not-for-profit organization with a STEM engagement program that aims to inspire young people to be science and technology leaders.)

In addition, state superintendents, heads of teacher organizations, school board presidents, and principals of public and private schools in the four NAEP regions were contacted directly and asked to propose qualified nominators for teacher and non-teacher educator panelists. Based on previous experiences in recruiting NAEP ALS panelists for Writing 2011 and Science 2009, the ratio of nominators from public schools to nominators from private schools was targeted at 9:1.

The process of recruiting panelists for the non-teacher educator group included contacting deans of a representative sample of technology and engineering higher education institutions, as well as leaders of STEM education centers. The goal focused on reaching 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 was made to balance the geographic representation of institutions of higher education. Pearson then contacted a representative sample of institutions to nominate people to represent the non-teacher educator panelists. A list of the institutions contacted is contained in Appendix B.

The group of panelists representing the general public was targeted to consist of individuals who were educated in and/or work directly in areas relevant to TEL. This included 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 involved contacting individuals in human resources or education offices of companies engaged in technology and engineering activities in each state. Nominators from nationwide companies were asked to nominate qualified nominees from each of the four NAEP regions who represented the diversity required of the panels. Companies were identified from the engineering and technology sectors to represent a broad array of occupations requiring training and experience in engineering and technology. A list of the organizations and companies contacted to nominate representatives of the general public is included in Appendix C.

Based on previous experiences in recruiting NAEP ALS panelists, Pearson staff estimated that 20 percent of the nominators would respond by submitting at least one nominee for consideration. In addition, they predicted 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, they established 1,275 as the target number of NAEP TEL-related organizations, companies, and other institutions of the types listed above to ask to provide nominations of panelists for the initial recruitment cycle. Once the decision was made to conduct a second pilot study, people who were qualified to be panelists but who were not selected for the original studies were contacted and asked to serve as panelists for the operational

ALS. Pearson staff also contacted nominators from the first phase and asked them to nominate additional people.

Selection of Panelists

Nominees were asked to complete an online questionnaire regarding their qualifications and experiences for serving on a panel. Candidates having the credentials required of panelists were contacted by phone to collect any missing information, to verify the information provided, and to confirm the willingness of the candidate to serve on the panel if selected. The goal was to select the most qualified panelists who were knowledgeable about TEL, while maintaining the goal to recruit 55 percent teachers, 15 percent non-teacher educators, and 30 percent members of the general public to compose each of the panels. Panelists recruited in each panelist group met the following qualifications:

Teacher panelist:

- At least five years of overall teaching experience,
- At least two years of experience teaching TEL in grade 8, 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,
- Curriculum director or content specialist at a 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),
- Not a former educator, and
- Familiar with students in grade 8 (e.g., as a parent or volunteer).

Pearson evaluated and rated potential panelists based on the number and importance of their professional credentials presented in their information materials. Persons having no distinguishing credentials were rated low. Persons having extensive professional 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, were rated very high. The rating scheme differed for each panelist type (teacher, non-teacher educator, and general public). Persons with the highest ratings were given top priority by placing the best-qualified candidates at the beginning of the candidate list. Persons were selected to reach the targets listed above, with those having the highest qualifications being the first selected each time. All panels were selected to have approximately equal proportions of males and females and equal proportions of persons from each of the four NAEP regions. Pearson also attempted to draw panels so that 20 percent of the persons self-identified as a minority.

Panelists' expenses were reimbursed for travel, meals, and lodging according to federal travel regulations. In addition to covering the direct expenses for panelists (consistent with federal travel regulations), panelists were given an honorarium of \$300 each for the four-day Pilot 1 study and \$500 each for five-day Pilot 2 and Operational studies to cover incidental expenses during their stay at the panel meetings. School districts were reimbursed for the cost of substitute teachers. Pearson acknowledged that the funds available to panelists were not commensurate with their contribution and emphasized to panelists that their participation in the NAEP TEL ALS represented an exceptional contribution to technology and engineering education in the United States (see Appendix D for samples of panelist recruitment letters and Appendix F for the lists of panelists by study).

[Advance Materials and Preparation of Panelists](#)

This section describes the briefing materials mailed to panelists prior to the pilot and operational ALS meetings. Panelists selected for a given study were sent letters inviting them to participate in the ALS process. Panelists who agreed to participate received correspondence thanking them for agreeing to participate and providing them with information about the meeting dates and location, and travel information. In addition, panelists received the items listed below for review prior to their assigned event. The agenda, briefing document, policy definitions, and ALDs are contained in the appendices. These materials were intended to serve as a foundation for successfully carrying out the ALS process.

- A draft Agenda (see Appendix N for the ALS meeting agenda)
- A briefing document overviewing the ALS process and describing the activities to be accomplished each day (Appendix E)
- 2014 NAEP TEL Framework

- NAEP TEL Policy Definitions and ALDs (Appendices H and G)
- Links to the full and abridged TEL Framework⁴

The letter sent to the panelists underscored the importance of the TEL Framework and the ALDs to the process and urged panelists to become familiar with those two documents prior to the ALS meeting.

Approximately two weeks prior to each study, an email and a letter were sent to panelists providing more detailed information regarding logistics, hotel and city information, transportation to and from the airport, check-in procedures, as well as a confidentiality agreement and taxpayer forms.

Panelist and Item Pool Division and Assignment of Forms

Division of Panelists into Subgroups

The NAEP ALS Process has used a split panel design since the 1991 studies. In past NAEP ALS meetings, the standard setting panel was typically split into two rating groups. Because of the amount of time predicted for review of and interaction with the technology-based items on the NAEP TEL assessment, the TACSS recommended that Pearson split the standard setting panel in the NAEP TEL pilot and operational ALS studies into three rating groups. For Pilot 1, there were 14 panelists with four or five panelists in each subgroup. For Pilot 2 and the operational ALS studies, there were 29 and 31 panelists, respectively, divided into three subgroups. To the extent possible, each of the groups of panelists was equivalent in terms of the designated panel attributes. Given the smaller size of each group in Pilot 1 (four to five per rating group), complete equivalence across all attributes was not possible.

For the Pilot 2 study and the Operational ALS study, Pearson divided each of the three panelist subgroups into two table groups of approximately five panelists each. This additional subdivision supported panelists' individual work by providing the space each panelist needed, and it facilitated table discussion. The demographic attributes used to recruit panelists were also used to assign panelists to subgroups and table groups, with the goal of maximizing the equivalence of the subgroups as well as the equivalence across table groups.

4

http://nagb.gov/content/nagb/assets/documents/publications/frameworks/naep_tel_framework_2014.pdf

<http://nagb.gov/content/nagb/assets/documents/publications/frameworks/tel-abridged-2014.pdf>

Division of Items into Subsets and Assignment of Forms

As was done for previous NAEP ALS studies, items were divided into item rating sets to limit the number of items reviewed by each panelist, reducing both the time required for the process and the potential for panelist fatigue. In order to construct lists of items ordered by difficulty, the NAEP TEL item pool of 20 scenarios and 97 discrete items was divided into three unique sets, A, B, and C, and a common set of items.⁵ Each unique item set was combined with the common set to form an OIL, which was assigned to one of three groups of panelists. Figure 1 depicts the structure of the item rating sets and their assignment to panelist groups. The item sets were constructed to be as equivalent as possible in terms of content area, item type, and item difficulty.

Items were assigned to each set based on (a) the item's assessment area (Technology and Society, Design and Systems, Information and Communication Technology), (b) item type (scenario-based or discrete), and (c) item difficulty⁶. Items remained in the organizational units (e.g., scenarios) used for administration of the assessment. The common item set consisted of items that were on one administration form and were selected for possible release to the public. The common item set was designed to provide a source for examples that facilitators could use during group discussions with panelists and to offer an empirical basis at Round 1 to evaluate how well the groups were functioning as pseudo-replications.

As shown in Figure 1, each of the three panelist subgroups was assigned a unique set of items to review and rate. In addition, all panelist subgroups reviewed and rated the same common set of items.

⁵There were 98 discrete items but one was dropped from the final scaling of the items.

⁶ Item difficulty was calculated for dichotomous items using each item's scale value for which a correct response probability of 0.67 was expected. Item difficulty was calculated for each score point of a polytomous item where the probability of being awarded that score point or higher was 0.67. The response probability of 0.67 was based on an Item Response Theory (IRT) model and is an accepted convention in standard setting.

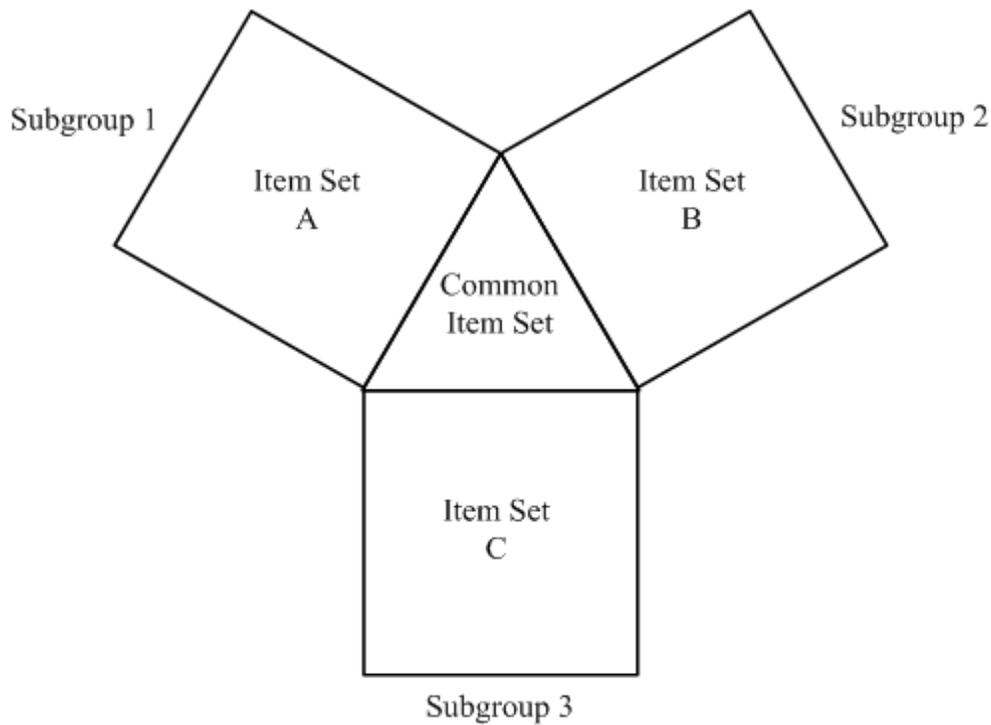


Figure 1: The Assignment of Item Sets to Panelists Subgroups

Characteristics of the three item sets are presented in Table 10. As mentioned in the Introduction, two computers were used for item presentation, recording of panelist input, and sharing feedback. Given the presentation of items through an electronic format, the usual phrase used with item mapping procedures, "ordered item booklet," was replaced with "ordered item list" (OIL).

Selected response items and other items with a single correct score point are represented one time in the OIL at their RP67 scale value. However, each score category greater than the lowest one is represented in the OIL for polytomous items. The correct response for dichotomous items and the score categories greater than the lowest one for polytomous items constitute the rating elements in the OIL. The first section of the table shows the number of rating elements contained in each OIL by TEL assessment area and total combined. The next section shows the number of distinct scenarios and the number of scenario-based and discrete (non-scenario-based) rating elements in each OIL. The third section of the table shows the number of dichotomous and polytomously scored items and the total number of items in each OIL. The last section presents information about the average, median, and range rating element scale locations (RP67) for each OIL.

Table 10: Characteristics of Ordered Item Lists

		OIL 1	OIL 2	OIL 3
Number of Elements per Assessment Area	Design and Systems	42	44	47
	Information and Communications Technology	39	48	45
	Technology and Society	48	38	39
Total Number of Elements for Ratings		129	130	131
Scenario-Based and Discrete Elements	Distinct Scenarios	8	8	8
	Scenario-Based	85	85	84
	Discrete Items	44	45	47
Number of Unique Items by Type	Dichotomous	27	19	23
	Polytomous	62	71	67
	Total	89	90	90
Scale Score Information	Average/Median Scale Score	157.8/152	156.5/150	158.5/153
	Minimum/Maximum Scale Score	36/300	35/300	57/274

ALS Methodology

For the NAEP TEL ALS studies, several possible standard setting methodologies were considered for recommending cut scores for the NAEP TEL. Pearson’s evidence-based standard setting procedure (McClarty et. al., 2013) was initially considered. The Governing Board had originally requested that an evidence-based approach be used to provide evidence of external validity for the standard setting results. However, the TEL Framework is not widely taught in most schools as a stand-alone instructional curriculum. Rather, aspects of TEL are addressed in a wide variety of educational experiences and courses. Pearson attempted to identify sources of relevant external validity evidence and concluded that 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 foregoing the external validity evidence as part of the ALS process and the Governing Board staff agreed.

Pearson’s proposal was to implement an item mapping process in which panelists would make criterion-referenced, content-based cut score recommendations over

three rounds of standard setting (Lewis, Mitzel, Mercado, & Schulz, 2012). The item mapping approach satisfied the main considerations when choosing an appropriate standard setting methodology. According to Hambleton and Pitoniak (2006), the method is appropriate for the item types and for item scaling, the judgments are likely to be completed in a reasonable amount of time, and the procedure is widely accepted in the measurement field and supported by current validity evidence.

In the item mapping methodology, items are arranged along a continuum using IRT-based item difficulty estimates. Standard setting panelists receive extensive training and preparation for applying the ALS procedures to recommend performance level thresholds or cut points. They locate the cut points by identifying items that examinees performing at different achievement levels should be able to answer correctly with a specified level of probability. Variations of item mapping methodologies have been implemented in previous NAEP ALS studies (Math 2005, Grade 12; Science 2009; Judgmental Standard Setting Studies, 2011). The general procedures were adapted to the characteristics and complexities of the NAEP TEL assessment, specifically the scenario-based, interactive nature of the items.

The following sections provide a general description of the training, standard setting activities, and feedback that were used for the pilot studies and the operational ALS meeting. Specific differences in how they were implemented in each study follow the general description. The facilitator guided the panelists through all training and standard setting activities using PowerPoint slides. The slides used for the operational ALS meeting are contained in Appendix O.

Orientation to the Methodology

Dr. Steve Fitzpatrick⁷, the project director, opened the first day by welcoming the panelists and introducing the Pearson staff, the process and content facilitators, the Governing Board representative(s), and the observers. He clarified the roles and responsibilities of each person and to whom the panelist should address particular questions. Panelists were informed, for example, that while the observers have expertise in standard setting, they were not resources for panelists' questions regarding the ALS processes being implemented.

Following the introductions, Dr. Sharyn Rosenberg provided an overview of NAEP and the role of the Governing Board. She reviewed historical information about NAEP achievement levels, the Governing Board policy on setting achievement

⁷ Dr. Fitzpatrick led the pilot activities; however, due to illness, he was not able to attend the Operational ALS meeting. Dr. Tim O'Neil, from Pearson, was acting director of the Operational meeting and provided the opening on the first day.

levels, the purpose of setting achievement levels, and their importance to the Governing Board's overarching goals and responsibilities.

Dr. Fitzpatrick then provided an overview of the NAEP TEL ALS process in which the panelists would be engaged over the next several days, including an overview of standard setting and a review of the full agenda. He provided detailed information regarding the recruitment and selection of panelists and the composition of the panel and emphasized the security of NAEP TEL materials and the need for panelists to adhere to the meeting parameters such as putting away all personal electronic devices during all ALS sessions.

Training and Preparation

On the first day of the study, each panelist received a folder containing materials to support panelists' work such as an updated agenda, paper copies of login directions for various aspects of the ALS process, the abridged NAEP TEL Framework, and the TEL ALDs. Panelists were instructed to use these materials throughout the ALS process.

Panelists received an orientation to and instructions regarding the dual-computer arrangement. They received information regarding the purpose and use of each computer as well as support from the facilitators while exploring the dual-computer arrangement. The computers were identified to panelists as the *NAEP computer* and the *ALS computer*. The NAEP computers were from the 2014 operational administration of TEL. They were reconfigured by NCES Alliance contractors to include an actual form of the NAEP TEL assessment and to display all scenario-based tasks and interactive discrete items so that panelists could view test items in context. The ALS computer contained all of the tools, data files, and other documents and files panelists needed to use throughout the fully computerized ALS process.

Taking a NAEP TEL Assessment

After general orientation to the ALS process and training on the use of the computers, panelists used the NAEP computers to take a form of the 2014 NAEP TEL assessment as a student would. NAEP computers were used to administer a form of the 2014 assessment under the same conditions as that experienced by grade 8 students and this allowed panelists to experience the NAEP TEL from a student's perspective. Panelists were encouraged to pay attention to the different types of assessment items they encountered and to think about the relative difficulty and complexity of the items for eighth grade students. After all panelists had completed the TEL assessment, they used the ALS computer to review the items on the form relative to the scoring criteria and rubrics for the items. While panelists did not have access to their responses to score them, the opportunity to review the items and scoring criteria directly after completing the assessment

provided insights into how the items were scored and the types of performances expected.

Understanding the NAEP TEL Framework and the NAEP TEL ALDs

Panelists were guided through the process of developing a deeper understanding of the assessment framework for the NAEP TEL. The content facilitator, Dr. Moye, described key aspects of the framework and its role as the foundation for both the TEL assessment and the ALS process as applied to TEL.

Panelists engaged in discussions about the framework, with an emphasis on understanding the interaction between the three TEL content areas (Technology and Society, Design and Systems, and Information and Communication Technology) and the three TEL practices (Understanding Technological Principles, Developing Solutions and Achieving Goals, and Communicating and Collaborating). These interactions provided a foundation from which panelists were better able to understand the complexity and functioning of TEL assessment items. Having clarity regarding the framework also supported panelists in developing an understanding of what students need to know and be able to do to perform at each TEL achievement level.

To complete panelists' preparation for participating in the ALS process, the content facilitator led panelists through an in-depth discussion of each performance level of the TEL ALDs. Through this discussion, the content facilitator ensured that panelists understood the logic behind the ALDs and were prepared to apply that understanding as the ALS process commenced.

Describing the Knowledge and Skills for Responding to Items

The process and content facilitators worked collaboratively to provide panelists with training on how to develop knowledge and skills statements for items within the OIL. They talked with panelists about the importance of gaining a detailed, *structured* understanding of the assessment, and about what the assessment required of students as they progressed up the continuum of student achievement, as measured by the items. Through the training and context provided, panelists gained an understanding of the purpose of the review as establishing a strong base of understanding about the NAEP TEL assessment that they would apply at each step of the ALS process. The facilitators took panelists through the process of describing knowledge and skills using a set of examples that represented different types of items students encounter on the NAEP TEL. This provided panelists with a shared experience of using the item, the scoring information, and their knowledge of the TEL framework to describe the knowledge and skills a student would need to get an item correct or perform at a particular score code for a polytomous item.

