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

Technical 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: Technical Report

Steve Fitzpatrick

Morgen Hickey

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 Chief Executive Officer
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 Sneider
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.

Technical Report
Table of Contents

List of Tables

List of Figures

Executive Summary

Overview

This report provides information about the technical aspects of procedures associated with the achievement levels-setting process 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 (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.

An item mapping procedure, based on the 2009 Science NAEP achievement levels-setting project, was designed for the 2014 NAEP TEL. Several studies were implemented to develop and refine procedures for developing the TEL achievement level cut scores and selection of exemplar items to recommend to the National Assessment Governing Board for use in reporting student performance on the NAEP TEL assessment. Table 1 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.

Table 1: 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

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 this report.

Complete information about procedural aspects and outcomes of studies conducted for this project are reported in the Process Report (Pearson, 2016).

Background

The entirely computer-based and highly interactive 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 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 Governing Board policy 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 assessment comprising both discrete items and performance-based scenarios, and 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.

Technical Advice

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. 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 served on a Technical Advisory Committee for Standard Setting (TACSS) and were convened for six in-person meetings and five webinars to provide advice for every key point in the process. They provided feedback on plans and materials before activities were implemented, and they reviewed results of the process and analyses afterward.

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 TACSS and COSDAM, Dr. Sharyn Rosenberg, the Governing Board's Assistant Director for Psychometrics and Contracting Officer's Representative (COR) for this contract, provided technical advice to Pearson throughout the project and participated in all TACSS meetings. Dr. Susan Loomis, who served as Project Director for several NAEP ALS meetings and as the Governing Board COR for several ALS activities, served as a technical consultant to Pearson. 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.

Coordination with NAEP Operations

The NAEP program is administered by the National Center for Education Statistics (NCES). The achievement levels-setting activities of the Governing Board require extensive resources from representatives of the NAEP Alliance contractors of NCES. Coordination of activities for the ALS project became more formalized through monthly meetings, via conference calls and webinars, involving key NCES NAEP staff and NAEP Alliance representatives meeting with Governing Board staff and members of the ALS contractor's staff. Some of the materials, data, and equipment used to conduct this project were provided through NCES by NAEP Alliance member companies. The following organizations provided resources for this project:

- Educational Testing Service (ETS)
- Pearson[1]
- Westat
- Fulcrum IT

[1] Pearson has a scoring contract with NCES separate and independent of the contract with the Governing Board to conduct the achievement levels-setting work.

Psychometric 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 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 that would 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. The item rating sets were presented to panelists as ordered item lists in an Excel file rather than in a paper ordered item book. Thus, they were referred to as OILs[2] rather than OIBs.

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.

Computation of Item Scale Values for a Response Probability of 0.67

All items in the assessment were calibrated together on an overall score scale by Educational Test Service (ETS), the Data Analysis and Reporting contractor for NCES, and provided to Pearson for this ALS project. Pearson used these data to calculate the scale score location of the dichotomous items and score point locations for the polytomous items using an RP value of 0.67. For multiple-choice (MC) and dichotomously scored short constructed-response (CR) items, the 3 parameter logistic (3PL) item response theory model was used; for

[2] The standard setting process was conducted using a digital interface and digital materials wherever possible. Very few paper materials were used during the meeting. For this reason, the usual paper-based terminology of an ‘ordered item book’ was replaced with the digitally oriented phrase ‘ordered item list’.

polytomously scored constructed-response items, the Generalized Partial Credit (GPC) item response theory model was used. A scale value was computed for every score point greater than 0 for each polytomous item.

A pseudo-NAEP scale score was used for all panelist materials that included a scale score. This pseudo-scale was used to disguise the true NAEP scale and to avoid the risk of having the true NAEP achievement level cut scores released before intended. The transformation of the NAEP scale scores to the pseudo-NAEP scale was simply to add 200 to the actual scale value. This produced scale values ranging from 200 to 500.

Cut Score Computation

The cut score for each panelist was computed as the scale value at the midpoint between the location of the item or score point where the panelists placed the bookmark and the scale value of the next higher item or score point in that OIL. The group cut score (calculated for each table and the whole group) was calculated by taking the median of the cut scores of the panelists in that group.

Ordered Item Lists

Using the item rating sets, the dichotomous items and polytomous item score points were arranged together in order of their RP67 values from lowest to highest based on student performance on the NAEP TEL assessment in 2014. These ordered item lists (OILs) were provided to panelists in a Microsoft Excel workbook. Each item rating group of panelists had a specific item rating set in their OIL.

Feedback to Panelists

Cut Score Distribution Charts

A chart showing the cut score distribution for the table group and the whole group was prepared as feedback for panelists after Rounds 1 and 2. These cut score distributions, also known as rater location charts, were presented as bar graphs that showed both the frequency that panelists had specific cut scores at each achievement level and the group cut score (median) for each level.

Item Maps

An item map was produced for use by panelists, starting with feedback after Round 2. The pseudo-NAEP scale score at which the item had a 0.67 probability of a correct response was used to map the items and score points. Items were ordered by difficulty with easiest at the

bottom and hardest at the top. Items were color coded in the item map to identify the OIL that included the item. They were also displayed in a separate column for each assessment area.

Each item was represented on the map by a unique identifier. Polytomous items with multiple score codes were displayed once for each score code above the minimum. Identifiers for polytomously scored items included an underscore "_" followed by the score code at which RP67 was reached.

Consequences Data and Questionnaire

As feedback, panelists were provided consequences data, also called impact data, which was based on their cut scores for Rounds 2 and 3. The data provided for the NAEP TEL were the percentages of grade 8 students in the 2014 assessment that scored within the range of each achievement level. Data indicating the estimated percentage of students scoring at each scale score were provided to Pearson by ETS for use in the process to establish the percentages for each achievement level.