Panelists used both computers to complete their item reviews and to create knowledge and skills statements for their assigned items. The NAEP computer

provided access to the scenarios and to the discrete interactive items, so panelists could review these items in the context they were administered. The ALS computer was used for review of all discrete, non-interactive items and for recording of panelists' knowledge and skill statements for all reviewed items.

Panelists reviewed an assigned subset of the items in their OIL. All items in an OIL were common to the panelists at the same table and assigned for review by at least two panelists. After all item reviews were complete, panelists at the same table shared information about reviewed items to ensure all table members had information on every item in their OIL. Throughout this item review process, facilitators emphasized the importance of this work to ensure that panelists developed the knowledge foundation and the deep understanding of the items and of the assessment as a whole that they would need for their later work on identifying cut scores.

Developing Borderline Achievement Level Descriptions

The content and process facilitators guided panelists through the process of creating borderline achievement level descriptions (BALDs) for the Proficient, Basic, and Advanced achievement levels. Prior to developing the BALDs the process facilitator provided training and information to ground the panelists in their understanding of what the BALDs represent and their importance to the achievement levels-setting process. Figure 2 is one illustration used to help panelists conceptualize borderline performance.

Conceptualizing Borderline Performance

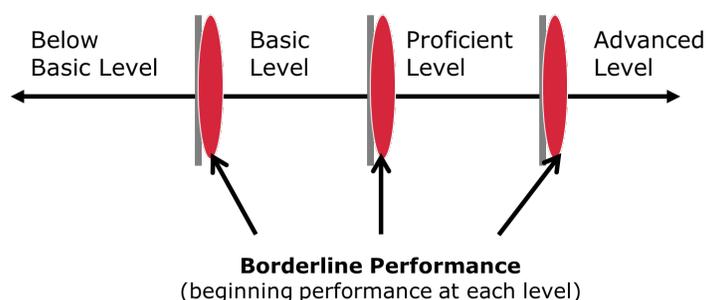


Figure 2: Illustration of Borderline Performance

Panelists then worked iteratively as individuals and in groups to come to an agreement on the descriptors for performance that is "just good enough" to belong in a particular achievement level as well as the associated knowledge and skills that define the transition of performance across adjacent categories. To develop the borderline description for Proficient, followed by Basic, and then Advanced, panelists used the NAEP TEL grade 8 ALDs and the NAEP policy definitions, as well as their understanding of the knowledge and skills required to perform well on the

NAEP TEL items. (The policy definitions and ALDs provided to the panelists are available in Appendix H and Appendix G, respectively.)

Using the ALS Methodology to Set Cut Scores

Once panelists completed the training and preparation activities, the process facilitator provided training on the modified item mapping methodology used for the NAEP TEL ALS. Panelists gained a conceptual understanding of the NAEP TEL achievement levels-setting process used to establish cut scores on the continuum of possible NAEP TEL scores. They also gained an understanding of how actual student performance data from the 2014 NAEP TEL administration was used to order the items, by difficulty, in the OILs that panelists would use for identifying cut scores. As part of their training on the OILs, panelists gained insight into the placement of score codes for polytomous items at multiple locations in the OIL, with the number of placements dependent on the total number of score codes possible for a particular item.

Next, panelists learned about the mapping of items and students onto the same scale and were provided graphics to illustrate this concept (see figure 3). They also became familiar with the use of 0.67 as the response probability used for NAEP TEL to place items or score codes on the score scale to create an item map. This response probability was described as representing a reasonably high probability (i.e., 2/3) of a correct response (or of achieving a particular score code) without being overly demanding.

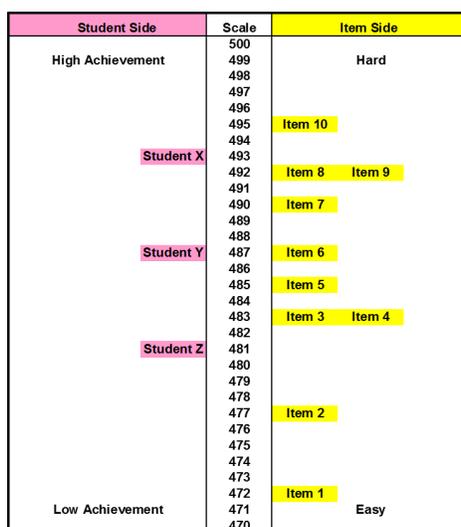


Figure 3: Mapping Student Achievement and Item Difficulty

Panelists then received training on using the information about item order and what the items represent in terms of student knowledge and skills to operationalize the minimally qualified performance described in the BALDs and recommend cut scores to represent the lower borderline of performance at each achievement level. The

term 'bookmark' was used to help panelists conceptualize the placement of their recommended cut scores at specific points in the OIL.

Panelists participated in a practice round followed by three rounds of standard setting. For each round of the ALS process, panelists were encouraged to review the ALDs, BALDs, and NAEP policy definitions to conceptualize performance at the border of a particular achievement level. Panelists then started with the easiest item in the OIL. For each item, they were instructed to:

- Ask themselves "Does borderline performance that is at the just barely Proficient (or Basic or Advanced) level correspond to a 2/3 chance or better of a correct response to this item?"
- Identify the set of items that constitute the zone of borderline performance for that achievement level. That is, identify the area of transition within which the responses to the question above shift from mainly "Yes" to a mix of "Yes" and "No."
- Consider the items in the zone holistically as the area within which the cut score is set.
- Carefully review each item in the zone and identify the last item in that zone that represents a solid 2/3 probability of a correct response with regard to the borderline performance. (That is, this item represents the hardest item the borderline students will have a 2/3 chance of answering correctly.)
- Bookmark this item to represent the location used to compute the cut score for that achievement level (see figure 4).

Zone	
Order #	Answer
9	Y
10	N
11	Y
12	Y
13	N
14	Y
15	N

Bookmark the LAST confident YES (i.e., item 12)

Figure 4: Identifying the Bookmark in the Zone⁸

Facilitators directed panelists to work independently to determine their individual recommendations for each round. Between each round, facilitators asked panelists

⁸ The items are arranged in the list from the easiest item at the top to the most difficult item at the bottom.

to discuss their understanding of the borderline descriptions and the knowledge and skills required at the lower border of each achievement level, and to make adjustments as needed to the borderline descriptors. Feedback from the previous round was provided prior to rounds two and three of the ALS procedure.

Round 1: Understanding the Assessment and Student Achievement

Panelists applied the item mapping process to the items and score codes in their OIL to place three bookmarks—Proficient, Basic, and Advanced—at the lower border of each of these achievement levels. Again, panelists were instructed to use the BALDs, ALDs, Policy Definitions, knowledge and skills statements, and item scoring information when answering the question about the probability of a correct response to each item.

Round 2: Using Feedback

Prior to making Round 2 judgments, panelists had the opportunity to discuss the table and whole group information provided from Round 1. Panelists received feedback regarding the following (see Figures 5 and 6):

- Individual and whole group cut scores presented as the median value for the table or the group as a whole.
- Table Level Rater Location Chart: Showed the distribution of cut scores by Table.
- Whole Group Rater Location Chart: Showed the distribution of cut scores for the whole group and the group cut score (dashed line shows the group median for each cut score).

Prior to discussing the Round 1 feedback, Dr. Nebelsick-Gullett explained how each panelist's bookmarks were converted to a cut score on the pseudo-score scale⁹ and that the median scale score was used to compute the cut score for a table or the whole group.

⁹ Pseudo-NAEP scales were used for all components of the study (Pilots and Operational) to avoid the risk of early release of cut score information. As in past ALS studies, the pseudo-NAEP scale was a linear transformation of the NAEP reporting scale.

Panelist ID	Item Order Number			Scale Score		
	Basic	Proficient	Advanced	Basic	Proficient	Advanced
MDS1NE3	27	58	93	323	348	380
NJB1PA1	47	69	115	342	356	412
SBD1AL2	9	34	129	300	332	500
SJJ1GA5	23	68	118	318	354	420
WER1UT4	23	39	73	318	336	361
Table Median				318	348	412
Group Median				317	342	381

Figure 5: Table Level Feedback

NAEP TEL Achievement Level Setting
Round 1, Table 1

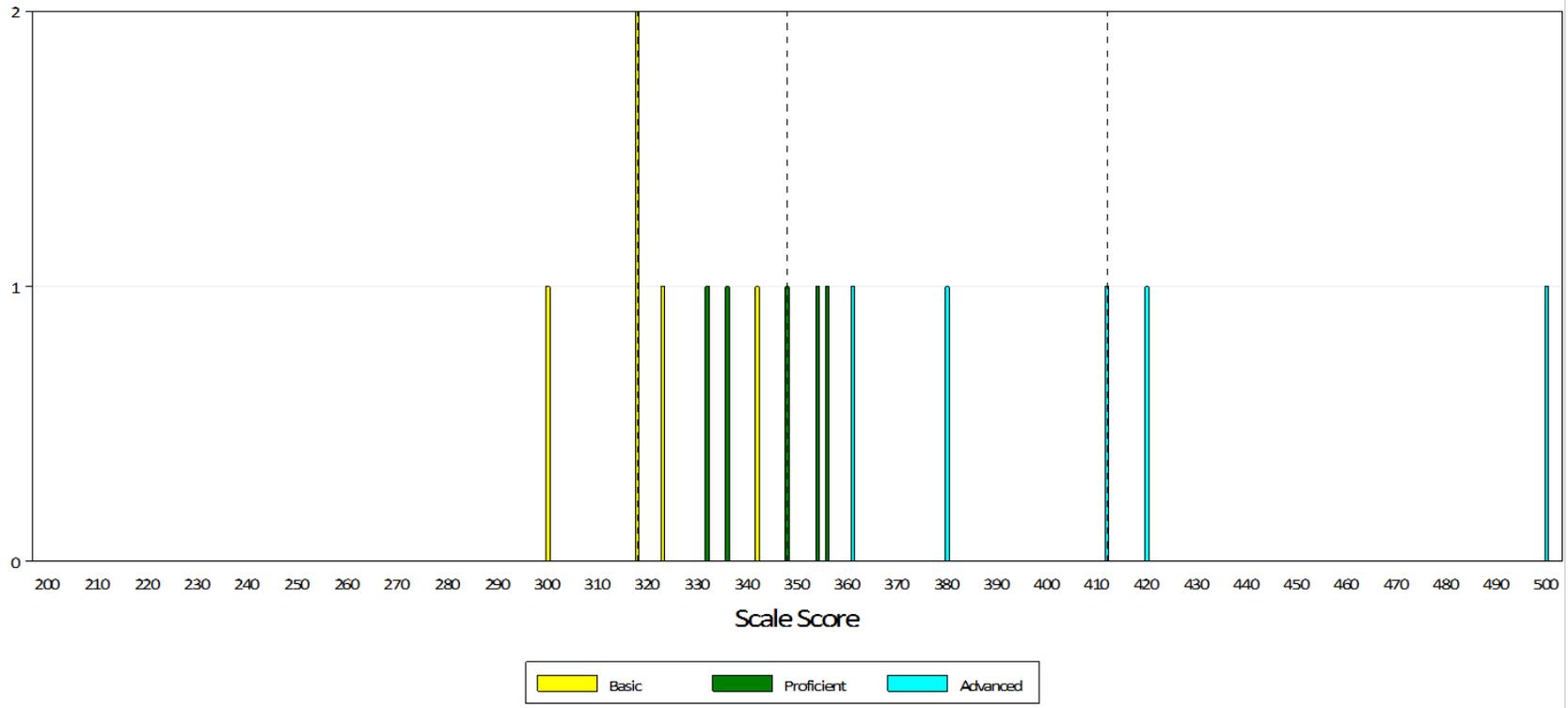


Figure 6: Rater Location Chart – Table Level

Feedback from Round 1 was discussed only at the table level. In preparation for Round 2, table groups were directed to:

- Examine the range of cut scores selected for each level.
- Examine the data showing the median values of the cut scores for panelists at their table and for the whole group.
- Note areas where cut scores for one achievement level overlap with another achievement level.
- Discuss the cut scores selected and the rationales for the judgments.
- Discuss in order from lowest to highest the cut scores selected by panelists for each achievement level.

After reviewing and discussing the Round 1 data at their table, panelists had the opportunity to make modifications to the BALDs at any or all levels. The process facilitator reminded panelists that it was acceptable to modify some, all, or none of their bookmark placements during Round 2. Panelists used the refined BALDs along with all the other sources of information used in Round 1, and the data from Round 1, to individually select bookmarks for Round 2.

Round 3: Understanding and Using Consequences Data

The process followed for Round 2 was replicated for Round 3 with the following additions:

- Use of the NAEP TEL Item Map. Dr. Nebelsick-Gullett provided panelists with training on interpreting and using the information on the item map. The item map depicted the location of each item or score code along the score scale in increments of three scale score points. Items were grouped into columns for each assessment area and were color coded to indicate the unique and common items in each OIL. She explained how to use the item map to understand the relative difficulty of items in the OIL within and across the three TEL content areas. Each panelist used a paper copy of the Item Map (see Appendix I), along with the information from Figure 5 showing cut scores for each table group and the whole group, to mark the location of the score level representing each of the three cut scores for the table group and for the whole group.
- Use of the Consequences Data Chart. This chart (see Figure 7) was constructed by applying the Round 2 whole group recommended cut scores to the NAEP TEL performance distribution data from 2014. The percentages shown represented the percent of students in each achievement level based on the Round 2 recommendations.

NAEP TEL Achievement Level Setting
Percentage of Students at Each Achievement Level
Round 2

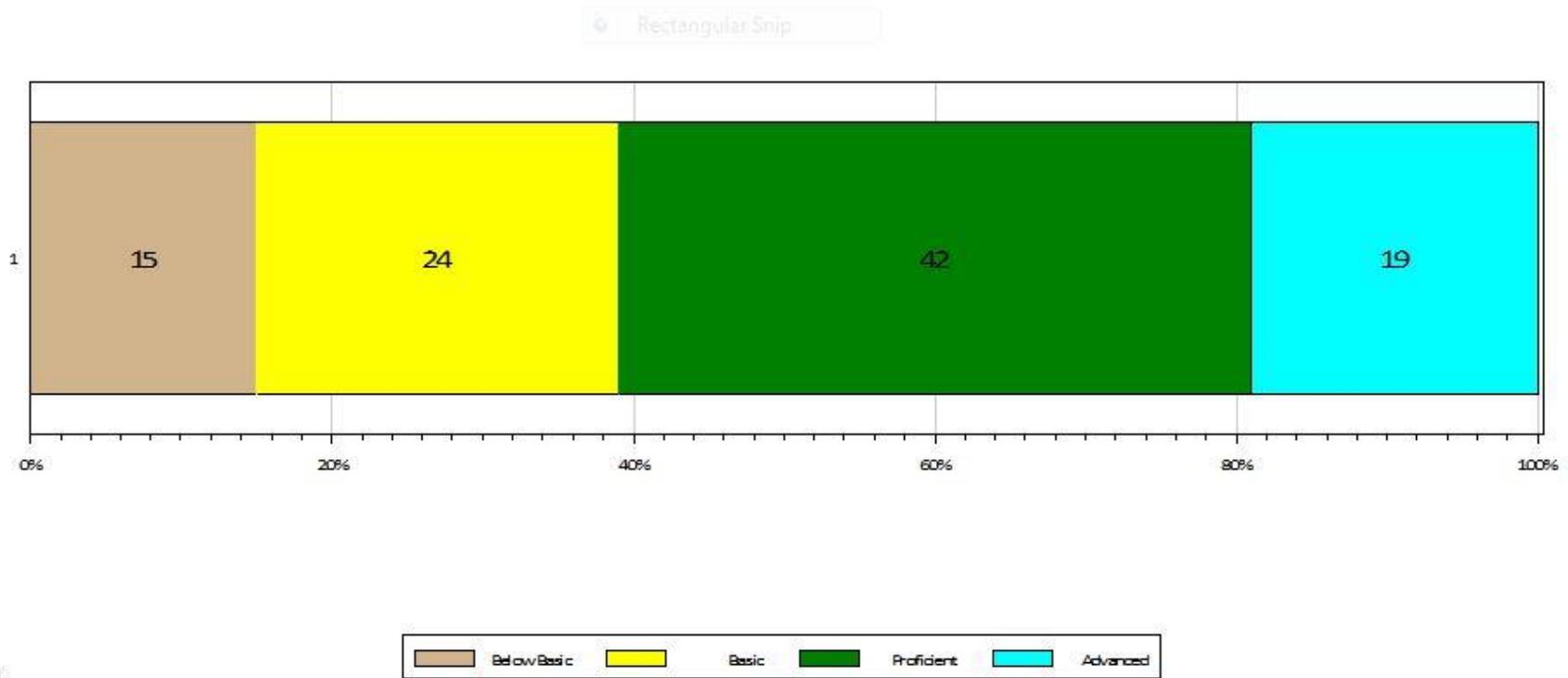


Figure 7: Consequences Data

The item map and consequences data were discussed as a whole group. The group was directed to discuss the cut scores using information about their location on the item map as well as with respect to the consequences data, asking themselves if the consequences data seemed reasonable, given the ALDs, the assessment items reviewed, and what was known about instruction in the area of technology and engineering literacy.

After reviewing and discussing the Round 2 data, panelists had the opportunity to make modifications to the BALDs at any or all levels. Again, the process facilitator reminded panelists that it was acceptable to modify some, all, or none of their bookmark placements during Round 3. Panelists used the refined BALDs along with all the other sources of information used in Rounds 1 and 2, and the data from Round 2, to individually select bookmarks for Round 3.

Completing the Consequences Review

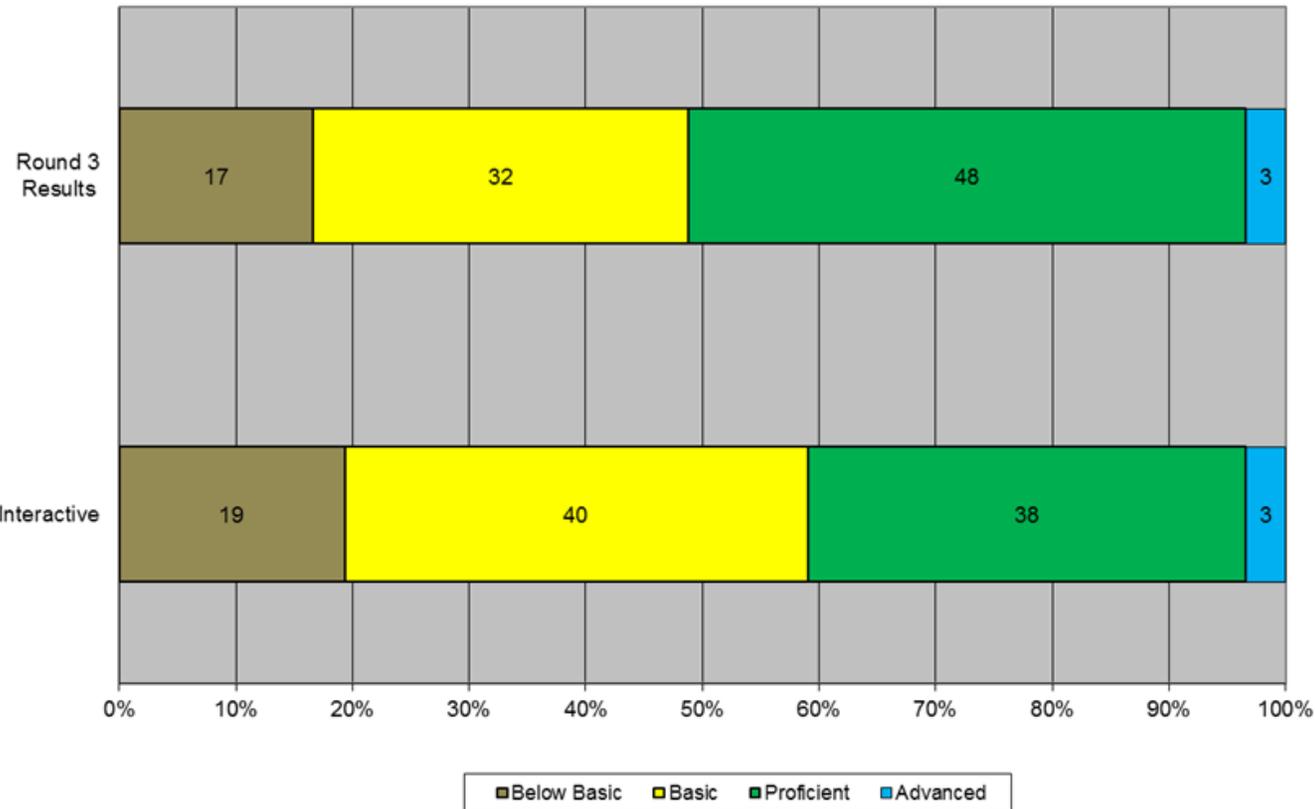
Following the Round 3 rating activity, panelists reviewed and discussed the whole group feedback from Round 3, which was similar to that provided for the other rounds. This review was briefer as panelists would not be going through a fourth round of the ALS process. Instead, panelists completed the consequences review process. For this process, panelists first completed the *Pre-Consequences Questionnaire* provided in Appendix J. This questionnaire asked panelists to rate their understanding of the consequences data from Round 3 and then to evaluate the percentage of students in each performance category, based on the Round 3 recommended cut scores. For each achievement level—Proficient, Basic, and Advanced—panelists indicated their level of agreement that the percentage of students that would fall in the particular category reflected their expectations. Finally, panelists were asked if they would change the Round 3 cut scores for one or more of the achievement levels.

Once the pre-consequences questionnaire was complete, panelists moved to the second component of the consequences review process. The process facilitator provided panelists with the Round 3 consequences data (Figure 8) in a format that allowed a panelist to make adjustments to a cut score and see the consequences of that change in terms of the percentage of students now in the achievement level categories impacted by the adjustment. The percent of students at each achievement level based on the Round 3 cut scores was depicted in the top bar of Figure 8. The row in the table at the bottom of the figure labeled *New Scale Score Value* allowed the panelist to enter a set of cut scores different from the Round 3 scores and observe the effect on the percent of students in each level. The lower bar showed the results of changes panelists made in the table at the bottom of Figure 8. If a panelist made a change to the Round 3 cut score by placing a different cut score in the bottom row of the table (labeled *New Scale Score Value*), the impact of that change was displayed in the lower bar. The Basic and Proficient

cut scores have been changed in the figure. Panelists were given the following directions to complete this process:

- Consider all of the information used in Round 3, the Round 3 feedback, and the post Round 3 discussion.
- Consider the BALDs for each level.
- Decide if you believe an adjustment is needed to any of the cut scores to more accurately reflect the BALDs. Before adjusting the cut scores, examine the items that fall in the area of the cut score you are considering against the BALD for that borderline performance level.
- If you still believe the new cut score is justified, make the adjustment to the cut scores using the table below the chart and examine the impact on the percentages in each category. (If you change a cut score, you will provide a rationale on the final questionnaire for that decision.)
- Save your final recommendations in the document and record them on the *Consequences Table* in your folder.
- Complete the *Post-Consequence Review* (Appendix J).
 - Enter your final cut score recommendations even if they did not change from the group recommendations.
 - Complete the question asking you for the rationale behind any changes made to the Round 3 cut scores.

Percentage of Students at Each Achievement Level



	Basic	Proficient	Advanced
Round 3 Cut Scores	316	351	409
New Scale Score Value	320	360	409

Figure 8: Consequences Review Chart

Selection of Exemplar Items

For the final step in the NAEP ALS process, panelists identified potential exemplar items for each achievement level from a pool of items designated for release to the public. Facilitators provided panelists with information regarding the items or score codes aligned with each achievement level and instructed panelists to decide whether an item or score code aligned with a particular level would be a good representation of performance within the entire range of that achievement level.

Exemplar items are a part of the official set of information recommended to the Governing Board as part of the achievement levels-setting process. These items serve to communicate to the public the types of knowledge, skills, and abilities that are required for performance within the Basic, Proficient, and Advanced NAEP achievement levels. The role of exemplar items in communicating performance on the NAEP TEL is especially important because this is a 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 blocks marked for public release by the National Center for Education Statistics (NCES) determined the set from which exemplar items were selected. The released blocks included four scenarios, each consisting of a set of items related to the scenario, and 20 discrete items. Some of the blocks marked for release were common to panelists in each of the three panelist rating groups.

Selection of exemplar items was the last of the judgments that panelists were asked to make in the ALS process. At this point, panelists were very familiar with both the achievement level descriptions and the items in their OIL—some of which were in the released blocks. They were well prepared to make the judgments regarding items that may serve to represent the achievement levels. The facilitators provided instructions regarding (a) the purpose of the task and the statistical criteria used for selecting the items for panelists' consideration, (b) the use of an item's statistical information to aid in panelists' evaluation of the item as a potential exemplar, and (c) the fact that the most important consideration was the relationship between the achievement level descriptions and the knowledge, skills, and abilities assessed by each item.

The panelists were provided with the set of items designated for release categorized into the achievement level at which the item had a response probability of 0.67 at a scale score point within the Basic, Proficient, or Advanced achievement level cut

score range. Each panelist reviewed only the potential release items that appeared in that panelist's OIL. The items were presented in an Exemplar Item List to show the following:

- The position of the item or score code in the panelist's OIL,
- The scale score location (i.e., the scale score at which the item or score point had a 0.67 response probability),
- The item ID with a link to a document showing
 - The item image
 - The item label, score code, and maximum score code,
 - The answer key for a multiple-choice item or a link to the scoring rubric for a polytomous item, and
 - The knowledge and skill statements for each time as well as other comments the panelists had entered in their OIL.

Items and score codes were assigned to achievement levels based on their RP 0.67 scale scores. Items with an RP 0.67 scale score equal to or greater than the Basic cut score and less than the Proficient cut score were assigned to the Basic level. Assignment of items to the Proficient level were made using the same process. Items with an RP 0.67 scale score equal to or greater than the Advanced scale score cut were assigned to the Advanced level. Pearson also calculated both the average response probability within the assigned achievement and the response probability at each of the three scale score cut point; both values were shared for each item in a panelist's potential exemplar group. The process facilitator gave panelists the following specific directions:

- Examine the item-level data provided. The items shown are items designated for release that were in the set of items panelists reviewed in their OIL.
- Make sure to also review the comments associated with each item or score code.
- Consider the ALD for the achievement level the item or score code is associated with as a possible exemplar.
- Decide whether this item or score code is a good representation of solid performance within the entire range of the achievement level. Remember that for this activity, the focus is the entire range of the achievement level, not solely the borderline portion.
- Rate the item as "Should be Used," "Might be Used," or "Should not be Used" as an exemplar for the specified achievement level.

Process Evaluation Procedure

The validity of standard setting outcomes depends, in part, on evidence of the procedural validity of the processes implemented. One source of evidence of procedural validity for the NAEP TEL assessment resulted from process evaluation questionnaires given to panelists at twelve points in the ALS process. The

questionnaires included both selected-response and open-ended questions that addressed the panelists' understanding and evaluation of instructions, tasks, and materials as well as their comfort level with particular processes and their confidence in the results. Process Evaluation Questionnaires were completed using the Survey Monkey¹⁰ online survey tool at the following points of the ALS process:

- End of Day 1 Questionnaire
- End of Day 2 Questionnaire
- Practice Round Questionnaire
- Pre and Post Round 1 Questionnaires
- Pre and Post Round 2 Questionnaires
- Pre and Post Round 3 Questionnaires
- Pre and Post Consequences Questionnaires
- Final Process Evaluation

Most responses were collected on Likert scales, but several responses were narratives that addressed specific aspects of the process. The Likert scales were either five-point agreement scales with 5 representing "strongly agree" and 1 representing "strongly disagree," or three point scales focused on the amount of time or amount of detail provided. For the three-point scale, a rating of 4 represented "too long" or "too much detail," a rating of 3 represented "about right," and a rating of 2 represented "too short" or "too little detail." Most questions on the *Pre-Consequences Questionnaire* and one on the *Final Process Evaluation* were yes/no responses scored as 1 or 0.

Access to the questionnaire was a two-step process for Pilot 2 and the ALS meeting during which the panelist accessed a document on the ALS computer that contained hyperlinks to each questionnaire. The hyperlink for each questionnaire took the panelist to a login page for Survey Monkey that required panelists to enter the panelist ID provided in their packets. Subsequent pages contained the survey questions. Panelist responses were reviewed at the end of each day or between steps in the process. Sources of confusion were identified for clarification with individual panelists or the group as a whole. The process evaluation questionnaires are presented in their entirety in Appendix J.

Facilitator Guide and Training

The ALS meeting for the NAEP TEL assessment included a content facilitator and a process facilitator. The content facilitator was selected for his TEL expertise and

¹⁰ <https://www.surveymonkey.com/>

experience as one of the members of the Steering Committee of the TEL Framework development team. The process facilitator was selected for her expertise and experience in conducting ALS meetings.

The content and process facilitators were trained to implement the procedures as designed. In collaboration with Pearson and the Governing Board COR, both facilitators developed PowerPoint presentations to use throughout the ALS implementations. In addition, the NAEP TEL Facilitator Guide was developed for facilitator use. The guide included a script for providing instructions, a description of the activities, and examples of tables, graphs, and other feedback.

Facilitators attended a half-day, web-based training prior to each of the three studies. The project director led the training. In addition, the facilitators and the project director did a walkthrough of the entire agenda at the meeting site the day before each of the three studies was conducted.

Pilot Studies

Pilot 1

This section of the report describes only those features of Pilot 1 that differ from the procedures described under the previous section on general procedures. More specific information is given in this section about how those procedures were implemented.

The panel of 14 teachers, educators, and the general public selected for Pilot 1 was convened March 16-19, 2015, to participate in a pilot implementation of the achievement levels-setting process planned for the grade 8 NAEP TEL assessment. This pilot study was designed to test the methods planned for the ALS operational study and to evaluate the outcomes of implementing the procedures.

Panelists

Table 11 summarizes Pilot 1 panel members' demographic information.

Table 11: Pilot 1 Panelist Demographics

Demographic Variable	Attributes	Percent
Panelist Type	Teachers	72%
	Non-Teacher Educators	14%
	General Public	14%
Gender	Female	43%
	Male	57%
Race/Ethnicity	White/Non-Hispanic	65%
	Black/Non-Hispanic	14%
	Native American/Mixed	7%
	Hispanic	7%
	Other	7%
NAEP Region	Midwest	7%
	Northeast	29%
	South	43%
	West	21%
Students with Disabilities	Experienced	93%
	Not experienced	7%
English Language Learners	Experienced	100%
	Not experienced	0%

Pearson organized the panelists into three table groups as depicted in Figure 9.

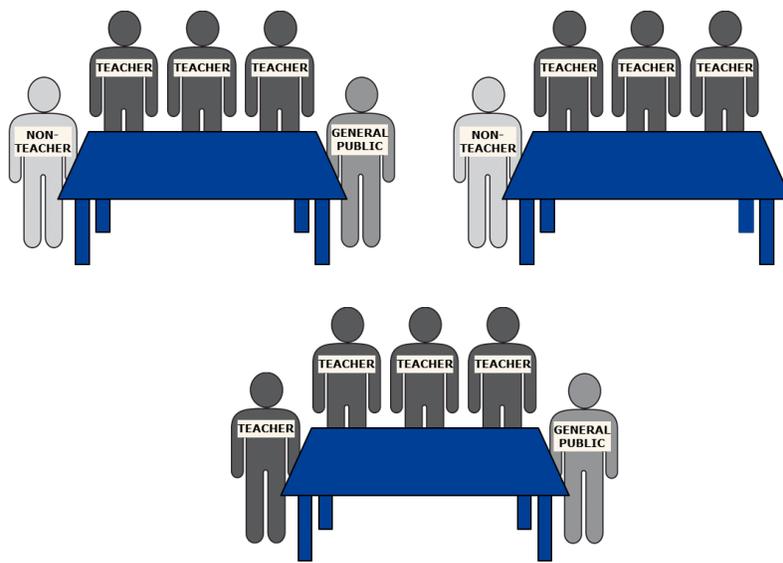


Figure 9: Composition of Table Groups for Pilot 1

On-Site Materials

The materials in panelists' folders for Pilot 1 included:

- Meeting Agenda
- Overview Booklet
- NAGB Policy Definitions and the NAEP TEL Achievement Level Descriptions
- Technology and Engineering Literacy Framework for the 2014 NAEP

Logistics

Room Configuration

In addition to the panelist tables, the room was configured (see Appendix L) such that the facilitators were positioned at the front of the room and all others (observers, Governing Board representative, and Pearson team members) were positioned at tables in the back and sides of the room.

Observers

Four observers were present for Pilot 1. They included two members of the TACSS, Dr. Barbara Dodd and Dr. Mary Pitoniak, and two personnel from the National Center for Education Statistics, Dr. Bill Tirre and Dr. Amy Yamashiro. Dr. Sharyn Rosenberg from the Governing Board was also present.

Procedures

The judgmental procedure used was an item mapping process as previously described. The process was implemented using web-based software, named OPLS, developed by Measurement Incorporated. The OPLS software was considered appropriate to the TEL ALS procedures, based on its functioning during the dual-computer usability study and its use in other major standard setting studies. The software functioned as anticipated for item review. However, it was not developed to group the scenario-based items by scenario, and this caused confusion for panelists as they completed the item reviews. When panelists began implementing the standard setting rounds, unanticipated interruptions in the functioning of the software occurred. These interruptions were apparently related to the complexity of the TEL ALS procedures, such as the use of three item-rating groups and three separate item lists. The item order and links to items were misaligned in some portions of the item lists in the system. Between rounds, Pearson staff corrected the alignment of items and links to items. In addition to these problems, there were challenges to:

- the processing of panelists' cut scores,
- producing feedback needed for panelists' immediate use between rounds of standard setting, and
- the processing of panelist's questionnaire response data.

To address these challenges, panelists were asked to record their bookmark selections and questionnaire responses on paper. Pearson team members then collected these, processed the data, and produced results outside of the software system. Feedback from each round was provided on paper.

All planned ALS activities were completed, although the schedule was adjusted to accommodate the interruptions to the process. A summary of the day-by-day activities is provided in Table 12. The summary reflects the actual schedule implemented for Pilot 1 rather than the planned schedule. On the first day, panelists took a version of the grade 8 TEL assessment, received instruction in the TEL framework and developed understanding of the achievement level descriptions (ALDs) and their application to ALS. They then began the process of item review. On the second day, panelists completed the item review process. The item review process required more time per item than anticipated, leading to assignment of items at the end of day. On Day 2, two or three panelists at each table reviewed each item contained in the table group's OIL that had not been reviewed on Day 1. Following the completion of item review, facilitators led panelists through the process of developing borderline achievement level descriptions (BALDs) for each of the three levels. This activity also took longer than anticipated and added another source of disruption to the original schedule planned for Pilot 1. However, the process facilitator was able to guide the panelists through the remaining activities in a manner that ensured their understanding, but allowed for all activities to be completed by the end of day four.

Table 12: Summary of Schedule and Activities

Date	Activity
March 16	Take the NAEP TEL assessment Review performance on the NAEP TEL assessment Review the NAEP TEL Framework Review the Achievement Levels Descriptions Item Review
March 17	Item Review continued Item Review discussions Develop Borderline Achievement Levels Descriptions Instruction in Item Mapping Methodology
March 18	Review and discussion of Item Mapping procedures Round 1 bookmarking session Round 1 Feedback – Interpretation and Discussion Round 2 bookmarking session
March 19	Round 2 Feedback – Interpretation and Discussion Round 3 bookmarking session Round 3 Feedback – Interpretation and Discussion Consequences Questionnaire Identification of Exemplar Items

With regard to the ALS process itself, the panelists worked effectively with one another in both the table groups and together as a whole group. In general, panelists engaged in meaningful debate and were able to come to a working agreement on critical tasks. While consensus was not required for any activity, facilitators worked to ensure panelists reached general agreement and developed common understandings at each step of the process. Panelists were instructed to consider the discussions around each activity as a broader context within which they were to make their individual ratings.

Results and Recommendations

Following Pilot 1, and in collaboration with the Governing Board and the TEL TACSS, Pearson made the decision to develop and implement an alternative digital approach to support the NAEP TEL ALS. The extent to which the OPLS software would need to be modified to effectively carry out the standard setting activities was considered to be too great to be accomplished in the available time before the next panel meeting. Pearson instead created digital OILs in Excel, selected a web-based survey product to collect panelists' ratings and questionnaire responses, and

designed an alternative process for sharing feedback with panelists electronically. In addition, the content facilitator revised panelists' instruction and training on the TEL Framework, the TEL ALDs, and the creation of knowledge and skills statements to ensure all panelists had information about developing the knowledge and skills statements needed to place bookmarks.

Given the number of adjustments to the planned schedule that were needed throughout Pilot 1 and the extent of change to the ALS procedures, the TACSS recommended and the Governing Board staff approved inclusion of a second pilot study to evaluate the implementation of the new process components and tools as well as the adjusted schedule for Pilot 2. Based on the timing data collected throughout Pilot 1, all parties supported the addition of one day to both Pilot 2 and the Operational ALS study. The extended schedule allowed for longer breaks to be incorporated into the schedule planned for Pilot 2 and the Operational study. This change was in response to panelists' request for time to address outside responsibilities at multiple points throughout the day, which helped maintain their focus on their responsibilities during all ALS procedures. As previously noted, the honorarium was also increased (from \$300 to \$500) to account for the longer time commitment.

Pilot 2

This section of the report provides general results for Pilot 2 and describes only those features of Pilot 2 that differ from the procedures described under the section on general procedures. More specific information is given in this section about how those procedures were implemented.

The panel of teachers, educators, and the general public that had originally been selected for the operational ALS meeting were used for Pilot 2. Pearson contacted them to inform them that the meeting would be extended one additional day and asked if they would still be able to attend. Two of them indicated that they would not. Pearson then contacted people who had earlier volunteered and were qualified to serve as panelists but were not selected during the initial recruitment process in order to recruit replacement panelists with similar background characteristics. Dr. Rosenberg then communicated with each panelist through email (see Appendix K), explaining that the originally planned ALS study was converted to a second pilot study. No panelist expressed concerns in response to receiving this communication. The study was convened June 1-5, 2015. Pearson designed this pilot study to test the updated tools and processes planned for the ALS operational study and to evaluate the outcomes of that process.