Following Round 3, a consequences questionnaire was administered to panelists to collect their evaluations of the consequences data for Round 3 and their recommendations to the Governing Board regarding whether the panelist felt that the cut scores should remain as set in Round 3 or be modified for any or all three achievement levels.

For the post-Round 3 review, consequences data were presented in an interactive form so that panelists could adjust the cut score location and immediately see the change in the percent of students in each achievement level associated with the new cut score.

Exemplar Item Selection

Potential exemplar items that fell into one of the achievement levels identified by the Round 3 cut scores were selected from a set of items recommended for public release by NCES. Within each OIL, items that had a response probability of 0.67 at a score point within each achievement level range were presented to panelists as potential exemplar items. Each panelist only judged the appropriateness of the potential exemplar items that had been included in their item rating set; consequently, some items were viewed by only one-third of the panelists and some items were viewed by all panelists. The items were presented in an Exemplar Item List as an Excel file that indicated the item ID with a link to the item image, the item label, score code, maximum score code, the answer key for a multiple-choice item or a link to the scoring rubric,

and comments the panelists had entered in the OIL on the Item Review spreadsheet. Additional data about each item were provided for the exemplar selection process to assist panelists in judging the appropriateness of the items as exemplars of the achievement level: the scale score at which the item reached a 0.67 probability of a correct response, the average probability of a correct response across the achievement level range, and the probability of a correct response at the lower boundary (cut score) of each achievement level. 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. Pearson reviewed the panelists' ratings and identified a set of items for Governing Board staff that might be used as exemplars. Governing Board and NCES staff then reviewed the pool of potential exemplar items to select the items to include in the reports that would represent a variety of item types and formats across a broad range of content included in the assessment.

Process Evaluation Procedures

Process evaluation questionnaires were administered throughout the ALS process using Survey Monkey[3]. The questionnaires included both selected-response and open-ended questions that addressed the panelists' understanding 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 Process

Most responses were collected on a five-point Likert scales, but several responses were narratives that addressed specific aspects of the process. Responses were collected and reviewed after each questionnaire administration in order to make corrections and clarifications if needed.

Results of the NAEP TEL ALS Process

Table 2 shows the scale score cuts resulting from the Round 3 rating process of the

[3] https://www.surveymonkey.com/

Operational ALS meeting and Pilot 2. NAEP achievement is typically reported in terms of the percentages of students performing within and at or above each achievement level. An important reason for conducting a pilot study is to produce results that could be compared with the Operational ALS results and evaluated for reliability of the results. As can be seen in Table 2, the results for Pilot 2 and the Operational ALS are very similar.

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

Level	Cut Scores and Percentages					
	Operational ALS Panel Results			Pilot 2 Study 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%

Reliability Estimates

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. The cut score for each achievement level was computed as the median of the cut scores for the whole group of panelists. 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 underlining distribution of the scores, which is generally unknown. For that reason, 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 is the bootstrap method (Efron & Gong, 1983) and the second is Maritz-Jarrett procedure (Maritz & Jarrett, 1978). Table 3 shows the bootstrap and Maritz-Jarrett estimates of the standard error of the median for each round and achievement level.

Table 3: 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

Recommendations and Governing Board Action

Complete results of the ALS process were reviewed with the TACSS. The positive evaluation by panelists of the process provided evidence in support of the results of the process. In addition, the fact that the cut scores resulting from two independent sets of panelist, Pilot 2 and the Operational ALS, were so similar provided additional evidence of the consistency of the application of the procedures and the resulting outcomes.

Pearson presented the results to CODSAM during a webinar meeting on November 3, 2015. COSDAM 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 COSDAM 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.

Introduction

Under contract with the National Assessment Governing Board (Governing Board) initiated in July 2014, Pearson conducted a project to produce a set of recommendations for the Governing Board to consider in establishing achievement levels for reporting student performance on the grade 8 Technology and Engineering Literacy (TEL) assessment, one of the assessments in the National Assessment of Educational Progress (NAEP). The wholly computer-based 2014 NAEP TEL is based on the TEL Framework developed by the Governing Board. The first-ever TEL assessment was administered to a nationally representative sample of more than 20,000 grade 8 students in 2014.

The assessment is based on a diverse curriculum and is designed to measure the three interconnected areas of Technology and Society, Designs and Systems, and Information and Communication Technology. Furthermore, it is designed to measure three ways of thinking and reasoning that are used when solving a problem. These three ways of thinking are referred to as Practices and are expected to be demonstrated through Understanding Technological Principles, Developing Solutions and Achieving Goals, and Communicating and Collaborating.

The TEL assessment consists of both scenario-based tasks and discrete items. Scenario-based tasks assess students through their interaction with multimedia tasks using constructed-response and selected-response items to monitor student actions as they manipulate components of the systems and models that the tasks comprise (TEL Framework, 2013). There are short, medium, and long scenario-based tasks designed to take 10, 20, and 30 minutes of testing time, respectively. Discrete items are independent, stand-alone items that are not tied to a scenario. They may be either selected-response items or short constructed-response items.

The contract called for two reports, a technical report and a process report. This technical report provides information on the computational procedures and technical aspects of materials developed for the achievement levels-setting (ALS) meetings held for this project. The Process Report (Pearson, 2016) contains a detailed description of the methods and procedures used for the achievement levels-setting process, preparatory usability and pilot studies, and a presentation of the outcomes of the process.

Methodology

This section provides an overview of the standard setting method that was used for the achievement levels-setting process and the use of computers throughout the process. An overview of preliminary usability and pilot studies that were performed in preparation for the achievement levels-setting meeting is also provided.