Panelists

Table 13 summarizes Pilot 2 panel members' demographic information.

Table 13: Pilot 2 Panelist Demographics

Demographic Variable	Attributes	Percent
Panelist Type	Teachers	55%
	Non-Teacher Educators	14%
	General Public	31%
Gender	Female	55%
	Male	45%
Race/Ethnicity	White/Non-Hispanic	66%
	Black/Non-Hispanic	21%
	Native American/Mixed	3%
	Hispanic	10%
	Other	0%
NAEP Region	Midwest	21%
	Northeast	14%
	South	48%
	West	17%
Students with Disabilities	Experienced	72%
	Not experienced	28%
English Language Learners	Experienced	62%
	Not experienced	38%

As evidenced in Table 13, the panelists for Pilot 2 met the criteria for representation to a greater degree than the set of panelists who participated in Pilot 1. This was partially due to the larger size of the panel, twice that of Pilot 1, which made it easier to select a diverse set of qualified panelists. The panelists for Pilot 2 were organized into six table groups (two per OIL form) as depicted in Figure 10.

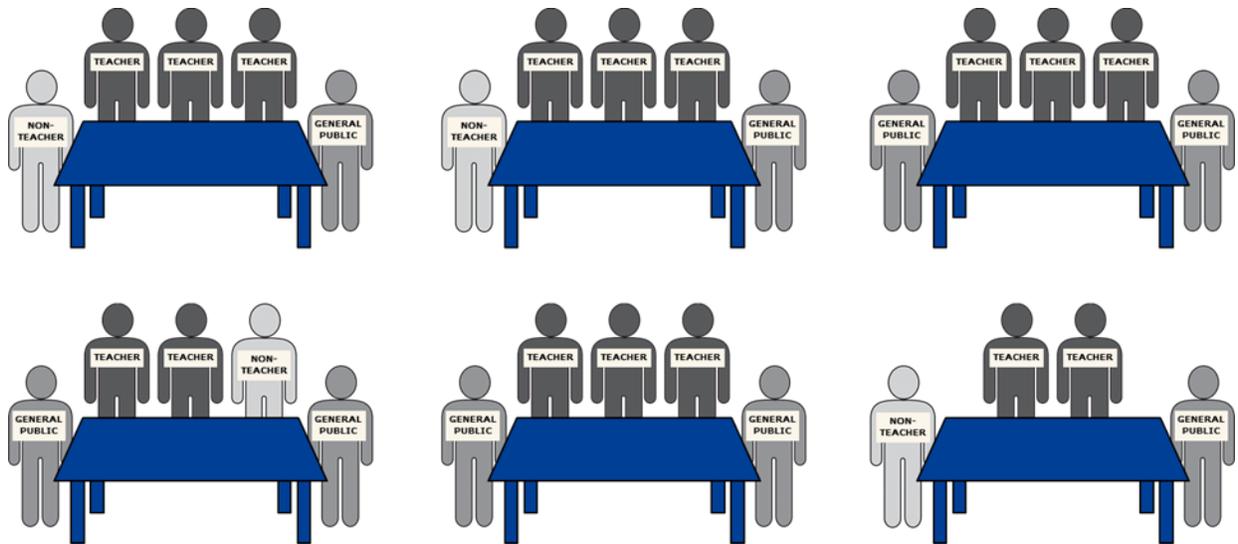


Figure 10: Composition of Table Groups for Pilot 2

On-Site Materials

The materials in panelists' folders for Pilot 2 included:

- Panelist Roster
- NAEP computer login instructions
- Overview booklet
- Agenda
- TEL ALDs by Sentences
- Terms and Definitions
- Table 3.1 from Framework
- Abridged Framework

Logistics

Room Configuration

In addition to the panelist tables, the room was configured (see Appendix M) such that the facilitators were positioned at the front of the room and all others (observers, Governing Board representatives, and Pearson team members) were positioned at tables in the back and sides of the room. The configuration for Pilot 2 differed from Pilot 1 to accommodate the change from three to six panelist table groups and also included larger tables.

Observers

Three observers were present for Pilot 2. They included two members of the TACSS, Dr. Barbara Dodd and Dr. Mary Pitoniak, and one staff member from the National Center for Education Statistics, Dr. Amy Yamashiro. Dr. Sharyn Rosenberg and Michelle Blair from the Governing Board were also present.

Procedures

The item mapping process previously described was successfully implemented for Pilot 2. The process was implemented using the software options recommended and designed by Pearson after Pilot 1 and approved by the TACSS and the Governing Board. Before the newly developed digital OILs were used in Pilot 2, they were independently reviewed and tested by Pearson staff not associated with the NAEP TEL project. A total of nine Pearson personnel provided a review of the three OILs. Each group of three reviewers consisted of a research associate, a research scientist, and a content expert. In addition, people within these job families who also had extensive Excel experience were especially encouraged to serve as reviewers. The OILs were divided among the reviewers who were given detailed written instructions about how to check each cell on each tab of the spreadsheet and to enter comments and ratings where appropriate and how to review all of the item images, rubrics, key annotations, and exemplars. Reviewers captured their comments and returned their completed OILs to the NAEP TEL team who then updated elements of the spreadsheet and supporting materials as necessary.

A summary of the day-by-day activities is provided in Table 14. The schedule shown in the table corresponds to that planned for Pilot 2. However, the creation of the initial BALDs, prior to Round 1 of the ALS procedures, took longer than scheduled. This time was made up through the remainder of the activities prior to and during Round 1, allowing the study to stay on track overall.

The activities summarized in Table 14 align with the descriptions of procedures provided previously with the following differences from Pilot 1:

- An example set of items was selected from the set of items common to all OILs for use during training of the panelists for the item review process. The item set included a variety of item types and incorporated both discrete items and items that were part of a scenario. The facilitators guided the panelists through the process of item review and the creation of knowledge and skills statements for each of the items in the example set. The guided practice included whole group discussion to help ensure a common understanding by all panelists. The example items were presented and discussed at the beginning of Day 2.
- Each item in a table group's OIL was assigned to three panelists in that table group for item review.
- During training for development of the BALDs, more focus was given to the BALDs as a tool for distinguishing specific key differences in the knowledge and skills required to transition from one achievement level to the next. In addition, the process facilitator included an example of a BALD for each achievement level that panelist used as a starting point for their work.

Table 14: Summary of Pilot 2 Schedule and Activities

Date	Activity
June 1	Take the NAEP TEL Assessment Review Performance on the NAEP TEL Assessment Review the NAEP TEL Framework Review the achievement levels descriptions Instructions for recording descriptions of Knowledge and Skills
June 2	Item Review Group discussion of K&S statements
June 3	Develop borderline achievement levels descriptions Instruction in Item Mapping Methodology Practice round of setting cut scores Round 1 bookmarking session
June 4	Round 1 Feedback – Interpretation and Discussion Round 2 bookmarking session Round 2 Feedback – Interpretation and Discussion Round 3 bookmarking session
June 5	Round 3 Feedback – Interpretation and Discussion Consequences Review Identification of Exemplar Items

As was true for Pilot 1, the panelists participating in Pilot 2 worked effectively with one another in both the table groups and together as a whole group. In general, panelists engaged in meaningful debate and were able to come to a working agreement on critical tasks. A number of problems arose related to full implementation of a technology-based process to support the ALS procedures. Each of these was addressed as it occurred and did not prevent implementation of the full process nor the completion of all tasks.

- On Day 1, issues arose with some of the NAEP computers. All of these computers had undergone extensive quality control checks. However, error messages appeared on the first day that neither Pearson staff nor NCES contractors had previously encountered. Replacement computers were available and substituted. Prior to the operational meeting, the anti-virus software was removed from the NAEP computers, as it was the suspected cause of the issue. This was an acceptable solution since at no time were the NAEP computers connected to the Internet. In addition, representatives from

Fulcrum IT and Westat were identified to attend the operational meeting and provide onsite support.

- One OIL had errors for a few items such that the link to an item image and/or rubric were incorrect. The panelists at the two tables using this OIL were dismissed while Pearson staff corrected the OIL. The panelists agreed to return at 7:30 a.m. the next day to make their Round 1 ratings. Panelists at the tables using the other two OILs were told to return at 9:45 a.m. the following day to resume the ALS activities.
- An issue with version control was discovered for a limited number of item images and scoring rubrics. The final versions of some of the item images and scoring rubrics inadvertently had not been transmitted to Pearson prior to Pilot 2. The updated versions were located in the NAEP item management system during the evening and inserted as replacements to the outdated versions. NCES and their contractors re-verified all materials prior to the operational study.

Cut Score Results

Table 15 presents the pseudo-scale score cuts after the Round 3 final ratings, before panelists made adjustments using the consequences review process.

Table 15: The Pilot 2 Panel Recommendations for NAEP TEL after Round 3

Level	Cut Score and Impact		
	Pseudo-Scale Score	Percent In Level	Percent at or Above
Basic	319	30%	81%
Proficient	351	46%	51%
Advanced	404	5%	5%

After viewing the consequences data from the Round 3 final ratings, panelists completed the consequences review. Table 16 summarizes the number and percentages of panelists who chose to make a change to the whole group Round 3 recommendations.

Table 16: Number and Percentages of Pilot 2 Panelists Who Changed Cut Score Recommendations during Consequences Questionnaire Activity

	Basic	Proficient	Advanced
Count	6	4	8
Percent	21%	14%	28%

Table 17 shows the pseudo-scale score cuts after the consequences review round. The group-level cut score (median across all panelists) was not changed for any

achievement level as a result of panelists’ recommendations in the Consequences Review.

Table 17: Pilot 2 Panel Recommendations for NAEP TEL Cut Scores

Level	Cut Score and Impact		
	Pseudo-Scale Score	Percent In Level	Percent at or Above
Basic	319	30%	81%
Proficient	351	46%	51%
Advanced	404	5%	5%

Identification of Exemplar Items

For the final step in the NAEP ALS process, panelists implemented the process previously described to identify exemplar items from the pool of items designated for potential release to the public.

Process Evaluation

Some survey questions were repeated across rounds. Panelists appeared to become more comfortable with the ALS process as the rounds progressed as seen in the increasing numbers of “strongly agree” responses to these repeated questions. See Table 18 for the number and percentages of responses by round to each of the relevant questions.

Table 18: Pilot 2 Panelist Changes to Repeat Survey Questions Across Rounds

	Strongly Agree / Agree		
	Round 1	Round 2	Round 3
I was comfortable using the two computers while placing my bookmarks.	24 (83%)	24 (83%)	22 (76%)
The instructions on how I was to select my bookmarks were clear.	25 (86%)	29 (100%)	29 (100%)
I had a good understanding of how to use the Borderline Achievement Level Descriptions to select my bookmarks.	25 (86%)	29 (100%)	29 (100%)
I understand the difference between borderline performance and typical performance within an achievement level.	28 (97%)	29 (100%)	29 (100%)
When choosing my bookmarks, I was comfortable taking into account how the Technology and Engineering Literacy principles and practices related to the achievement level.	23 (79%)	24 (83%)	27 (93%)
When choosing my bookmarks, I was comfortable thinking about and using the idea of a borderline performance within an achievement level.	26 (90%)	29 (100%)	29 (100%)
The Technology and Engineering Literacy principles and practices required by the items around my bookmarks are appropriate for the borderline of the corresponding achievement level.	22 (76%)	22 (76%)	27 (93%)
I was comfortable using the description of performance at the borderline of Proficient when I selected my Round 1/2/3 bookmarks.	26 (90%)	28 (97%)	29 (100%)
I was comfortable using the description of performance at the borderline of Basic when I selected my Round 1/2/3 bookmarks.	24 (83%)	28 (97%)	29 (100%)
I was comfortable using the description of performance at the borderline of Advanced when I selected my Round 1/2/3 bookmarks.	20 (69%)	20 (69%)	26 (90%)
I am confident in my Round 1/2/3 bookmark placements.	14 (48%)	28 (97%)	29 (100%)

Panelists completed a final process evaluation questionnaire to provide overall feedback on their achievement levels-setting experience. The following questions from that questionnaire are key to evaluating the success of the ALS procedure as they provide information regarding panelists' overall confidence in the results.

Table 19 provides the responses from the panelists to these questions. The number in each cell represents the number of panelists selecting each of the response categories.

- 4. This ALS process produced achievement levels that are defensible.
- 5. This ALS process produced reasonable achievement levels.
- 18. I would be willing to sign a statement (after reading it of course) recommending the use of the cut scores resulting from this ALS process. (Yes/No)

Table 19: Number of Pilot 2 Panelists in Each Response Category for Questions 4, 5, and 18 on the Final Process Evaluation

Question	Response N (%)					Total N
	Strongly Agree/ Yes	Agree	Neutral	Disagree	Strongly Disagree/ No	
Item 4	11 (37.9%)	18 (62.1%)	0	0	0	29
Item 5	12 (41.4%)	17 (58.6%)	0	0	0	29
Item 18	29 (100%)				0	29

Recommendations

The ALS procedures planned for Pilot 2 were implemented as planned and according to schedule, with one exception. More time was needed for development of the initial BALDs prior to Round 1. Based on the timing data, the process facilitator and project director recommended, and the TACSS agreed, that for the Operational ALS study, this activity should be introduced with a greater emphasis on the draft nature of the Round 1 BALDs. This would ensure the panelists focused on developing clear BALDs while understanding the BALD revisions prior to Round 2 would be informed by panelist experiences during Round 1.

The Operational Achievement Levels-Setting Study

Panelists

The addition of the second pilot study required Pearson to add a second cycle of panelist recruitment and selection for the Operation ALS study. The target number of panelists for the Operational meeting was 30. However, Pearson staff identified and selected 33 panelists to ensure that at least 30 would attend the meeting if some were unable to participate at the last minute. The nomination and recruitment process continued until 33 qualified and available panelists were identified.

Communication with the nominators and nominees was conducted through email and supplemented by telephone calls as needed to optimize the recruitment process. Shortly before the meeting, two panelists notified Pearson that they would not be able to attend, resulting in 31 panelists for the Operational ALS study.

Table 20 shows the demographic characteristics of the 31 panelists for the ALS meeting.

Table 20: ALS Panelist Demographics

Demographic Variable	Attributes	Percent
Panelist Type	Teachers	55%
	Non-Teacher Educators	16%
	General Public	29%
Gender	Female	68%
	Male	32%
Race/Ethnicity	White/Non-Hispanic	55%
	Black/Non-Hispanic	26%
	Hispanic	10%
	Asian	3%
	Other	6%
NAEP Region	Midwest	26%
	Northeast	3%
	South	55%
	West	16%
Students with Disabilities	Experienced	87%
	Not experienced	13%
English Language Learners	Experienced	73%
	Not experienced	27%

As evidenced in Table 20, the panelists for the Operational study met the overall criteria for representation of panelist type and had a higher percentage of

candidates in race/ethnicity categories other than “White/Non-Hispanic,” as well as a higher percentage of panelists indicating experience with students in special populations. The table also shows that the northeast region was not as well represented on this panel. The representation of panelists by type was given higher priority.

The panelists were organized into six table groups (two per OIL form) as depicted in Figure 11.

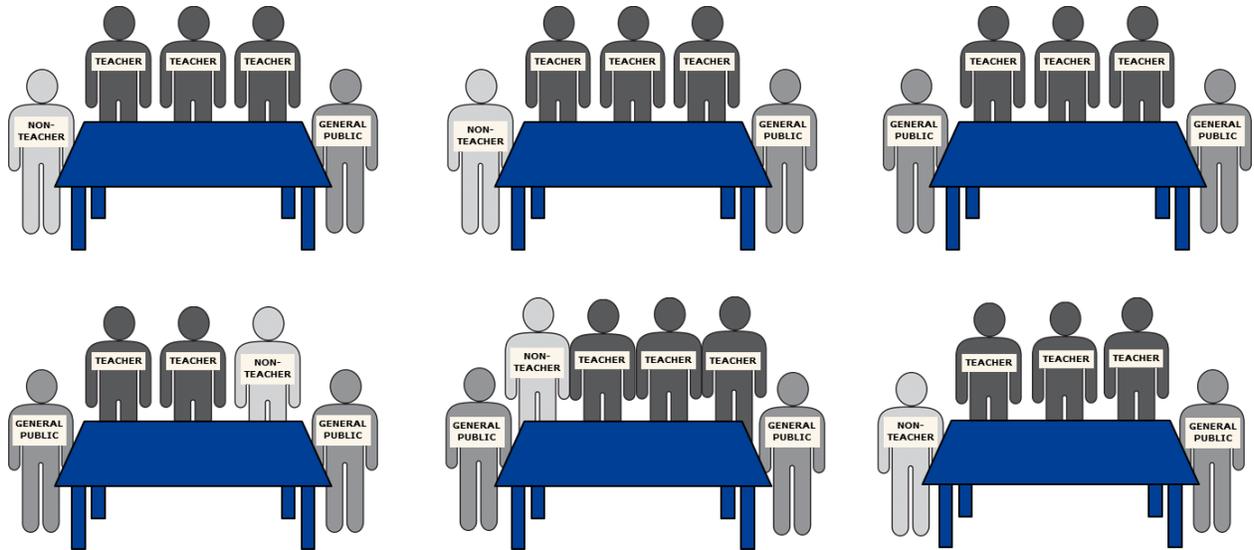


Figure 11: Composition of Table Groups for Operational ALS Study

Materials

Advance Materials

All panelists received a packet in the mail containing the same materials sent for the pilots. These materials were intended to provide panelists with a foundation for successfully carrying out the ALS process.

The letter sent with the materials underscored the importance of the TEL Framework and the ALDs to the process and urged panelists to become familiar with those two documents prior to the ALS meeting.

On-Site Materials

The materials in panelists’ folders for the Operational ALS study were the same as those used for Pilot 2.

Logistics

Facilitation

Dr. Nebelsick-Gullett and Dr. Moyer continued as the process and content facilitators for the operational study. They worked collaboratively throughout each stage of the process to support panelists' knowledge and application of the ALS procedures implemented.

Room Configuration

The room configuration was the same as that displayed in Appendix M and used for Pilot 2.

Observers

Four observers were present for the Operational ALS Study, including one member of the TACSS, Dr. Mary Pitoniak, and one staff member from the National Center for Education Statistics, Dr. Amy Yamashiro. To ensure all systems operated smoothly throughout the Operational ALS, James Carrion, IT Systems Manager for NAEP at Fulcrum IT, and Dan Weber, Systems Analyst at Westat, were present for the first three days of the ALS event. Dr. Sharyn Rosenberg from the Governing Board was also present.

Orientation to the Methodology

Figure 12 presents a high-level overview of the day-by-day activities for the meeting. A complete copy of the agenda can be found in Appendix N. All activities for the Operational ALS study occurred on schedule and as planned. Consequently, only a brief description is included for most of the sections of the Operational ALS study, with additional information added for clarification (e.g., examples of tools used for the Operational study) as well as extensions of and differences from what was previously described.

Date	Activity
September 28	Take the NAEP TEL assessment Review performance on the NAEP TEL assessment Review the NAEP TEL Framework Review the Achievement Levels Descriptions Instructions for recording descriptions of Knowledge and Skills (K&S)
September 29	Item Review to develop K&S statements Group discussion of K&S statements
September 30	Develop Borderline Achievement Levels Descriptions Instruction in Item Mapping Methodology Practice Round of setting cut scores Round 1 bookmarking session
October 1	Round 1 Feedback – Interpretation and Discussion Round 2 bookmarking session Round 2 Feedback – Interpretation and Discussion Round 3 bookmarking session
October 2	Round 3 Feedback – Interpretation and Discussion Consequences Review Identification of Exemplar Items

Figure 12: Summary of Schedule and Activities.

Training and Preparation

Each panelist received a folder during registration on the first day of the study event. The folder contained materials to support panelists’ work such as updated agendas, paper copies of login directions for various aspects of the ALS process, the abridged NAEP TEL Framework, and the TEL ALDs. Panelists were instructed to use these materials throughout the ALS processes.

Panelists received an orientation to the use of the dual-computer arrangement. They received information regarding the purpose and use of each computer as previously described and received support from the facilitators while exploring the functions of each computer.

Taking a NAEP TEL Assessment

After orientation and training, panelists used the NAEP computers to take a form of the 2014 NAEP TEL assessment. After all panelists had completed the TEL assessment, they used the Test Review worksheet in the Excel document that served as the OIL on the ALS computer to review screenshots of the items on the

form alongside the scoring criteria and rubrics for the items. The *Test Form Review* tab presented the core items in the NAEP sample test in the order they were encountered. Figure 13 shows a screenshot of the *Test Form Review* tab.

Item Order	Item ID	TEL Assessment Area	TEL Practice	Label	Max Score Code	MC Key or Rubric
1	VH007628	D&S	DS&AG	Iguana Home - Item 1	2	C
2	VH007631	D&S	DS&AG	Iguana Home - Item 2	3	Rubric
3	VH007635	D&S	DS&AG	Iguana Home - Item 3	2	D
4	VH007663	D&S	DS&AG	Iguana Home - Item 4	3	Rubric
5	VH007665	D&S	DS&AG	Iguana Home - Item 5	2	Rubric
6	VH007672	D&S	DS&AG	Iguana Home - Item 6	3	Rubric
7	VH007674	D&S	DS&AG	Iguana Home - Item 7	3	Rubric
8	VH007675	D&S	DS&AG	Iguana Home - Item 8	2	Rubric
9	VH007678	D&S	DS&AG	Iguana Home - Item 9	2	C
10	VH007680	D&S	DS&AG	Iguana Home - Item 10	2	Rubric
11	VH007572	ICT	C&C	Recreation Center - Item 1	2	C
12	VH007574	ICT	C&C	Recreation Center - Item 2 Cluster	3	Rubric
13	VH007575	ICT	C&C	Recreation Center - Item 3	2	D
14	VH007576	ICT	C&C	Recreation Center - Item 4	2	A
15	VH007577	ICT	C&C	Recreation Center - Item 5 Cluster	2	Rubric
16	VH007579	ICT	C&C	Recreation Center - Item 6	2	B
17	VH007580	ICT	C&C	Recreation Center - Item 7	4	Rubric
18	VF205106	T&S	DS&AG	Citizen Journalism Cluster Item	3	Rubric
19	VF420418	T&S	DS&AG	E-Books Cluster Item	3	Rubric
20	VF203596	T&S	DS&AG	Hybrid vs Gas Vehicle - Set Member 1 of 2	2	Rubric
21	VF203617	T&S	DS&AG	Hybrid vs Gas Vehicle - Set Member 2 of 2	2	Rubric

Figure 13: Segment of Test Form Review Spreadsheet.

The columns in the Test Form Review worksheet provided the following information:

- Column one, *Item Order*, provided the number associated with an item or score codes' order by difficulty.
- Column two, *Item ID*, provided a link to the item image that panelists could click to review a screenshot of the item.
- Columns three and four provided the codes for the associated TEL assessment and practice areas. The panelist could hover over the *TEL Assessment Area* or *TEL Practice* cell to see the full label associated with that item or score code.
- Column five, *Label*, provided the identifying label used in the OIL.

- Columns six and seven provided information on the scoring code associated with the item or score code on that row; and the correct response for a selected-response item or a link to the scoring guide/rubric the panelists reviewed to understand what is required of a student to answer an item correctly or fully.

Understanding the NAEP TEL Framework and the NAEP TEL ALDs

The content facilitator guided panelists through the process of developing a deeper understanding of the assessment framework for the NAEP TEL and the NAEP TEL ALDs. He engaged panelists in extensive discussion to ensure a common understanding of the framework and the ALDs with a focus on: (a) understanding the interaction between the three TEL content areas and the three TEL practices, (b) developing a foundation for evaluating the complexity and functioning of TEL assessment items, and (c) understanding what students should know and be able to do to perform at each TEL achievement level.

Describing the Knowledge and Skills for Responding to Items

The process and content facilitators worked collaboratively to provide panelists with training on how to develop knowledge and skills statements for items within the OIL. The training was completed at the end of Day 1 and included walking panelists through the process of accessing the information via the computers (see pages 14-16 of the NAEP TEL Facilitator Guide in Appendix A) and applying the review process using a subset of the common items. The process facilitator explained that, beginning the next day, each panelist would review an assigned subset of the items from the table group's OIL.

Item review was a time consuming and intensive process that grounded panelists in the nuances of the items and what they required of students. While this process is typically intense, the complexity of the TEL items and the need for panelists to review scenarios and discrete interactive items using the NAEP computer added to the typical workload for this activity. Constructed-response items on the TEL assessment were presented in a variety of different formats and required panelists to understand not only the impact of a given item structure but also the differences in knowledge and skills required to perform at varying levels on these items. Student performance on NAEP TEL constructed-response items was evaluated against criteria at up to five levels of performance, one for each credited response. Panelists used all of the time allotted on Day 2 to complete their individual item reviews and to share information with their table group to ensure all table group members had knowledge and skills information recorded for each item in their OIL.

Figure 14 shows the item review worksheet in the OIL Excel document located on panelists' ALS computers. This Excel file contained an *Item Review* tab that presented all items in a panelist's ordered item list (OIL) organized by item category (scenario-based item, discrete constructed-response item, discrete

selected-response item). All items in a scenario were grouped together and presented in the same order they occurred in the scenario.

View these items on the NAEP Computer in the scenario site													
Item ID	Item Image	NAEP Computer Label	Ordered Item List Label	TEL Assessment Area	TEL Practice	MC Key or Rubric	Knowledge and Skills Statements	P1	P2	P3	P4	P5	
VH007628	Image	Iguana Home	Iguana Home - Item 1	D&S	DS&AG	C							
VH007631	Image	Iguana Home	Iguana Home - Item 2	D&S	DS&AG	Rubric							
VH007635	Image	Iguana Home	Iguana Home - Item 3	D&S	DS&AG	D							
VH007663	Image	Iguana Home	Iguana Home - Item 4	D&S	DS&AG	Rubric							
VH007665	Image	Iguana Home	Iguana Home - Item 5	D&S	DS&AG	Rubric							
VH007672	Image	Iguana Home	Iguana Home - Item 6	D&S	DS&AG	Rubric							
VH007674	Image	Iguana Home	Iguana Home - Item 7	D&S	DS&AG	Rubric							
VH007675	Image	Iguana Home	Iguana Home - Item 8	D&S	DS&AG	Rubric							
VH007678	Image	Iguana Home	Iguana Home - Item 9	D&S	DS&AG	C							
VH007680	Image	Iguana Home	Iguana Home - Item 10	D&S	DS&AG	Rubric							

View these multi-score-point (polytomous) items by selecting the item image link												
Item ID	Item Image	Ordered Item List Label	TEL Assessment Area	TEL Practice	Item Rubric	Knowledge and Skills Statements	P1	P2	P3	P4	P5	
VF205106	Image	Citizen Journalism Cluster Item	T&S	DS&AG	Rubric							
VF420418	Image	E-Books Cluster Item	T&S	DS&AG	Rubric							

View these single-score-point items by selecting the item image link												
Item ID	Item Image	Ordered Item List Label	TEL Assessment Area	TEL Practice	MC Key or Rubric	Knowledge and Skills Statements	P1	P2	P3	P4	P5	
VF203596	Image	Hybrid vs Gas Vehicle - Set Member 1 of 2	T&S	DS&AG	Rubric							
VF203617	Image	Hybrid vs Gas Vehicle - Set Member 2 of 2	T&S	DS&AG	Rubric							

Figure 14: Segment of the Item Review Worksheet

The columns on the Item Review worksheet provided the following:

- Column one, *Item ID*, provided the item identification code associated with an item.
- Column two, *item image*, provided a pdf copy of the item.
- Column three, *NAEP Computer label*, provided the item label used by NAEP.
- Column four, *OIL Label*, provided the identifying label used in the OIL.
- Columns five and six provided the codes for the associated TEL assessment and practice areas. The panelist could hover over the *TEL Assessment Area* or *TEL Practice* cell to see the full label associated with that item or score code.
- Column seven, *MC Key or Rubric*, provided the correct response for a selected-response item or a link to the scoring guide/rubric the panelists reviewed to understand what is required of a student to answer an item correctly or fully.

- Column eight, *K&S Comments*, provided the space for recording the knowledge and skills statements developed during item review.
- Columns nine through thirteen provided the assignment information for each panelist. The panelists reviewed the items shaded in grey under their panelist number.

Developing Borderline Achievement Level Descriptions

The process and content facilitators guided panelists through the process of creating BALDs to delineate what students should know and be able to do at the lower border of each achievement level, providing a key resource for panelists to distinguish the lowest performance that qualifies a student for inclusion in a particular achievement category. Prior to developing the BALDs, the process facilitator provided training and information to ground the panelists in their understanding of what the BALDs represent and their importance to the standard setting process (see detailed description under the general procedures section of this report). For the Operational ALS study, she also provided an example of a BALD statement for each achievement level. The development of the BALDs provided a context for panelists to gain a common understanding of the borderline performances and to determine key differences in the knowledge and skills a student would need to transition between adjacent levels of achievement.

Using the ALS Methodology to Set Cut Scores

The process facilitator provided panelists with training and support to understand and apply the item mapping procedure for completion of three rounds of selecting and modifying cut score recommendations. As previously described, the item mapping was a judgmental decision-making process utilizing an OIL. Each of the three item sets included a subset of common items that were on the test form the panelists had seen on the first day. The facilitator also provided instruction regarding the use of a response probability level of 0.67 when matching item performance to the borderline ALDs in the ALS process.

During the ALS process, panelists placed bookmarks in their OIL to represent the cut score for each achievement level. In their application of the item mapping procedure, panelists utilized the Policy Definitions and the ALDs as well as the BALDs to inform their judgments. For each round of the ALS process, panelists independently identified a location in the OIL to select as the bookmark to represent the location used to compute the cut score for each of the three achievement levels.

Cut scores were identified one achievement level at a time beginning with Proficient, followed by Basic, and then Advanced. Facilitators instructed panelists to use the ALDs, the borderline ALDs, and their understanding of the response probability of 0.67 when determining which item to select for each cut point.

Appendix P. contains a segment of the OIL worksheet in the OIL Excel document located on panelists' ALS computers. This Excel file contained an *Ordered Item List* tab that presented all items/score codes in a panelist's ordered item list (OIL) in order of difficulty from least to most difficult. Figure 15 shows a sample worksheet.

Item Order	Location	Item ID	Zone/Bookmark			TEL Assessment Area	TEL Practice	Label	Score Code	Max Score Code	MC Key or Rubric	Sample Responses	Key Annotation	Viewed Item	K&S Comments Entered during Item Review	Additional Comments
			Basic	Prof.	Adv.											
1	257	VH007576				ICT	C&C	Recreation Center - Item 4	2	2	A					
2	300	VH007572				ICT	C&C	Recreation Center - Item 1	2	2	C					
3	302	VH007580				ICT	C&C	Recreation Center - Item 7	2	4	Rubric	Sample Responses	Key Annotation			
4	307	VH007635				D&S	DS&AG	Iguana Home - Item 3	2	2	D					
5	313	VH007631				D&S	DS&AG	Iguana Home - Item 2	2	3	Rubric	Sample Responses	Key Annotation			
6	315	VH007678				D&S	DS&AG	Iguana Home - Item 9	2	2	C					
7	319	VH007674				D&S	DS&AG	Iguana Home - Item 7	2	3	Rubric	Sample Responses	Key Annotation			
8	321	VH007574				ICT	C&C	Recreation Center - Item 2 Cluster	2	3	Rubric					
9	324	VH007663				D&S	DS&AG	Iguana Home - Item 4	2	3	Rubric	Sample Responses	Key Annotation			
10	331	VF205106				T&S	DS&AG	Citizen Journalism Cluster Item	2	3	Rubric					
11	336	VH007577				ICT	C&C	Recreation Center - Item 5 Cluster	2	2	Rubric					
12	337	VF420418				T&S	DS&AG	E-Books Cluster Item	2	3	Rubric					
13	340	VH007674				D&S	DS&AG	Iguana Home - Item 7	3	3	Rubric	Sample Responses	Key Annotation			
14	341	VH007575				ICT	C&C	Recreation Center - Item 3	2	2	D					
15	342	VH007579				ICT	C&C	Recreation Center - Item 6	2	2	B					
16	348	VH007665				D&S	DS&AG	Iguana Home - Item 5	2	2	Rubric	Sample Responses	Key Annotation			
17	348	VH007580				ICT	C&C	Recreation Center - Item 7	3	4	Rubric	Sample Responses	Key Annotation			
18	350	VH007628				D&S	DS&AG	Iguana Home - Item 1	2	2	C					
19	358	VH007672				D&S	DS&AG	Iguana Home - Item 6	2	3	Rubric	Sample Responses	Key Annotation			
20	360	VF205106				T&S	DS&AG	Citizen Journalism Cluster Item	3	3	Rubric					
21	371	VF203617				T&S	DS&AG	Hybrid vs Gas Vehicle - Set Member 2 of 2	2	2	Rubric	Sample Responses	Key Annotation			
22	373	VH007680				D&S	DS&AG	Iguana Home - Item 10	2	2	Rubric	Sample Responses	Key Annotation			
23	380	VH007580				ICT	C&C	Recreation Center - Item 7	4	4	Rubric	Sample Responses	Key Annotation			
24	391	VF203596				T&S	DS&AG	Hybrid vs Gas Vehicle - Set Member 1 of 2	2	2	Rubric	Sample Responses	Key Annotation			
25	393	VF420418				T&S	DS&AG	E-Books Cluster Item	3	3	Rubric					
26	412	VH007675				D&S	DS&AG	Iguana Home - Item 8	2	2	Rubric	Sample Responses	Key Annotation			
27	413	VH007663				D&S	DS&AG	Iguana Home - Item 4	3	3	Rubric	Sample Responses	Key Annotation			
28	431	VH007574				ICT	C&C	Recreation Center - Item 2 Cluster	3	3	Rubric					
29	462	VH007672				D&S	DS&AG	Iguana Home - Item 6	3	3	Rubric	Sample Responses	Key Annotation			
30	465	VH007631				D&S	DS&AG	Iguana Home - Item 2	3	3	Rubric	Sample Responses	Key Annotation			

Figure 15: Sample Ordered Item List

The columns on the OIL worksheet provided the following:

- Column one, *Item Order*, provided the number associated with an item or score codes' order by difficulty using RP 0.67.
- Column two, *Location*, provided the location of the item or score code on the NAEP pseudo-score scale.
- Column three, *Item ID*, provided a link to the item image that panelists could click to review a screenshot of the item.
- Columns four to six, *Zone/Bookmark*, provided one column for each bookmark decision at Basic, Proficient, and Advanced. A drop down menu was provided for each cell in a column to allow panelists' to enter their "Yes/No" decision by item or score code.
- Columns seven and eight provided the codes for the associated TEL assessment and practice areas. The panelist could hover over the *TEL Assessment Area* or *TEL Practice* cell to see the full label associated with that item or score code.
- Column nine, *Label*, provided the identifying label used in the OIL.
- Columns ten to twelve provided information on the scoring code associated with the item or score code on that row; the maximum score codes possible for the item associated with the row; and the correct response for a selected-response item or a link to the scoring guide/rubric the panelists reviewed to understand what is required of a student to answer an item correctly or fully.
- Columns thirteen and fourteen provided, for constructed-response items, additional examples of student responses at each score code.
- Column fifteen, *Viewed Item*, provided the panelist with a place to keep track of completed reviews.
- Column sixteen, *K&S Comments*, provided the knowledge and skills statements developed during item review.
- Column seventeen, *Additional Comments*, provided a place for panelists to add during the process of examining item information for setting bookmarks.

The facilitator directed panelists to work independently to determine their individual recommendations for each round. Between each round, facilitators asked panelists to discuss their understanding of the borderline descriptions and the knowledge and skills required at the border of each achievement level as well as the feedback that was provided from the previous round as discussed in the sections below devoted to each round of the ALS process.

Round 1: Understanding the Assessment and Student Achievement Overview of Round 1 Cut Score Placement

Facilitators directed the panelists to use the following process to apply the item mapping methodology to the items in their ordered item list:

- Log into the ALS computer, open the *Ordered Item List* Excel document, and click on the *Ordered Item List* tab.
- Begin the process with a focus on determining the cut score for the Proficient level.
- Review the BALDs, ALDs and policy definition for Proficient.
- Review each item, including the knowledge and skills statements and the scoring information for the item or score code.
- In the *Zone/Bookmark* column for Proficient, use the drop down menu associated within each cell to select if an item is “Yes” or “No” in answer to the question: Does borderline performance that is at the just barely Proficient level correspond to a 2/3 chance or better of a correct response to this item?
- Identify the zone within which you believe the bookmark falls.
- Use the dropdown menu in the cell you identify for your bookmark to mark the item as *Prof* for that item you have determined is your last confident “Yes.”
- Repeat this process to select bookmarks for Basic and Advanced.
- Record the item order number on the Round 1 handout in your folder for each cut score.
- Open the *Questionnaire* document on your ALS computer desktop.
 - Click on the *Round One Bookmark Selection* questionnaire.
 - Complete the questionnaire, including recording the placement of your bookmarks.

The first question asked the panelists to enter their panelist identification number, and the second question served as the bookmark selection form by asking the panelists to enter the item order number representing their bookmark for each achievement level. Before the panelists could continue, a facilitator verified the panelists’ entries against the item marked with a bookmark in the OIL for quality control. Figure 16 shows a screenshot of the online survey question used to collect the panelists’ bookmarks.

Round 1 Bookmark Selection and Questionnaire

2. Please enter the **Item Order Number** from the **Ordered Item List** that you select as the bookmark for each achievement level.

Basic

Proficient

Advanced

3. Please raise your hand so one of the facilitators can verify your entry.

Prev

Next

Figure 16: Bookmark Recording Form.