Several possible standard setting methodologies were considered for recommending cut scores for the NAEP TEL. Pearson's evidence-based standard setting procedure (Beimers et al., 2012) 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 something that is widely taught in 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 came to the conclusion 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 forgoing the external validity evidence as part of the ALS process and the Governing Board staff agreed.

Pearson proposed to implement an item mapping process in which panelists made 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 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. There also is precedent for using the method with NAEP assessments (i.e., Mathematics 2005, Grade 12; Science 2009)

Use of Computers

The innovative characteristics of the TEL assessment that make it unique, such as the complex scenarios and the item interactivity, are characteristics that are also novel to the achievement levels-setting (ALS) process. The TEL is only the second NAEP assessment developed to be computer based. The first was the NAEP writing assessment administered in

2011. Unlike that assessment, the TEL includes several highly innovative features that required highly innovative procedures for standard setting. Given this consideration and the desire to conduct the standard setting activities using a digital platform, a dual-computer setup was necessary for the panelists' activities.

To evaluate the items in a standard setting process, panelists need to view the scenario-based items taking the scenarios as a whole in the context of other related information. For this reason, static screenshots were not sufficient to give panelists a full understanding of the requirements of the test items. Given the need to access the items as administered to students, it was necessary to use the computer software developed for the assessment. That assessment administration software could not be modified to accommodate software for standard setting, so it was necessary for panelists to use two computers throughout the ALS process. Secure computers provided by NCES contractor Westat contained the test items and tasks and allowed panelists to view the scenarios in action. These computers were referred to as the NAEP computers. Pearson provided computers that were used as the interface to the panelists for the standard setting activities. These computers were referred to as the ALS computers.

The NAEP computer consisted of a sample form of the assessment as experienced by students. 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 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 maintained the interactive nature of the items in the scenario but was different from how the students experienced the assessment. This was referred to as the "out of system" software on the NAEP computer. Nine discrete items were also included in the out-of-system software because they had an interactive component that Pearson judged to be difficult to understand when represented as PDF screenshots. The ALS computer provided by Pearson consisted of everything else that panelists needed to engage in the standard setting process. This included the ordered item lists (OILs), item review materials, and links to the website used to collect bookmarks and questionnaire responses and to present feedback to the panelists. These materials are described in a later section of this report and more fully in the Process Report (Pearson, 2016).

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 given the unique scenario-based and interactive nature of the test items. Table 4 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.

Table 4: 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

An initial usability study was conducted in December 2014 to investigate the way in which the unique assessment features would function within the standard setting procedures. Pearson licensed the use of a web-based standard setting software product that had been developed by Measurement Incorporated and successfully used in setting standards for another large-scale assessment. The primary purpose of the initial usability study was to observe how a small group of panelists would interact with the NAEP computer to review scenarios and the ALS computer to view individual items, record comments about items, and select bookmarks. The goal was to determine whether the dual-computer setup hindered the ALS process in any way. The driving concern was whether the participants would be able to switch back and forth between the two computers to navigate the items, including viewing them within their scenarios, and provide bookmark ratings using the dual-computer setup. The amount of time taken by the participants to perform an item review and bookmarking task for a small set of items was recorded and used to help plan the time allotted for those activities with the full set of items during the Operational ALS meeting. Following the initial usability study, modifications were made to the licensed standard setting software to accommodate the full requirements of the ALS

process including adapting the way cut scores were calculated, and customizing the feedback to panelists to align with the requirements specific to the NAEP TEL ALS project. More details about the usability study are provided in the Process Report (Pearson, 2016).

A pilot study (Pilot 1) was conducted in March 2015 to test the planned item mapping standard setting procedures and the standard setting software. The study provided the opportunity to try out the ALS panelist setup and allowed for planning and modifications prior to the Operational ALS meeting. The standard setting software did not perform as expected during Pilot 1. It functioned as anticipated for item review. However, it did not group the scenario-based items by scenario, which introduced some confusion for panelists as they completed the item reviews. When panelists began implementing the standard setting rounds, the complexity of the TEL ALS procedures led to unanticipated interruptions in the functioning of the software. In addition, there were complications with the processing of panelists' standard setting ratings, producing the feedback needed for panelists' immediate use between rounds of standard setting, and the processing of panelist's questionnaire response data. In light of these issues, the decision was made to abandon the plan to use this software and for Pearson to develop Excel-based procedures for panelists to use instead.

Because of the challenges encountered during Pilot 1 in March 2015, it was deemed necessary to conduct a second pilot study (Pilot 2) to include one additional day for the ALS panel meeting and pilot the implementation of an alternative digital interface. The new digital interface included (a) the use of an Excel workbook to display screenshots of individual items for panelists and to allow panelists to record information at each step in the standard setting process, (b) the housing of materials on the desktop of each panelist's laptop computer, (c) access for panelists to a secure FTP site for posting feedback results as the standard setting process progressed, and (d) the use of an online survey tool to collect each panelist's responses to questionnaires.

Pilot 2 was conducted in June 2015 and supported the use of the revised agenda planned for the Operational ALS Meeting. The standard setting procedures and digital interface for the materials worked very well during Pilot 2. Only minor revisions to procedures and the time allotted for some activities were made for the Operational ALS meeting in September 2015.

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 Moye served as the content facilitator and onsite content expert.

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) and were 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 meetings of the TACSS are included in Appendix A to provide additional details of the technical considerations and deliberations regarding the procedures implemented for this ALS project.

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 on 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 contract 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 Huff[4]
Vice President of Strategy & Execution, ACT

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

Dr. Marianne Perie[5]
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)

Summary of Key TACSS Recommendations for the NAEP TEL ALS Procedures

- The TACSS recommended a special study to focus on panelists' ability to use and respond to tasks using two computers in a standard setting context. The TACSS members expressed concern that going back-and-forth between ratings software and viewing items on the same computer might disrupt the standard setting process. The TACSS recommended a study examining panelists' use of two computers. The study would have panelists use the two computers as proposed in the Design Document[6]. This study should be completed before the pilot study.
- For items that were modified during scaling (some pairs of items were combined and scaled as one; scoring categories for some polytomous items were collapsed) the TACSS

[4] 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.
[5] Marianne Perie resigned from the TACSS after the February, 2015 meeting.
[6] A Design Document was developed in the initial phase of the project describe that described in detail the NAEP TEL achievement levels-setting activities to be implemented. This document was intended to provide the foundation for all ALS activities.

recommended that the items be used in the ALS process and that panelists be given a clear explanation of the treatment.

- The TACSS recommended that all survey response options have a label and that none of the labels reflect a positive bias. They recommended rewording questions with response options reflecting degree of clarity and a middle response category that used neither/nor wording to place the concept of clarity or adequacy in the stem rather than the options and use a consistent set of response options such as "Strongly Agree" to "Strongly Disagree" with "Neutral" as the middle category. Some questions asked about the appropriateness of time spent on certain activities using five response options ranging from "Far too long" to "Far too short." The TACSS suggested eliminating the two extreme options and reducing the number of response options to three. They also recommended that questions be included about the computer configuration and other aspects of the standard setting meeting unique to the TEL assessment.
- The TACSS expressed concern that the revised mechanism for presenting the ordered item lists (OILs) to the panelists was so different than it had been during Pilot 1. They felt strongly that the Excel-based OIL should be tried out with a sample of people before the ALS panel meeting. It was also necessary to add at least another whole day to the agenda for the ALS meeting. Because of the extent of the changes after Pilot 1, the group agreed that another pilot study should be conducted to include all of the standard setting activities using the revised procedures and the revised agenda. As this would not be feasible to conduct before the planned June ALS meeting, the TACSS strongly recommended that the June panel meeting be treated as second pilot study and that a third panel meeting be scheduled to serve as the Operational ALS meeting.
- The TACSS recommended that achievement levels results from other NAEP subjects not be presented to panelists to inform their judgments in the ALS process. The appropriate audience for this type of comparison is the Governing Board that is charged with policy decisions regarding the NAEP achievement levels.
- The TACSS recommended that consequences data be presented to panelists as the percentage of students performing within, as opposed to at or above, each level. Data reporting performance at or above the level may be a more appropriate statistic for policy groups than for ALS panelists.
- The original plan for collecting public comment on the outcomes of the ALS process could not be implemented after the originally planned ALS meeting was converted to a second pilot study. Both the timing and confidentiality considerations of collecting public comment on the results of the rescheduled ALS meeting presented challenges leading to a TACSS recommendation to omit this activity from the process.

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
- 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

Pearson[7]

- 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

Psychometric Procedures

This section describes the technical procedures implemented during and after the Operational ALS meeting. The subsections describe the item pool, the division of the items into three ordered item lists (OILs), computation of item scale values for a response probability of 0.67, the use of pseudo-NAEP scales throughout the process, the method of collecting panelists' bookmarks judgments, round-by-round feedback, mapping potential exemplar items to achievement levels, the selection of potential exemplar items, reliability estimates, and collection of process evaluation questionnaires.

[7] 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.

Description of Item Pool

Items, items statistics, and student performance data from the 2014 administration of TEL were used in the achievement levels-setting process. The grade 8 assessment was administered using 231 items arranged into 21 assembly units. Based on the initial item analysis, ETS eliminated one item from assembly unit 10, combined three item pairs into three clustered items due to item dependencies, and collapsed the score categories of 13 polytomous items. This resulted in a total of 227 items available for the ALS process. Table 5 shows the items that were clustered together, and Table 6 shows the items for which score categories were collapsed. Table 7 summarizes the item bank by assembly unit, assessment area, item type, and score points.

Table 5: Items Combined to Form Cluster Items

Assembly Unit	NAEP ID	NAEP IDs of items contributing to cluster
14	D0100TL	D0100CL, D0101CL
14	D0102TL	D010201, D010301
17	D0117TL	D0117CL, D0119CL

Table 6: Items with Collapsed Score Categories

Assembly Unit	NAEP ID	Original Categories	Collapsed Categories
01	D0001CL	0,1,2	0,0,1
01	D0002CL	0,1,2	0,0,1
01	D0229CL	0,1,2	0,0,1
03	D0020CL	0,1,2	0,0,1
03	D002601	0,1,2	0,1,1
06	D0052CL	0,1,2	0,1,1
06	D0054CL	0,1,2	0,0,1
10	D0078CL	0,1,2	0,0,1
15	D010801	0,1,2	0,1,1
17	D0117TL	0,1,2,3,4	0,0,1,1,2
18	D0219CL	0,1,2	0,1,1
18	D0178CL	0,1,2	0,1,1
19	D0128CL	0,1,2,3	0,0,0,1
20	D0174CL	0,1,2	0,0,1

Table 7: Summary of TEL Item Pool by Assessment Unit.