Round 2: Using Feedback

Understanding Feedback from Round 1

The process facilitator led panelists through the review and discussion of the Round 1 feedback (see Appendix Q to view Round 1 feedback). Panelists accessed the feedback by connecting to a secure site on their ALS computers and downloading the relevant documents. The facilitator then directed panelists to have a discussion with their table group that focused on: (a) examining the range of bookmarks selected for each achievement level, (b) examining the data showing the median values of the cut scores for panelists in the table group and for the whole group, (c) noting areas where bookmarks for one achievement level overlapped with another achievement level, and (d) discussing the bookmarks selected by table members from lowest to highest within each achievement level and the rationales for those judgments.

After completing the review of Round 1 feedback, panelists worked as a whole group to review, and as necessary revise, the BALDs prior to the placement of Round 2 bookmarks. Facilitators asked panelists to consider individually how well the BALDs worked when placing the Round 1 bookmarks and then led a whole group discussion of the BALDs focused on making revisions to improve their utility. The changes made focused on refining BALD statements to ensure there was separation between the expectations for achievement levels at the lower border of Basic, Proficient, and Advanced. The discussion focused on one or two aspects of the BALD for each achievement level with the exception of the BALD for the lower border of the Advanced level. Panelists had the most difficulty discerning this lower border, partly due to the limited number of items at the upper end of the score scale.

Overview of Round 2 Cut Score Placement

The process facilitator directed panelists to follow a process similar to Round 1, except that for Round 2 panelists used the revised BALDs with the feedback and discussions based on Round 1 to consider adjustments to their Round 1 cut points. Panelists started the Round 2 process with Proficient decision by deciding if they needed to extend or shift the zone they identified in Round 1 for that bookmark. Within the Round 2 zone, panelists followed the same procedure of considering each item within the zone and then bookmarking the last confident "Yes." The process was repeated for the Basic and Advanced cut points. Panelists then recorded their decisions using the same process as described for Round 1.

Round 3: Using Consequences Data

Understanding Feedback from Round 2

The process facilitator led panelists through the review and discussion of the Round 2 feedback (see Appendix Q to view Round 2 feedback). Panelists accessed the feedback by again connecting to the secure site on their ALS computers and downloading the relevant documents. She then directed panelists to have a discussion with their table group that focused on the same topics as occurred prior to Round 2.

The process facilitator then directed panelists to take out the paper copy of the Item Map provided after Round 2. She walked panelists through the process previously described that resulted in each panelist having an Item Map with lines designating the median cut points for the panelists' table group and for the whole group.

Panelists were then directed to open the file on the secure site that contained the consequences data chart. The process facilitator explained that this chart was created by applying the Round 2 group cut scores to the NAEP TEL data from 2014 and showed the percent of the scores that would have been placed in each achievement category using the whole group Round 2 recommended cut scores.

At this point, the facilitators led a whole group discussion of the information panelists gleaned from the item map activity and the consequences data. They asked panelists to consider if the consequences data seemed reasonable given their knowledge of (a) the ALDs, (b) the assessment items, and (c) instruction and performance in the TEL area.

After completing the review of Round 2 feedback, panelists worked as a whole group to review, and as necessary revise, the BALDs prior to the placement of Round 3 bookmarks. Facilitators again asked panelists to consider individually how well the BALDs worked when placing the Round 2 bookmarks and then led a whole group discussion of the BALDs focused on making revisions to improve their utility. The changes made at this point were minor and focused mainly at the BALD

focused on the lower border of Advanced. Panelists continued to have difficulty discerning this lower border; however, the discussion at this point allowed panelists to reach more of a common ground in their understanding of the borderline performance for the Advanced achievement level.

Overview of Round 3

The process facilitator directed panelists to follow a process similar to Round 2, except that for Round 3 panelists used the revised BALDs with the feedback and discussions based on Round 2 to consider adjustments to their Round 2 cut points.

Post-Round 3 Activities

Whole Group Feedback from Round 3

The facilitators opened the fifth and final day of the ALS study by providing panelists an opportunity to review and briefly discuss the feedback provided from the Round 3 standard setting activities. They followed a similar process to that described under Round 2 above but had access to feedback at only the whole group level, including an updated consequences data chart (see Appendix Q to view the Round 3 feedback).

Consequences Review

Panelists then received information and instructions for completing the consequences review described in the general procedures section of this document. This process allowed panelists to consider the group's recommendations from Round 3, provide individual feedback regarding the recommended cut scores, and use an interactive tool to understand the impact of adjusting the Round 3 cut score recommendations. After completing these activities, panelists completed a post-consequences questionnaire in which they either confirmed the final group-level recommendations or indicated alternate recommendations and a rationale for the suggested change(s). The process facilitator shared that this was another key piece of information shared with the Governing Board and that panelists' rationales were critical to understanding any changes made to the Round 3 whole group recommendations (see Appendix J for panelists reasons for recommending changes to cut scores).

Ratings of Exemplar Items

As a final activity for the ALS study, the facilitators provided panelists with instructions and support for the selection of exemplar items. The need for and role of exemplar items as well as the overall selection process were described in the general procedures section of this report. As previously noted, panelists at this point in the ALS process were very familiar with both the achievement level descriptions and the items in their OIL. Consequently, they were well prepared to make the judgments regarding items that may serve to represent the achievement levels.

As a first step in this process, panelists were directed to download the *Exemplar Identification List* (EIL) Excel document from the secure site used to provide feedback during the ALS procedure. Appendix P provides an example of this document. The worksheet contained only the items designated for release from a panelist's OIL. The items were classified within the achievement level they fit best based on the item's/score point's response probability as previously described. Select columns on the *Exemplars* worksheet are described below.

- Column three, *Item ID*, provided a link to the item image panelists clicked to review a screenshot of the item.
- Column seven, *MC Key or Rubric*, provided information on the correct response or the scoring guide/rubric the panelists reviewed to understand what is required of a student to answer an item correctly or fully.
- Columns eight and nine showed the knowledge and skills statements developed during item review and the additional comments made during the ALS rounds.
- Columns ten through fourteen provided:
 - the achievement level the item or score code belonged in with regard to the RP value (the first item in a set had an RP at the cut for borderline of that achievement level that was at or below RP 0.67 and an RP above 0.67 at the cut for achievement levels above the one being considered),
 - the average probability of a correct response within the designated achievement level, and
 - the RP values at the cut for each achievement level (columns labeled *Prob at Basic Cut*, *Prob at Prof Cut*, *Prob at Adv Cut*).
- The final column, *Exemplar Rating*, provided a drop down menu for the panelists to provide their individual rating of the item/score code as an exemplar. The ratings were "Should be Used," "Might be Used," and "Should not be Used."

To complete the exemplar selection process, panelists were instructed to examine the item-level data provided, review the comments they had entered for the item or score code, and consider the ALD for which the item or score code was designated. Panelists worked individually to use this information to decide whether this item or score code was a good representation of solid performance within the entire range of the achievement level. The process facilitator reminded panelists they were no longer focused on the borderline; rather, it was important they continually think in terms of the entire range for an achievement level. The panelists' ratings of the potential exemplar items are summarized in Appendix R.

After completing the exemplar selection process, panelists provided the last component of the process evaluation by accessing and completing the final process evaluation questionnaire (see Appendix J for results).

Outcomes of the Achievement Levels-Setting Process

Process Evaluation

The intent and purpose of the process evaluation component of the Operational ALS Study was described in the general procedures section of this report. Panelists responded to a number of evaluation questions multiple times at different points of the ALS process. Comparison of panelists' responses to these questions showed an increase in the number and percentage of panelists choosing "strongly agree/agree" to the majority of these questions over time, indicating they became more comfortable with the ALS process as the rounds progressed (Table 21; see Appendix J for complete questionnaire results). The facilitator asked if panelists had any questions about the process or the feedback they received after each round. Time was allowed for group discussion and clarification of any questions or uncertainty before the next round began.

Table 21: Pattern of Response to Repeated Evaluation Questions across Rounds

Question	Strongly Agree / Agree		
	Round 1	Round 2	Round 3
The instructions on how I was to select my bookmarks were clear.	25 (81%)	31 (100%)	31 (100%)
I had a good understanding of how to use the Borderline Achievement Level Descriptions to select my bookmarks.	23 (74%)	31 (100%)	31 (100%)
I understand the difference between borderline performance and typical performance within an achievement level.	29 (94%)	31 (100%)	31 (100%)
When choosing my bookmarks, I was comfortable taking into account how the Technology and Engineering Literacy principles and practices related to the achievement level.	21 (68%)	29 (94%)	29 (94%)
When choosing my bookmarks, I was comfortable thinking about and using the idea of borderline performance within an achievement level.	22 (71%)	31 (100%)	31 (100%)
The Technology and Engineering Literacy principles and practices required by the items around my bookmarks are appropriate for the borderline of the corresponding achievement level.	19 (61%)	29 (94%)	29 (94%)
I was comfortable using the description of performance at the borderline of Proficient when I selected my Round 1/2/3 bookmarks.	21 (68%)	31 (100%)	31 (100%)

Question	Strongly Agree / Agree		
	Round 1	Round 2	Round 3
I was comfortable using the description of performance at the borderline of Basic when I selected my Round 1/2/3 bookmarks.	24 (77%)	31 (100%)	31 (100%)
I was comfortable using the description of performance at the borderline of Advanced when I selected my Round 1/2/3 bookmarks.	22 (71%)	28 (90%)	29 (94%)
I am confident in my Round 1/2/3 bookmark selections.	13 (42%)	28 (90%)	30 (97%)

The process facilitator asked panelists to complete a final process evaluation to elicit panelist feedback on their overall ALS experience. The results of the evaluation indicated that almost all panelists viewed the cut scores recommended by this group as defensible, reasonable, and applicable (Table 22), with only two panelists indicating they would not sign a statement recommending the use of the results.

Table 22: Number of Panelists in Each Response Category for Questions 4, 5, and 18 on the Final Process Evaluation

Evaluation Question	Response N (%)				
	Strongly Agree/ Yes	Agree	Neutral	Disagree	Strongly Disagree/ No
This ALS process produced achievement levels that are defensible.	20 (65%)	10 (32%)	1 (3%)	0 (0%)	0 (0%)
This ALS process produced reasonable achievement levels.	20 (65%)	10 (32%)	1 (3%)	0 (0%)	0 (0%)
I would be willing to sign a statement (after reading it of course) recommending the use of the cut scores resulting from this ALS process.	29 (94%)				2 (6%)

The remainder of this section on process evaluation focuses on providing information relative to key aspects of the evaluation of the ALS implementation and results. Results are reported as the average (mean) response to questions on items from surveys, administered throughout the study, which focused on a particular key aspect of the process evaluation. Support for procedural validity is indicated by the

consistent average responses at or about 4.0 on the 1-5 scale, particularly for the average response after Round 2 and Round 3, to questions repeated across rounds.

Clarity of Instructions and Presentations

Table 23 shows the average ratings for questions focused on panelists’ perceptions of the clarity of instructions and/or presentations. Three items had an average rating below 4.0 and all three were at or above 3.9. Two of the questions were answered at the end of the practice round and focused on the clarity of the training on the ALS method and the clarity of the explanation of the OIL. Both of these items had an average response rating 3.97. The third question had an average rating of 3.90 and focused on clarity of instructions for selecting bookmarks, this question was repeated across rounds, and the average response was higher with each iteration.

Table 23: Average Ratings of Clarity of Instructions and Presentations

The explanation/presentation/instructions/training/description of ... was clear (5=strongly agree, 4=agree, 3=neutral, 2=disagree, 1=strongly disagree)

	Question #	Orientation Topic	ALS Average Rating
End of Day 1	4	The overall NAEP program	4.35
End of Day 1	5	The purpose of the NAEP ALS meeting	4.16
End of Day 1	8	The TEL Framework	4.19
End of Day 1	16	How to use the NAEP computer to review items	4.55
End of Day 1	17	How to use the ALS computer to review items and record comments	4.42
End of Day 1	18	How to use the ALS and NAEP computers together	4.45
End of Day 1	21	How to identify the knowledge and skills necessary to answer an item	4.13
End of Day 1	22	Tasks to perform in the Item Review	4.13
End of Day 1	23	The score levels for polytomous items	4.06
Practice Round	3	Developing Borderline Achievement Level Descriptions	4.03
Practice Round	11	The method for setting achievement levels	3.97
Practice Round	12	The explanation of the information in the OIL	3.97
Post Round 1	4	How to select and record bookmarks	3.90

	Question #	Orientation Topic	ALS Average Rating
Pre Round 2	6	Task for selecting bookmarks in Round 2	4.42
Pre Round 3	10	Task for selecting bookmarks in Round 3	4.65
Post Round 2	4	How to select bookmarks	4.58
Post Round 3	4	How to select bookmarks	4.77
Pre Round 3	6	Item Map distributed with the Round 2 feedback	4.32
Pre Round 3	9	Using consequences data during Round 3	4.00
Pre-Consequences	7	Task for the consequences questionnaire	4.39
Final	15	The Exemplar Item Rating Task	4.39

Usefulness or Helpfulness of Activities and Information

Table 24 shows the average ratings for questions focused on panelists' perceptions of the usefulness or helpfulness of the activities and information included in the ALS procedures. Two items had an average rating below 4.0. One of these questions, with an average rating of 3.94, focused on the usefulness of the BALDs in placing bookmarks during the practice round, which did not incorporate a full application of these descriptors. The other question had an average rating of 3.19 and focused on whether panelists took into account the consequences data from Round 2 when selecting Round 3 bookmarks. The discussion surrounding the consequences data and the average responses to question 2 on the first part of the consequences questionnaire and question 12 on the final evaluation questionnaire suggests that while panelists understood the data and found it helpful during the ALS process, it did not necessarily influence their placement of the Round 3 bookmarks.

Table 24: Average Ratings of Usefulness or Helpfulness of Activities and Information

(5=strongly agree, 4=agree, 3=neutral, 2=disagree, 1=strongly disagree)

	Question #	Activity or Information	ALS Average Rating
Practice Round	14	The Borderline Achievement Level Descriptions will be useful in placing my bookmarks	3.94
Pre Round 3	7	The Item Map will be helpful in selecting my Round 3 bookmarks	4.03
Post Round 3	13	The Item Map showing the locations of the items by assessment area was helpful in selecting my bookmarks	4.00
Post Round 3	14	I took into account the consequences data from Round 2 when I selected my bookmarks	3.19
Final	7	The Achievement Level Descriptions were helpful during the ALS process	4.65
Final	8	The development of Borderline Achievement Level Descriptions was helpful during the ALS process	4.58
Final	9	The Ordered Item List was helpful during the ALS process	4.77
Final	10	The Item Map was helpful during the ALS process	4.42
Final	11	The Rater Location Charts were helpful during the ALS process	4.26
Final	12	The Consequences Data were helpful during the ALS process	4.10
Final	16	The exemplar items will be useful to describe the achievement levels	4.42

Amount of Time Allocated for Tasks

Table 25 shows the average ratings for questions focused on panelists' perceptions of the appropriateness of the time allocated to specific portion of the ALS process. The average ratings are based on a 3-point scale (4, 3, 2). An average rating of 3 indicated that the time allocated was, on average, about right. All but one of the average ratings were between 2.81 and 3.19.

Table 25: Average Ratings of Amount of Time Allocated for Tasks

The amount of time spent on/allotted to ... was (4=too long, 3=about right, 2=too short)

	Question #	Task	ALS Average Rating
End of Day 1	3	The General Orientation to the NAEP Program	3.19
End of Day 1	6	The General Introduction to the NAEP Achievement Level Setting process	3.06
End of Day 1	20	Orientation to the two computers	3.00
End of Day 2	2	Complete the Item Review	2.81
Practice Round	4	Developing Borderline Achievement Level Descriptions	3.06
Practice Round	10	Training for the Item Mapping method	2.87
Post Round 1	5	Select bookmarks during Round 1	2.81
Post Round 2	5	Select bookmarks during Round 2	2.94
Post Round 3	5	Select bookmarks during Round 3	3.10
Final	13	Complete the Consequences Questionnaire	3.00
Final	14	Complete the Exemplar Item Rating Task	2.97

Understanding of Concepts and Feedback

Table 26 shows the average ratings for questions addressing panelist understanding of concepts and panelist understanding of feedback. Two items had an average rating below 4.0. One of the questions focused on the appropriateness of the TEL principles and practices required by the items around the corresponding achievement level. This question was repeated across rounds and the average response increased from 3.65 after Round 1 to 4.39 after Round 3, corresponding to response data indicating panelists' confidence in their ratings and understanding of the process were higher across Round 2 and Round 3. The other item focused on panelist understanding of the Rater Location Chart, provided as feedback after Round 1 and Round 2. Prior to Round 2, the average rating for panelist understanding of this piece of feedback was 3.42. The average rose to 4.32 prior to Round 3, after the process facilitator made it clear which chart was associated with that title.

Table 26: Average Ratings of Understanding of Concepts and Feedback

I have a good understanding of ... (5=strongly agree, 4=agree, 3=neutral, 2=disagree, 1=strongly disagree)

	Question #	Concept or Feedback Area	ALS Average Rating
End of Day 1	10	Technology and engineering literacy as defined by the NAEP Framework	4.48
End of Day 1	15	Taking the NAEP TEL gave me a good idea of what is expected of students	4.74
End of Day 1	24	How to match items in the NAEP computer with their information in the ALS computer	4.39
End of Day 1	25	How to distinguish items with one correct answer from polytomous items	4.32
End of Day 1	26	How to locate the score categories of polytomous items in the ALS computer	4.35
End of Day 2	4	How to navigate through the scenarios on the two computers	4.65
End of Day 2	5	How to identify items within the scenarios on the two computers	4.58
Post Round 1	10	The Technology and Engineering Literacy principles and practices required by the items around my bookmarks are appropriate for the borderline of the corresponding achievement level	3.65
Post Round 2	10	The Technology and Engineering Literacy principles and practices required by the items around my bookmarks are appropriate for the borderline of the corresponding achievement level	4.16
Post Round 3	11	The Technology and Engineering Literacy principles and practices required by the items around my bookmarks are appropriate for the borderline of the corresponding achievement level	4.39
Pre Round 2	2	How my Round 1 cut scores were derived from my bookmarks	4.42
Pre Round 3	2	How my Round 2 cut scores were derived from my bookmarks	4.58

	Question #	Concept or Feedback Area	ALS Average Rating
Pre Round 2	3	How the Round 1 median group cut scores were derived from the panelists' bookmarks	4.35
Pre Round 3	3	How the Round 2 median group cut scores were derived from the panelists' bookmarks	4.61
Pre Round 2	4	The meaning of the Rater Location Feedback	3.42
Pre Round 3	4	The meaning of the Rater Location Feedback	4.32
Pre Round 2	5	What students at the Round 1 median cut scores should know and be able to do	4.19
Pre Round 3	5	What students at the Round 2 median cut scores should know and be able to do	4.52
Pre Round 3	8	The meaning of the consequences data.	4.16
Final	2	The purpose of this meeting	4.97

Understanding the ALDs and Borderline Performance

Table 27 shows the average ratings for questions focused on panelist understanding of ALDs and borderline performance. One item had an average rating below 4.0 after Round 1. This item, which focused on the use of BALDs to select bookmarks, was a repeated item. The average responses after Round 2 and Round 3 were 4.39 and 4.71, respectively, again, indicating an increase in panelist understanding and application of the process across rounds.