Assembly Unit #	All Items	Number of Items					All Score Points	Number of Score Points	
		Assessment Area[1]			Item Type[2]			Point Type	
		D&S	I&CT	T&S	MC	CR		MC	CR
AU01	12	0	12	0	2	10	16	2	14
AU02	6	1	0	5	1	5	13	1	12
AU03	11	0	2	9	3	8	15	3	12
AU04	10	10	0	0	3	7	14	3	11
AU05	11	0	4	7	0	11	13	0	13
AU06	9	0	9	0	1	8	14	1	13
AU07	8	0	4	4	1	7	14	1	13
AU08	9	0	5	4	2	7	13	2	11
AU09	12	4	0	8	4	8	18	4	14
AU10	12	7	5	0	4	8	17	4	13
AU11	11	0	7	4	4	7	16	4	12
AU12	9	0	4	5	2	7	17	2	15
AU13	11	5	6	0	1	10	15	1	14
AU14	10	7	0	3	3	7	20	3	17
AU15	11	5	0	6	5	6	15	5	10
AU16	10	6	0	4	0	10	15	0	15
AU17	13	9	0	4	3	10	18	3	15
AU18	12	5	0	7	4	8	19	4	15
AU19	13	6	7	0	3	10	16	3	13
AU20	16	0	10	6	7	9	18	7	11
AU21	11	2	9	0	2	9	14	2	12
Total	**227**	**67** 30%	**84** 37%	**76** 33%	**55** 24%	**172** 76%	**330**	**55** 17%	**275** 83%

[1] D&S = Design and Systems, I&CT = Information and Communications Technology, T&S = Technology and Society
[2] MC = Multiple Choice, CR = Constructed Response

ETS provided item images for constructed-response items. Pearson created images for the remaining item types (cluster, and selected-response). Scoring rubrics were provided by the Pearson scoring contractor for hand-scored items and by ETS for machine-scored items. In addition, a subset of scoring rubrics had to be re-created by Pearson to account for item treatments that occurred during analysis by ETS.

Assignment of Items to Rating Sets

The standard setting process was conducted using a digital interface and digital materials, wherever possible. Very few paper materials were used during the meeting. For this reason, the usual paper-based terminology of an 'ordered item book' was replaced with the digitally oriented phrase, 'ordered item list'. In order to construct the ordered item lists (OILs), the NAEP TEL item pool of 20 scenarios and 97 discrete items[8] was divided into three unique sets, A, B, and C, and a common set of items. Items were divided into item rating sets to limit the number of items reviewed by each panelist and minimize possible fatigue. Each unique item set, which was assigned to one of three groups of panelists, was combined with the common set to form an OIL. Figure 1 depicts the structure of the item rating sets and their assignment to panelist groups.

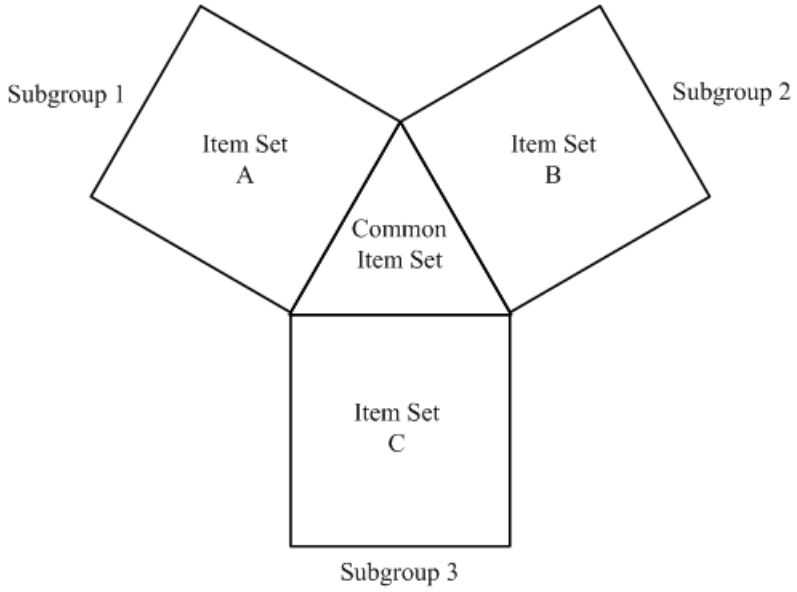


Figure 1: The Assignment of Item Sets to Panelist Subgroups.

The item sets A, B and C 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 assessment area (Technology and Society, Design and Systems, Information and Communication Technology), (b) item type (scenario-based or discrete), and (c) item difficulty. All items from a scenario were assigned to the same OIL. 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 OILs in order to have items to serve as examples in

[8] There were originally 98 discrete items but one was dropped from the final scaling of the items.

group discussions with panelists during the standard setting meeting. It was possible to follow this design because the item pool was found to support the construction of three item sets that were roughly equivalent in number of discrete item blocks and scenarios, representation across the three subscales, representation across item type, and median and range of item difficulty.

Table 8 shows the characteristics of each of the three item rating sets. 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 the 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 8: 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

After the creation of the item rating sets, the dichotomous items and polytomous item score points were arranged together in order of their RP67 values from lowest to highest based on student performance on the NAEP TEL assessment in 2014. These ordered item lists (OILs) were provided to panelists in a Microsoft Excel workbook.

Computation of Item Scale Values for a Response Probability of 0.67

All items in the assessment were calibrated together on an overall score scale by Educational Test Service (ETS), the Data Analysis and Reporting contractor for NCES. Pearson received a data file from ETS containing the item parameters for the items. These were used to calculate the scale score location of the dichotomous items and score point locations for the polytomous items using an RP value of 0.67.

The computation of item scale values in the item mapping method begins with the computation of score probabilities. Let U_i represent the score point on item i and let θ represent student achievement on the overall scale. For multiple-choice (MC) and dichotomously scored short constructed-response (CR) items, the 3 parameter logistic (3PL) item response theory model was used:

$$P(U_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + exp[-Da_i(\theta - b_i)]} \qquad (1)$$

where D is 1.7, a_i is the item discrimination parameter, b_i is the item difficulty parameter, and c_i is the pseudo-guessing parameter for multiple-choice items ($c_i = 0$ for dichotomously scored constructed-response items). For dichotomous items using this IRT model, the theta corresponding to a desired response probability can be obtained from a simple formula. Let r represent the desired response probability and denote the item parameters as above, then θ_{rp}, the theta corresponding to that probability, is

$$t_{rp} = b + \frac{\ln\left(\frac{r - c}{1 - r}\right)}{Da} \qquad (2)$$

For polytomously scored constructed-response items, the Generalized Partial Credit (GPC) item response theory model was used:

$$P(U_i = h \mid \theta) = \frac{\exp\left[\sum_{r=0}^{h} Da_i(\theta - b_i + d_{ir})\right]}{\sum_{k=0}^{m_i} \exp\left[\sum_{r=0}^{k} Da_i(\theta - b_i + d_{ir})\right]} \qquad (3)$$

where m_i is the maximum score on the item, and d_{ih} is the threshold parameter on item i for score point h, h=0,1,…,m_i, and $d_{i0} = 0$.

A scale value was computed for every score point greater than 0 for each polytomous item. There is no closed form solution for the theta that corresponds to a specified response probability in a target category or higher for the GPC model. For that reason, an iterative procedure was used to find the theta that produced a value of 0.67 (+/- 0.0001) in equation 4 below for each response greater than 0.

$$P(U_i \geq h \mid \theta) = \sum_{j=h}^{m_i} P(U_i = j \mid \theta) \qquad (4)$$

Item Scale Values

A pseudo-NAEP scale score was used for all panelist materials that included a scale score. This was done to disguise the true NAEP scale and to avoid the risk of having the true NAEP achievement level cut scores released before intended. The transformation of the NAEP scale scores to the pseudo-NAEP scale was to add 200 to the actual scale value. This produced scale values ranging from 200 to 500.

Round-by-Round Feedback

Round One

Following Round 1, the facilitator provided panelists with their individual cut scores, the distribution and median value for the cut scores identified by the panelists in their table group, and the distribution and median value for the cut scores identified by the panelists as a whole group. The cut scores for each panelist were computed as the scale value at the midpoint between the location of the item or score point where the panelists placed the bookmark and the scale value of the next higher item or score point in that OIL. The group cut scores (calculated for each table and the whole group) were calculated by taking the median of the cut scores of the panelists

in that group. The panelists at each table discussed the Round 1 results for their table and the total group. There was no whole group level discussion of the Round 1 results.

Round Two

Initial feedback from Round 2 consisted of the median cut scores for the table and the whole group and the cut score distribution for the table and the whole group. The facilitator asked the panelists to engage in discussion of the results at their table. Panelists were also given an item map showing the distribution of items and score points along the score scale separated by assessment area and color coded by item rating set. The process facilitator then instructed panelists to view a chart showing the consequences data from Round 2 and led the whole group in a discussion of the impact data.

Round Three

The Round 3 feedback provided to each panelist consisted of the panelist's cut scores, the cuts scores for each panelist at the table, the median cut scores for the whole group, and the distribution of cut scores for the whole group. The process facilitator walked the panelists through the results and then instructed them to view a chart showing the consequences data from Round 3 and led the whole group in a discussion of the impact data.

Consequences Review

After reviewing the Round 3 results, panelists were presented with an interactive bar chart displaying the percentage of students in each achievement level based on the whole group Round 3 cut scores. Panelists were able to enter different cut scores for each achievement level and evaluate the consequences of raising or lowering the cut scores. They could observe the effect of moving the cut scores on the percentage of students scoring in each level. They independently evaluated whether the Round 3 whole group cut score recommendations should be changed, and if so, they recorded their cut score recommendations.

Table 9 reports the number and percentages of panelists who recommended changes to one or more Round 3 group cut scores after viewing the consequences data from the Round 3 final ratings. When panelists proposed a change, they were asked to record their reason(s) in the consequences questionnaire. In general, the majority of the panelists provided rationales based on the content of the items they believe fell at the cut scores.

Table 9: Number and Percentage of ALS Panelists Who Recommended Cut Score Changes During Consequences Review

	Basic	Proficient	Advanced
Total Number of Changes Recommended (%)	8 (26%)	8 (26%)	12 (39%)
Number (%) Lowered	4 (13%)	4 (13%)	9 (29%)
Number (%) Raised	4 (13%)	4 (13%)	3 (10%)

Panelist modifications during the consequences review did not alter the group-level cut score recommendation (median across all panelists) for any of the three cut scores.

Summary of Results

Table 10 shows the median cut scores by round, and Table 11 shows the scale score cuts resulting from the Round 3 rating process of the Operational ALS meeting and Pilot 2.

Table 10: ALS Panel Median Cut Scores by Round

Round	Scale Score[9]		
	Basic	Proficient	Advanced
1	116	150	193
2	116	151	205
3	116	151	209
Consequences Review	116	151	209

Table 11: 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%

[9] As previously mentioned, different NAEP-like scales were used to avoid the risk of having the NAEP achievement level cut scores released before intended. As in past ALS studies, the NAEP-like scale was a linear transformation of the NAEP reporting scale. All results in this report have been transformed back to the actual NAEP reporting scale.

The results from the Operational ALS are consistent with those from Pilot 2 conducted in June. As with the Operational ALS, the cut score recommendations did not change as a result of the consequences review during Pilot 2.