Table 27: Average Ratings of Understanding the ALDs and Borderline Performance

I have a good understanding of ... (5=strongly agree, 4=agree, 3=neutral, 2=disagree, 1=strongly disagree)

	Question #	ALDs or Borderline Performance Level	ALS Average Rating
End of Day 1	11	The ALDs are clear descriptions about what students should know and be able to do at each achievement level	4.00
End of Day 1	12	The Basic achievement level description	4.03
End of Day 1	13	The Proficient achievement level description	4.10
End of Day 1	14	The Advanced achievement level description	4.03
Practice Round	6	The difference between borderline performance and typical performance within an achievement level	4.26
Post Round 1	7	The difference between borderline performance and typical performance within an achievement level	4.16
Post Round 2	8	The difference between borderline performance and typical performance within an achievement level	4.45
Post Round 3	8	The difference between borderline performance and typical performance within an achievement level	4.68
Post Round 1	6	How to use the Borderline Achievement Level Descriptions to select bookmarks	3.77
Post Round 2	7	How to use the Borderline Achievement Level Descriptions to select bookmarks	4.39
Post Round 3	7	How to use the Borderline Achievement Level Descriptions to select bookmarks	4.71

Comfort with processes and procedures

Table 28 shows the average ratings for questions addressing panelist comfort level with specified processes and procedures. The majority of items after the practice

round and Round 1 had an average response between 3.7 and 4.0, with most of the items focused on panelist comfort with the ALDs and their use. Level of comfort increased across Round 2 and Round 3, with all average response values at or above 4.48 by Round 3. One repeated item, which indicated feeling pressure to select bookmarks close to those of other panelists, had an average rating of 2.35 after Round 2 and 2.55 after Round 3. The lower average response indicated that, in general, panelists did not agree with the statement.

Table 28: Average Ratings of Comfort with Processes and Procedures

I am/was comfortable ... 5=strongly agree, 4=agree, 3=neutral, 2=disagree, 1=strongly disagree)

	Question #	Process or Procedure	ALS Average Rating
End of Day 1	19	Using the two computers together	4.65
End of Day 2	3	Working through the Item Review tab in the Excel file during the Item Review	4.35
Practice Round	5	Using the Achievement Level Descriptions to develop the idea of borderline performance	3.81
Practice Round	7	With the description of performance at the borderline of Basic	4.03
Practice Round	8	With the description of performance at the borderline of Proficient	3.90
Practice Round	9	With the description of performance at the borderline of Advanced	3.94
Practice Round	13	The Practice Round to select my bookmarks helped me feel comfortable with the process	3.90
Post Round 1	8	Taking into account how the Technology and Engineering Literacy principles and practices related to the achievement level	3.68
Post Round 2	9	Taking into account how the Technology and Engineering Literacy principles and practices related to the achievement level	4.23
Post Round 3	9	Taking into account how the Technology and Engineering Literacy principles and practices related to the achievement level	4.48

	Question #	Process or Procedure	ALS Average Rating
Post Round 1	9	Thinking about and using the idea of borderline performance within an achievement level	3.81
Post Round 2	10	Thinking about and using the idea of borderline performance within an achievement level	4.39
Post Round 3	10	Thinking about and using the idea of borderline performance within an achievement level	4.65
Post Round 1	11	Thinking about and using the idea that 2/3 of students with borderline performance would get the item correct or receive a score in the specified score category for a polytomous item	4.00
Post Round 2	12	Thinking about and using the idea that 2/3 of students with borderline performance would get the item correct or receive a score in the specified score category for a polytomous item	4.35
Post Round 3	12	Thinking about and using the idea that 2/3 of students with borderline performance would get the item correct or receive a score in the specified score category for a polytomous item	4.55
Post Round 1	12	Using the description of performance at the borderline of Proficient when I selected Round 1 bookmarks	3.74
Post Round 2	13	Using the description of performance at the borderline of Proficient when I selected Round 2 bookmarks	4.45
Post Round 3	15	Using the description of performance at the borderline of Proficient when I selected Round 3 bookmarks	4.61
Post Round 1	13	Using the description of performance at the borderline of Basic when I selected Round 1 bookmarks	3.84

	Question #	Process or Procedure	ALS Average Rating
Post Round 2	14	Using the description of performance at the borderline of Basic when I selected Round 2 bookmarks	4.39
Post Round 3	16	Using the description of performance at the borderline of Basic when I selected Round 3 bookmarks	4.58
Post Round 1	14	Using the description of performance at the borderline of Advanced when I selected Round 1 bookmarks	3.71
Post Round 2	15	Using the description of performance at the borderline of Advanced when I selected Round 2 bookmarks	4.23
Post Round 3	17	Using the description of performance at the borderline of Advanced when I selected Round 3 bookmarks	4.55
Post Round 2	6	I felt pressure to select my bookmarks close to those of other panelists	2.35
Post Round 3	6	I felt pressure to select my bookmarks close to those of other panelists	2.55

Reactions to Consequences Data

Table 29 shows the average ratings for questions focused on panelists' understanding of the consequences data provided at the end of the ALS process, based on Round 3 recommended cut scores. The data for these questions indicate that while most panelists understood the data and indicated agreement that the proportions reflected their expectations, the proportion marking "Yes" was notably lower regarding the proportion of students who would be at or above the Advanced cut score. In addition, nearly 40% of panelists (see Table 35) recommend a change to the Round 3 cut score recommended for the Advanced achievement level.

Table 29: Average Ratings of Reactions to Consequences Data

(5=strongly agree, 4=agree, 3=neutral, 2=disagree, 1=strongly disagree / 0=No, 1=Yes)

	Question #	Consequences Data Element	ALS Average Rating
Pre-Consequences	2	I understand the meaning of the consequences data from Round 3	4.55
Pre-Consequences	3	Does this percentage reflect your expectation about the proportion of students whose NAEP score would be at or above the Proficient cut score? (Yes/No)	0.87
Pre-Consequences	4	Does this percentage reflect your expectation about the proportion of students whose NAEP score would be at or above the Basic cut score? (Yes/No)	0.81
Pre-Consequences	5	Does this percentage reflect your expectation about the proportion of students whose NAEP score would be at or above the Advanced cut score? (Yes/No)	0.65
Pre-Consequences	6	Having seen the data on the percentages of students whose score on the NAEP was at or above the cut score your panel set for each achievement level, would you change the Round 3 cut score for one or more of the achievement levels if you could? (Yes/No)	0.35

Confidence with decisions and outcomes

Table 30 shows the average ratings for questions addressing panelist confidence in their bookmark selections, by round, and the outcomes of the ALS process. With the exception of a lower confidence in bookmark selections after Round 1, all average response values were above 4.25 for items rated on the five-point scale. Question number 18 asked panelists to indicate if they would be willing to sign a statement recommending the use of the cut scores resulting from the ALS process. The reported average is 0.94, which indicated that 94% of the panelists responded "Yes." As shown in Table 22, 29 panelists marked "Yes" to this item and two panelists marked "No." The two panelists who marked "No" did not provide a specific reason for doing so in their comments; however, comments by both in response to the consequences questionnaire indicated that they thought the Advanced cut score was too high. One of the two wrote that he/she "may be" willing to sign a statement.

Table 30: Average Ratings of Confidence with Decisions/Outcomes

(5=strongly agree, 4=agree, 3=neutral, 2=disagree, 1=strongly disagree /
0=No, 1=Yes)

	Question #	Decision or Outcome	ALS Average Rating
Post Round 1	15	I am confident in my Round 1 bookmark selections	3.32
Post Round 2	16	I am confident in my Round 2 bookmark selections	4.26
Post Round 3	18	I am confident in my Round 3 bookmark selections	4.68
Final	3	This ALS process provided me an opportunity to use my best judgment to recommend cut scores for the NAEP TEL assessment	4.84
Final	4	This ALS process produced achievement levels that are defensible	4.61
Final	5	This ALS process produced reasonable achievement levels	4.61
Final	6	The achievement levels capture meaningful distinctions in TEL performance as described in the ALDs	4.61
Final	17	The exemplar items I reviewed are appropriately matched to their achievement level	4.29
Final	18	I would be willing to sign a statement (after reading it of course) recommending the use of the cut scores resulting from this ALS process (Yes/No)	0.94

Summary of Cut Scores

Table 31 shows the median and mean cut scores and standard deviations of the total group cut scores by round. The standard deviation is a measure of the variability of the panelists' cut scores, and it is based on the mean. The table shows that the variability of the cut scores for each achievement level is highest at Round 1 and generally decreases over the rounds as panelists become more familiar and comfortable with the process and develop a better understanding of the concept of borderline performance.

Table 31: ALS Panel Median Cut Scores and Standard Deviations by Round

	Scale Scores ¹¹ and Standard Deviations (S.D.)								
	Basic			Proficient			Advanced		
Round	Median	Mean	S.D.	Median	Mean	S.D.	Median	Mean	S.D.
1	116	119	19.4	150	155	30.0	193	204	40.2
2	116	120	12.9	151	156	17.3	205	207	27.6
3	116	121	12.2	151	158	18.3	209	208	14.8

¹¹ Throughout the project, a pseudo-NAEP scale was used to avoid the risk of having the NAEP achievement level cut scores released before intended. As in past ALS studies, the pseudo-NAEP scale was a linear transformation of the NAEP reporting scale. The results in this section have been transformed back to the actual NAEP reporting scale.

Figure 17 shows the mean absolute difference of the panelists' cut scores from the whole group median by round.

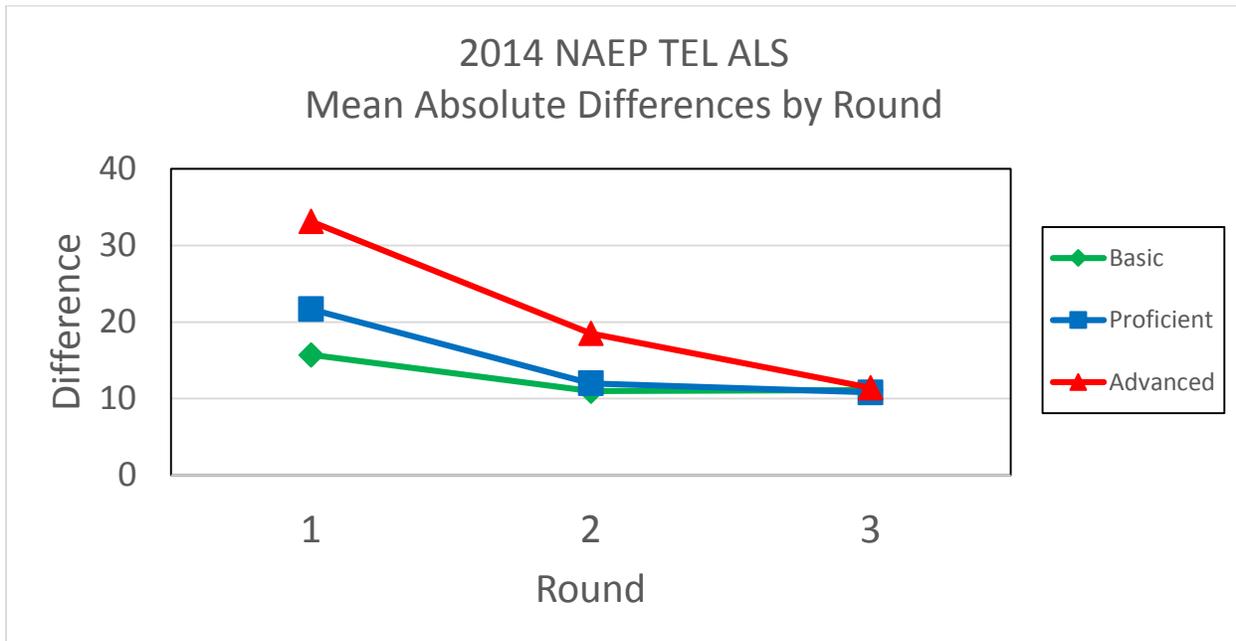


Figure 17: Mean Absolute Difference in Panelist Scores by Round

The data in Figure 17 show that the greatest discrepancy among the panelist cut scores occurred during Round 1, and especially at the advanced level. By Round 3, the deviation among the cut scores for all three levels was reduced and very similar across levels.

Table 32 shows the median cut scores for each achievement level and round disaggregated by OIL and table group. During the meeting, table groups 1 and 4 worked with OIL 1, table groups 2 and 5 worked with OIL 2, and table groups 3 and 6 worked with OIL 3.

Table 32: NAEP TEL ALS Median Cut Scores by OIL and Table Group

Group		Basic			Proficient			Advanced		
		R1	R2	R3	R1	R2	R3	R1	R2	R3
Total		116	116	116	150	151	151	193	205	209
OIL	1	108	108	110	142	143	147	212	205	212
	2	121	119	118	148	151	151	180	187	193
	3	126	133	134	158	176	165	220	218	218
Table	1	110	110	110	143	143	146	181	199	204
	2	123	133	133	136	150	152	178	181	185
	3	124	131	132	172	185	197	220	220	220
	4	98	104	106	137	150	149	222	213	213
	5	116	114	114	151	151	151	181	192	194
	6	138	138	138	152	153	153	211	211	211

Table 33 shows the number and percent of panelists who changed their bookmarks between rounds and whether that resulted in a change to their cut score between rounds.

Table 33: Number and Percent of Panelists Who Changed their Cut Scores between Rounds in the ALS

Achievement Level	Changes Between Rounds					
	1 to 2			2 to 3		
	Increase N (%)	No Change N (%)	Decrease N (%)	Increase N (%)	No Change N (%)	Decrease N (%)
Basic	15 (48)	6 (19)	10 (32)	8 (26)	18 (58)	5 (16)
Proficient	17 (55)	5 (16)	9 (29)	12 (39)	12 (39)	7 (23)
Advanced	18 (58)	2 (6)	11 (35)	13 (42)	13 (42)	5 (16)

Table 33 shows that most panelists made changes to their Round 1 bookmark placements in Round 2, and most of those changes were to increase the cut scores. By Round 3, most people made no change to their bookmark placement for the Basic cut score. The number of panelists that made no changes in Round 3 to their Round 2 bookmark placements is exactly equal to the number that changed their bookmark placements to increase the cut scores for both the Proficient and Advanced levels. Few panelists decreased their cut scores for any achievement level between Rounds 2 and 3.

Table 34 reports the scale score cut scores and the percent of students in and at or above each achievement level after the Round 3 rating process.

Table 34: Panel Recommendations for NAEP TEL Achievement Levels after Round 3

Level	Cut Scores and Percentages		
	Scale Score	Percent In Level	Percent at or Above
Basic	116	32%	83%
Proficient	151	48%	51%
Advanced	209	3%	3%

Table 35 reports the number and percentages of panelists who chose to modify the Round 3 group cut score recommendations after viewing the consequences data from the Round 3 final ratings. The first row shows the number and percent of panelists who did not recommend a change from the Round 3 cut scores. The second and third rows show the number and percent who suggested a decrease or increase, respectively, and the average change in the scale score cut resulting from the recommendation. When panelists proposed a change, they were asked to record their reason(s) in the post-consequences questionnaire. A record of all responses can be found in Appendix J. In general, the majority of the panelists provided rationales based on the content of items they judged to represent performance right at the lower border of the achievement level.

Table 35: Number and Percent of ALS Panelists Who Changed Cut Score Recommendations during Consequences Review

	Basic	Proficient	Advanced
N (%) Did Not Change	23 (74%)	23 (74%)	19 (61%)
N (%) Lowered	4 (13%)	4 (13%)	9 (29%)
N (%) Raised	4 (13%)	4 (13%)	3 (10%)
N (%) Who Changed Recommendation to Match Their Round 3 Cut Score	1 (3%)	6 (19%)	3 (10%)

The majority of panelists recommended no changes to cut scores. For those who would make changes, the largest percentage recommended lowering the Advanced cut score. The individual changes panelists suggested be made to the Round 3

whole group cut scores during the consequences review did not result in any changes to the Round 3 cut scores.

Reliability of Cut Scores

The reliability of cut scores resulting from a standard setting process is typically thought of with regard to how consistently the cut scores would be reproduced if the achievement levels-setting process were repeated with a different sample of panelists. Pilot 2 and the Operational ALS meetings represent such a replication reasonably well. The results from the two meetings were highly similar and are compared at the end of this section. But first, the cut scores from the ALS meeting are presented for different subgroups of panelists to show how the results vary depending on the OIL with which the panelists worked (Figure 18), as well as their table group (Figure 18). Differences in results at these levels are expected to be larger than would be expected for a full replication of the meeting because of the much smaller group sizes.

Figure 18 shows a plot of the cut scores by round for the panelists using each item rating set (OIL).

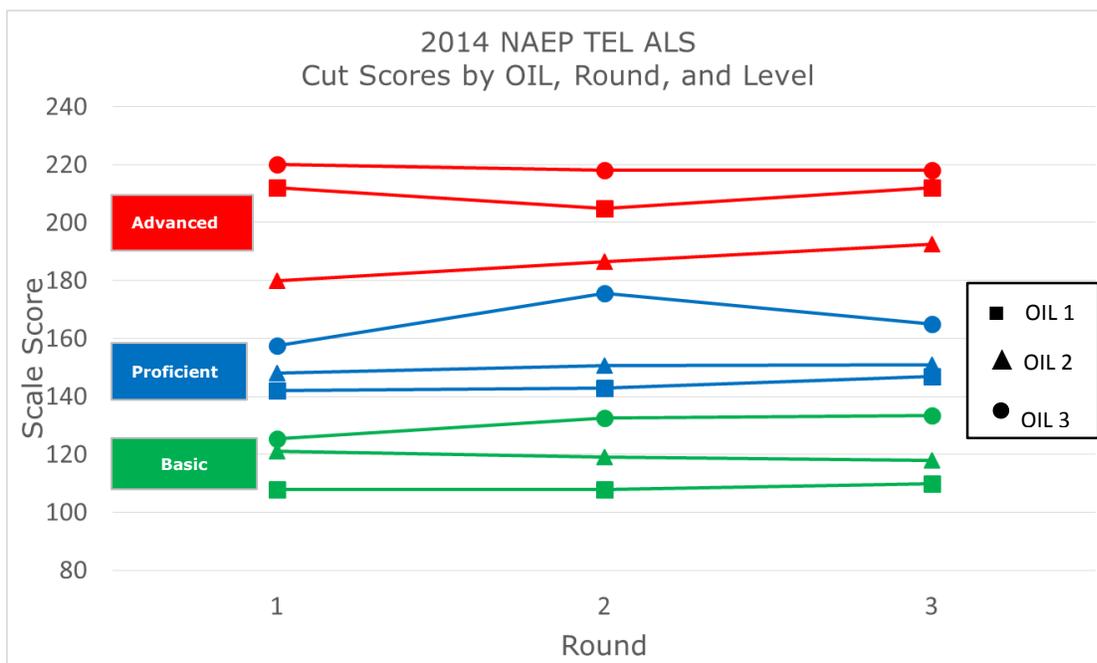


Figure 18: Cut scores by OIL and Round

Figure 19 shows the cut scores by round and achievement level for the table groups. The variability of the cut scores across the table groups is smallest for the Proficient level with the exception of Table 3, which produced a Proficient cut score more like the Advanced cut score for Tables 2 and 5.