Mapping Potential Exemplar Items to Achievement Levels

Exemplar items are a part of the official set of information that is to be recommended to the Governing Board as part of the achievement levels-setting process. Exemplar items serve to communicate to the public the types of knowledge, skills, and abilities that are required for performance within each of the three NAEP achievement levels. The role of exemplar items in communicating performance on the NAEP TEL is especially important because this is an entirely new, innovative area of assessment for the NAEP program. The items selected to represent 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.

Items and score points 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 scale score cut and less than the Proficient scale score cut were assigned to the Basic level. Items were similarly assigned to the Proficient and Advanced levels. The average response probability for each dichotomous item and polytomous score point greater than zero within the assigned achievement level was calculated along with the response probability at each of the three scale score cut points. The equation used to calculate the average response probabilities within an achievement level is shown below,

$$Average\ Probability = \left(\sum_{j=L}^{H} P_{j,k} * f_j\right) / \sum_{j=L}^{H} f_j \quad (5)$$

where L represents the scale score cut for an achievement level, H represents the highest scale score in the achievement level, $P_{j,k}$ represents the probability of a correct response at scale score j to a dichotomous item or the probability of a response in category k of a polytomous item, and f_j is the relative frequency of students at scale score j. The summation was over scale scores in increments of one.

Criteria for Selection of Exemplar Items Recommendations

The following criteria were used for the selection of items to recommend to the Governing Board as exemplars for each achievement level:

- 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.

The criteria for selecting items and the items under consideration for recommendation to the Governing Board as exemplars were vetted by the TACSS before they were presented to the Governing Board for approval. Items meeting the statistical criteria were first presented to TACSS to evaluate the results relative to the statistical criteria. The final set of items deemed appropriate for use as exemplars for each achievement level were recommended to the Governing Board in November 2015 for approval to use in reporting achievement levels for TEL.

Reliability Estimates

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. The cut score for each achievement level was computed as the median of the cut scores for the whole group of panelists. 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 underlining 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 is the bootstrap method (Efron & Gong, 1983) and the second is Maritz-Jarrett procedure (Maritz & Jarrett, 1978). For the bootstrap approach, 1,000 sample replications with replacement were used, stratifying on panelist type (teacher, non-teacher educator, and general public). The median cut score for each achievement level was calculated for each sample and the standard deviation of the medians across the samples was used as the bootstrap estimate of the standard error of the median. Stratification was used because the Governing Board policy specifies the percent of panelists from each group that must be included.

If a different sample of panelists had been used for the ALS meeting, it would have varied somewhat in representation of gender, ethnicity, NAEP region, and other demographic variables, but would have had the same breakdown of panelist type. The Maritz-Jarrett estimate of the standard error of the median was calculated for an odd number of judges as described in Price and Bonett, 2001.

Let *n* be the number of judges, which can be represented as *2m+1* when *n* is odd, and *X* be the vector of order statistics. The Maritz-Jarrett estimate of the variance of the median is

$Variance_{median} = A_{J2} - (A_{J1})^2$

where $A_{Jr} = \sum_{i=1}^{n} W_i X_i^r$

and $W_i = F_\beta\left[\left(\frac{i}{n}\right), (m+1), (m+1)\right] - F_\beta\left[\left(\frac{i-1}{n}\right), (m+1), (m+1)\right]$

where F_β denotes the cumulative Beta distribution function.

The estimate of the standard error of the median is then the square root of the variance. Table 12 shows the bootstrap and Maritz-Jarrett estimates of the standard error of the median for each round and achievement level.

Table 12: 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

Process Evaluations

Panelists responded to a number of questionnaires at different points during the ALS process using the Survey Monkey[10] online survey tool. The list below shows the nature of each questionnaire and when it was administered.

- 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

The panelist responses to the questionnaires were tabulated and reviewed in real time throughout the meeting. If the data indicated any lack of understanding or misconception about the process, the process facilitator addressed the issue before proceeding. In addition, the process facilitator asked if panelists had any questions about the process or the feedback they received after each round, and time was allotted for group discussion and clarification of any questions or uncertainty before the next round began.

Recommendations and Governing Board Action

Complete results of the ALS process were reviewed with the TACSS. The positive evaluation by panelists of the process provided evidence in support of the results of the process. In addition, the fact that the cut scores resulting from two independent sets of panelist, Pilot 2 and the Operational ALS, were so similar provided additional evidence of the consistency of the application of the procedures and the resulting outcomes.

Pearson presented the results to CODSAM during a webinar meeting on November 3, 2015. COSDAM 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),

[10] https://www.surveymonkey.com/

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 COSDAM 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.

Pearson also 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 be recommended to the Governing Board to serve as exemplars.

Materials

This section describes the technical aspect of preparing the materials provided to panelists during the achievement levels-setting (ALS) meeting. The subsections describe how panelists were selected and assigned to tables, the test form given to panelists, item review materials, item mapping, computation of cut scores, cut score feedback and rater location charts, consequences feedback and questionnaire, and the exemplar item rating form.

Other materials prepared for the meeting were sent to panelists in advance. This is consistent with the belief that distributing advanced materials is considered the first step in training panelists (Cizek & Bunch, 2007; Loomis, 2012). The Process Report (Pearson, 2016) provides additional details about all of the materials used in the ALS process and describes the communications used to prepare panelists for the ALS process.

Division of Panelists and Item Pools into Rater-Groups

Governing Board policy specifies the target composition for the ALS panel to be 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) in the subject area. The target number of panelists for the ALS meeting was 30. However, 33 were identified and selected to ensure that at least 30 would attend the meeting if some were unable to participate at

the last minute. Shortly before the Operational ALS meeting, two panelists notified Pearson that they would not be able to attend leaving a total of 31 panelists.