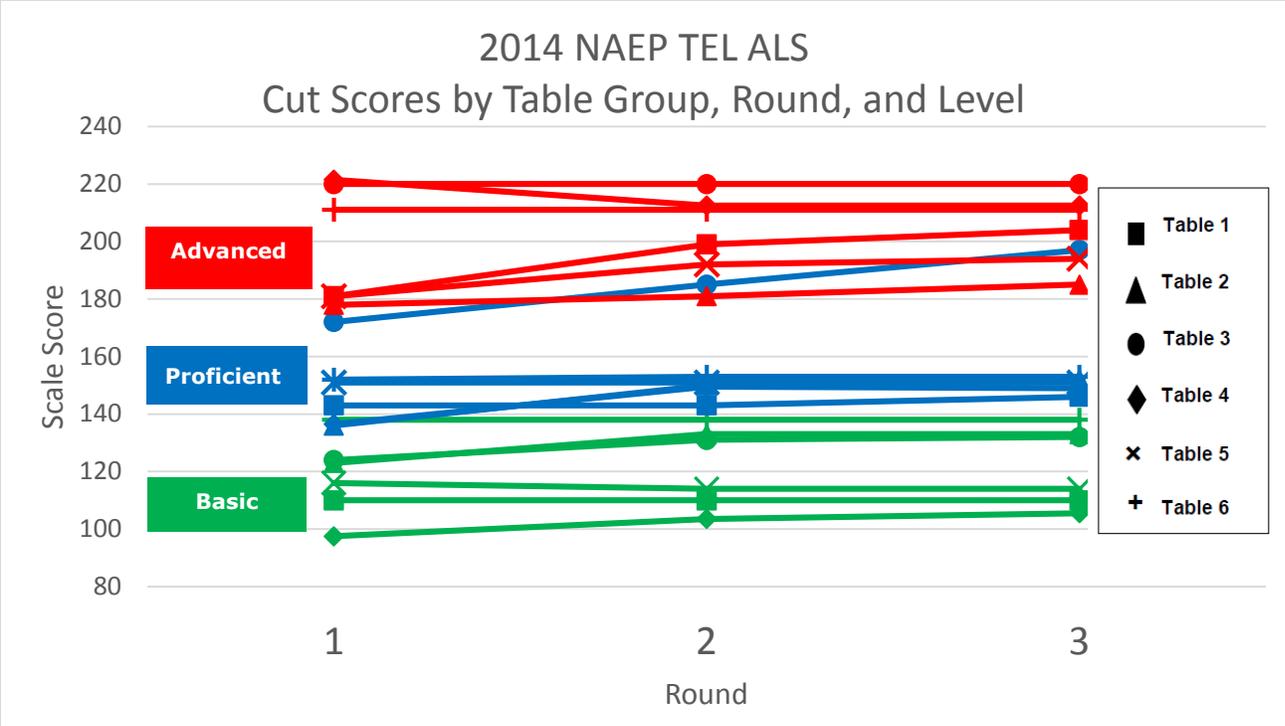


Figure 19: Cut Scores by Table Group, Round, and Level

The cut score for each achievement level was computed as the median of the cut scores for the whole group of panelists throughout the meeting and used in all feedback to the panelists. However, the TACSS and COSDAM expressed interest in the mean and standard deviations of the cut scores during review meetings, so they are presented here along with the median cut scores. Tables 36, 37, and 38 show the median and mean cut scores and standard deviations by round and table group for each achievement level. The table groups are arranged by the ordered item list (OIL) with which they worked.

Tables 36 through 38 show that the variability of the cut scores for the panelists at each of the tables was notably different especially after the first round, and generally decreased across rounds and became more similar for each achievement level.

Table 36: Median Cut Scores and Standard Deviations (S.D.) by Table and Round for the Basic Achievement Level

		Basic								
		Round 1			Round 2			Round 3		
Table	OIL	Median	Mean	S.D.	Median	Mean	S.D.	Median	Mean	S.D.
1	1	110	118	14.34	110	109	2.37	110	111	2.04
4	1	98	103	10.69	104	105	6.47	106	107	5.47
2	2	123	129	17.03	133	132	6.79	133	131	5.89
5	2	116	114	10.68	114	115	2.04	114	114	1.26
3	3	124	125	20.78	131	127	8.04	132	131	3.31
6	3	138	128	25.07	138	133	8.27	138	136	3.19
Total Group		116	119	19.43	116	120	12.89	116	121	12.23

Table 37: Median Cut Scores and Standard Deviations (S.D.) by Table and Round for the Proficient Achievement Level

		Proficient								
		Round 1			Round 2			Round 3		
Table	OIL	Median	Mean	S.D.	Median	Mean	S.D.	Median	Mean	S.D.
1	1	143	165	28.15	143	148	11.31	146	150	10.68
4	1	137	138	17.22	150	145	10.18	149	147	6.06
2	2	136	157	39.20	150	150	11.52	152	154	7.09
5	2	151	149	9.07	151	150	2.14	151	151	0.75
3	3	172	170	14.84	185	188	8.22	197	195	15.33
6	3	152	155	42.95	153	155	8.91	153	154	2.65
Total Group		150	155	29.96	151	156	17.28	151	158	18.32

Table 38: Median Cut Scores and Standard Deviations (S.D.) by Table and Round for the Advanced Achievement Level

		Advanced								
		Round 1			Round 2			Round 3		
Table	OIL	Median	Mean	S.D.	Median	Mean	S.D.	Median	Mean	S.D.
1	1	181	206	50.5	199	196	12.87	204	204	4.43
4	1	222	214	40.9	213	226	32.13	213	214	5.62
2	2	178	187	42.3	181	178	21.96	185	189	8.45
5	2	181	185	6.5	192	192	6.57	194	199	8.36
3	3	220	224	10.9	220	221	7.57	220	224	9.41
6	3	211	209	47.8	211	225	25.32	211	217	14.52
Total Group		193	204	40.2	40.2	207	27.63	209	208	14.80

The median cut scores and mean absolute differences for Round 3 are shown in Table 39, disaggregated by the background variables of panelist type, NAEP region, race/ethnicity, and gender. The different educator types of panelists produced similar cut scores with the largest difference being at the Basic cut for non-teacher educators. Unfortunately, the Northeast NAEP region was represented by only one panelist but the cut scores for panelists from the other regions were similar with the exception of the Basic cut score for panelists from the South region, which was lower than the cut scores for panelists in the Midwest and West regions. The cut scores for minority and non-minority panelists were very similar and females tended to recommend slightly higher cut scores than males.

Table 39: Medians and Mean Absolute Difference (MAD) of Round 3 Cut Scores by Demographic Variables

Group	N	Level					
		Basic		Proficient		Advanced	
		Median	MAD	Median	MAD	Median	MAD
Gen. Public	9	115.0	10.2	150.0	6.4	204.0	10.4
Non-Teacher	5	133.0	9.4	152.0	12.8	212.0	10.4
Teacher	17	116.0	11.0	152.0	12.4	209.0	11.8
Midwest	8	122.5	10.6	150.5	14.5	204.0	8.0
Northeast	1	112.0	0.0	171.0	0.0	212.0	0.0
South	17	114.0	11.9	151.0	9.2	205.0	13.1
West	5	125.0	8.2	153.0	8.0	211.0	9.6
Minority	14	119.5	11.9	152.0	11.3	208.0	11.8
Nonminority	17	116.0	10.5	151.0	10.3	209.0	11.1
Female	21	131.0	11.2	152.0	9.4	212.0	11.9
Male	10	113.0	8.9	150.0	13.3	204.0	8.0

The standard error is typically used as the estimate of the sampling variability of a statistic. However, the standard error of the median depends on the shape of the underlying distribution of the scores, which is generally unknown. Therefore, the standard error of the median must be approximated in some way. As in other NAEP ALS meetings, two methods were used in this study. The first was the bootstrap method (Efron & Gong, 1983), and the second was the Maritz-Jarrett procedure (Maritz & Jarrett, 1978). The details of the computation of these estimates are provided in the Technical Report. Table 40 shows the bootstrap and Maritz-Jarrett estimates of the standard error of the median for each round and achievement level.

Table 40: Standard Errors of Median Scores by Level and Round

Level	Round	Median	Bootstrap SE	Maritz-Jarrett SE
Basic	1	116	5.00	5.08
	2	116	4.48	4.25
	3	116	6.27	6.55
Prof.	1	150	4.39	4.54
	2	151	1.61	1.65
	3	151	0.83	1.01
Adv.	1	193	13.65	13.69
	2	205	3.94	4.11
	3	209	3.51	3.48

Table 41 shows the recommended scale score cuts after the consequences review from the ALS panel and from Pilot 2 conducted in June for comparison. The consequences review did not result in changes to the group cut scores for either panel. As can be seen in the table, the results from the Operational ALS were very consistent with those from Pilot 2.

Table 41: NAEP TEL Cut Score Recommendations by ALS and Pilot 2 Panels after Consequences Review

Level	Cut Scores and Percentages					
	Operational ALS Panel Results			Pilot 2 Panel Results		
	Scale Score	Percent In Level	Percent at or Above	Scale Score	Percent In Level	Percent at or Above
Basic	116	32%	83%	119	30%	81%
Proficient	151	48%	51%	151	46%	51%
Advanced	209	3%	3%	204	5%	5%

Exemplar Item Ratings

Appendix R provides a summary of panelists' exemplar item recommendations for the Basic, Proficient, and Advanced achievement levels. Several scenarios and items were designated for release by NCES. The goal in identifying possible exemplars was to maximize the number of items recommended to the Governing Board to increase the potential for more information to be available for demonstrating performance in each level. Pearson reviewed the criteria for selecting exemplar items originally specified in the Design Document in light of the eligible items once the cut scores from the ALS process were available and the panelist ratings of those items and suggested modifications to the selection criteria. The original criteria eliminated nearly all of the available items for use as exemplars. Therefore, the following modified criteria were proposed.

- 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.
- There should be a mix of items across the assessment areas, with at least one item from each of the three assessment areas.
- There should be a mix of items of each item format type, e.g., multiple choice and constructed response.

- Priority will be given to items with the highest frequency of panelists' ratings as "Should be Used" and with the lowest frequency of panelists' ratings as "Might be used."
- An item rated as "Should not be Used" by more than 20 percent of the panelists will be considered only if it is necessary to represent a particular feature of the assessment at a specific level of achievement.

These modified criteria were presented to the TACSS and accepted. They were then used to identify items that Pearson recommended to the Governing Board to serve as exemplars. The items that Pearson recommended to the Governing Board are presented in Appendix S.

Governing Board Action

The Governing Board reviewed the results of the ALS process during its meeting of November 19-21, 2015. Prior to that meeting, Pearson presented the results to CODSAM during a webinar meeting on November 3, 2015. CODSAM requested additional analyses from Pearson and Governing Board staff to address questions that arose during the meeting. Pearson provided the Round 3 mean panelist cut scores for the total group and panelist type (teacher, non-teacher educator, and general public), estimates of the standard errors of the median for the Round 3 median cut scores, and a list of TEL-related courses taught by teacher panelists. Governing Board staff compiled this information with additional information provided by them, and it was presented to CODSAM during a second webinar meeting held on November 17, 2015. The Governing Board adopted the cut scores that resulted from the operational ALS meeting for the Basic (116) and Advanced (209) achievement levels, and adopted a cut score of 158 for the Proficient achievement level (as compared to the operational ALS panel recommendation of 151) The deliberations of the Governing Board are described in the Addendum prepared by Governing Board staff and included at the end of this report.

As mentioned above, Pearson recommended a set of items to Governing Board staff for possible use as exemplars. Governing Board and NCES staff then reviewed the recommended items and identified the items that would serve as exemplars. The final set of items selected to serve as exemplars for each achievement level were approved by the Governing Board in November 2015 (see Appendix T).

Recommendations for Future Standard Settings

Pearson has a number of recommendations for future standard-setting studies focused on the review and evaluation of complex item types such as those on the NAEP TEL assessment. One recommendation is that when a usability study is conducted, it should comprehensively address multiple unique or complicated aspects of the study design. For example, the usability study for this project should have included a greater emphasis on implementing all aspects of the item review and ALS processes to ensure each step in the process was tested in the planned

electronic environment. Incorporating a fuller process into the usability study for the NAEP TEL ALS design would likely have revealed adaptations and revisions needed to the software that was replaced after Pilot 1.

A second recommendation is to ensure the process for evaluating the items and developing knowledge and skills statements is tested more thoroughly such that the time needed for reviewing complex item types within software that is new to panelists is clearly understood. This would allow for more accurate planning of agendas and timeframes from the start of the pilot work, avoiding adjustments that interrupt the flow of panelists' work.

A third recommendation addresses the need for adequate time to conduct pilot and ALS studies. Given the need to ensure a diverse group of panelists has adequate training and that NAEP item sets contain a large number of score points, Pearson recommends that NAEP ALS studies (both pilot and operational) should be planned as four and a half day events.

Time is also the focus of the fourth Pearson recommendation, to increase the contract timeline to allow for sufficient quality control checks of new software and other processes such as the use of computers and software from other contractors like the NAEP computers and software used in this study.

Another recommendation is to consider enhancing the process for collecting public comment on the Design Document by implementing processes to (a) better identify and communicate with the target group, (b) allow more time, and (c) follow-up with non-responders to differentiate those who have no comment from those who simply have not yet responded.

Finally, Pearson recommends that the Governing Board consider requiring one report at the end of the project in place of separate Process and Technical reports. The combined report would document all processes, technical aspects, and results of the ALS process.

References

- 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., & Lorié, W. A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives*, 2(2), 61–103.
- Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433–470). Westport, CT: Praeger.
- 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.
- 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.
- Loomis, S.C., and Bourque, M.L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G.J. Cizek (Ed.), *Setting Performance Standards: Concepts, methods, and perspectives* (pp 175-281). 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(2), 78-88.

Addendum: Governing Board Action on the Achievement Levels



**Governing Board Action on the
2014 Technology and Engineering Literacy Achievement Levels for Grade 8**

Sharyn Rosenberg, Ph.D.

Assistant Director for Psychometrics

During each quarterly meeting of the National Assessment Governing Board between August 2014 (shortly after the TEL achievement levels setting contract was awarded to Pearson) and November 2015 (when the Board took action on the TEL achievement levels), the Committee on Standards, Design and Methodology (COSDAM) received updates, provided guidance, and discussed the status of the TEL achievement levels setting project.

During the August 2015 quarterly Board meeting, COSDAM discussed the results from the June 2015 pilot study (Pilot 2). COSDAM members raised a concern about the Proficient cut score recommendation of 151, which would result in 51 percent of grade 8 students performing at or above Proficient in 2014. Although the results from each NAEP subject area framework are independent and should not be compared, it would be unprecedented to set a standard resulting in more than half of students performing at or above the NAEP Proficient level during the initial year of the assessment. The specific concern raised was that such a result is inconsistent with the Governing Board policy definition of NAEP Proficient as “competency over challenging subject matter” and may appear discordant with other indicators of student performance.

Working independently from Pilot 2, the operational achievement levels setting panel recommended the same Proficient cut score of 151. The results from the operational meeting were presented to COSDAM during a webinar on November 3, 2015, along with the conclusion of the Technical Advisory Committee on Standard Setting (TACSS) that there was no technical reason to recommend different cut scores. In addition to the procedural and technical information presented by Pearson (and summarized in the Process Report and Technical Report), COSDAM Chair Andrew Ho requested that Governing Board Assistant Director for Psychometrics Sharyn Rosenberg present information to COSDAM on: 1) the percent of students at or above Proficient on all NAEP assessments, based on the most recent administration at the national level; and 2) the history and context of adjustments that the Board made to panelist cut score recommendations from three previous achievement levels setting activities (1992 Mathematics; 1996 Science; 2009 Science).

COSDAM members did not raise any concerns about the panelist recommendations for the Basic and Advanced cut scores; discussion focused on the Proficient cut score. COSDAM members asked several questions about the procedures and results and requested the following pieces of additional information: teacher panelist background information; alignment of item difficulty and student ability distributions for additional NAEP subject areas; standard deviations and standard errors from other NAEP ALS activities; cut scores calculated by using the mean instead of the

median; cut score values after adjustment by one standard error upwards; and disaggregated cut scores by panelist type (teachers, non-teacher educators, and general public). The requested information was distributed to COSDAM on November 13, 2015.

On November 17, 2015, a second webinar was held to discuss the additional information. COSDAM Chair Andrew Ho prepared a memo to the committee and suggested that action on the Proficient cut score be framed as a binary decision: 1) accept the recommendation of the standard setting panel and set the Proficient cut score at 151 (51% at or above Proficient); or 2) acknowledge the recommended standard as a guideline and make a policy decision to set the Proficient cut score at 158, the mean of the panelists' judgments (43% at or above Proficient). The memo included an additional calculation of the standard error of the median (equal to 5.4 for Round 3) using bootstrapping and accounting for clustering by panelist tables. COSDAM members engaged in an extensive discussion of the information that was provided and the rationale for each option. The committee reached consensus on the second option.

On November 20, 2015, COSDAM reviewed the webinar discussions and unanimously approved a motion to recommend the following cut scores for full Board action: 116 (Basic), 158 (Proficient), and 209 (Advanced).

The full Board was first briefed on the TEL achievement levels setting procedures during the August 2015 meeting. On November 20, 2015, they were briefed on the results from the operational achievement levels setting meeting and on the COSDAM recommendation. The Board deliberated on the two options for the Proficient cut score. On the morning of November 21, 2015, the following cut scores were approved by a majority vote of 13 with three members opposing: 116 (Basic), 158 (Proficient), and 209 (Advanced). The Board also approved the exemplar items as recommended by staff.

Appendices

Appendix A: NAEP TEL ALS Facilitator Guide

Appendix B: Institutions Contacted to Nominate Non-teacher Educators

Appendix C: Organizations Contacted to Nominate General Public Representatives

Appendix D: Sample Panelists Recruitment Letters

Appendix E: Briefing Book

Appendix F: Lists of Panelists

Appendix G: NAEP TEL Achievement Level Descriptions

Appendix H: NAEP Policy Definitions

Appendix I: Item Map

Appendix J: ALS Questionnaires and Panelist Responses

Appendix K: Email Sent to Panelists Regarding Addition of Pilot 2

Appendix L: Pilot 1 Room Layout

Appendix M: Pilot 2 and ALS Room Layout

Appendix N: NAEP TEL ALS Operational Study Agenda

Appendix O: ALS PowerPoints

Appendix P: Excel Tools

Appendix Q: Feedback by Round

Appendix R: Summary of Exemplar Ratings

Appendix S: Exemplar Items Recommended to the Governing Board

Appendix T: Final Exemplar Item Set