Panelists were assigned to one of six tables such that six panelists sat at table number four, and five panelists were seated at the remaining tables for individual work and group discussion. 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.

The overall NAEP TEL item set was divided into 3 smaller item sets in order to reduce the number of items reviewed by each panelist and thus minimize the cognitive demand on the panelists. The task pools were constructed to be equivalent in terms of task types and range of difficulty. The procedure used to divide the items into three item rating sets is described above in the Psychometric Procedures section. Two table groups were assigned to each of the three item sets with tables 1 and 4, 2 and 5, and 3 and 6 assigned to the same set of items.

Test Form Administered to Panelists

On the first day of the meeting, panelists took an actual intact test form of the 2014 NAEP TEL assessment for grade 8 under conditions similar to those implemented for the student administration (timed, preceded by the same tutorial, and using the same computers). All panelists took the same test form. ETS provided Pearson with a list of the eight operational test forms where all core items[11] had been chosen for public release in the Nation's Report Card. The criteria to use released items for the sample test form was intended to provide in-depth exposure to some of the items that would subsequently be rated as potential exemplar items. The sample test form was selected to represent several different aspects of the assessment (i.e., to include both scenario-based tasks and discrete items, a variety of item types, and coverage across the content areas and practices).

After completing the assessment, panelists were provided an Excel spreadsheet with item images, answer keys for the selected-response items and scoring guides and rubrics for the

[11] The operational assessment consisted of 60 minutes of assessment time, where students received two assembly units of core items that contributed to the scoring of the results. When a student finished the core items well before the 60 minutes, additional items (called supplemental items) were administered but the scoring of such items did not contribute to the assessment results. There were no intact test forms where both core and supplemental items were all intended for public release, so only the core items were considered when choosing the sample test form.

constructed-response items in the core set. This information allowed panelists to review their performance on the assessment and the performance requirements for each item they completed. A sample of the Excel spreadsheet is shown in Figure 2.

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

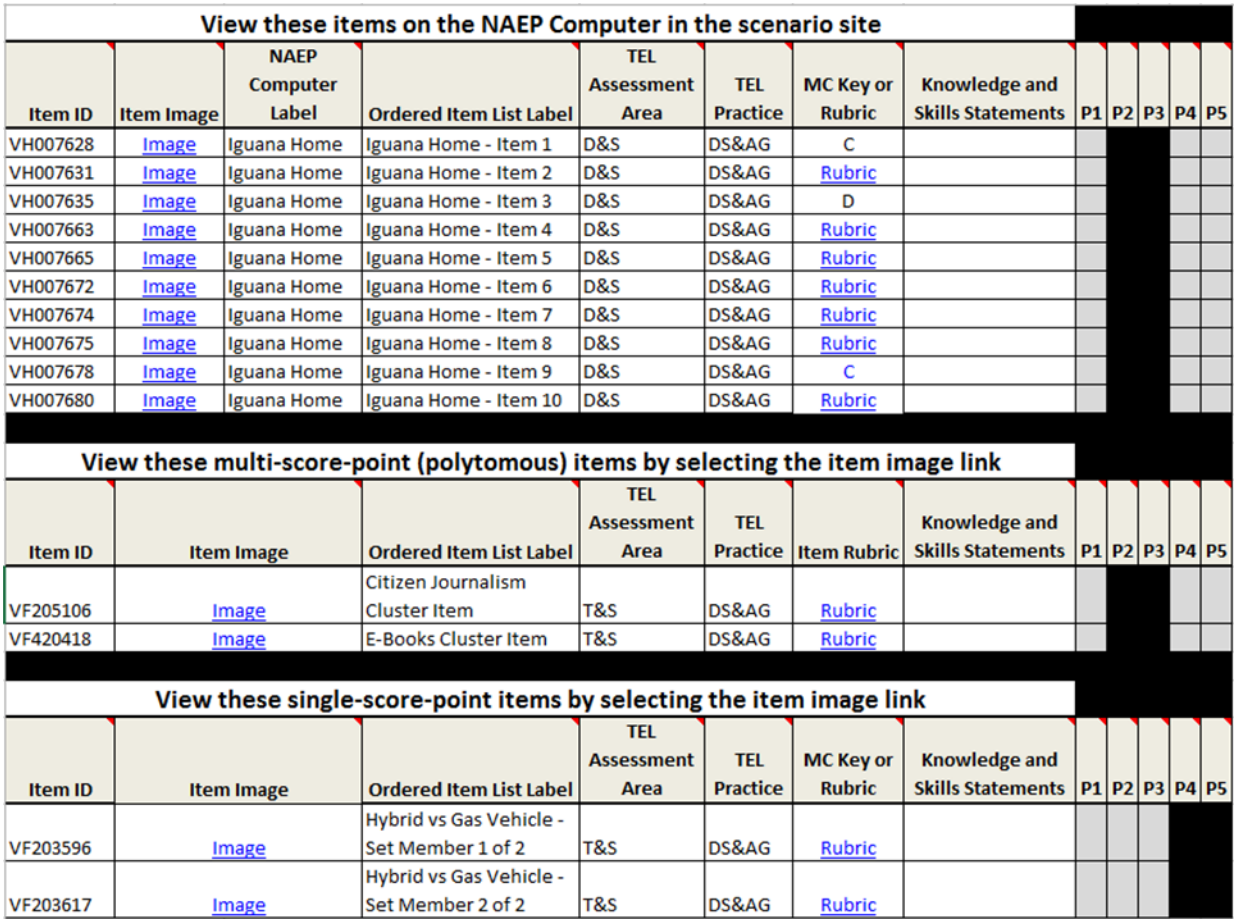
Figure 2: Segment of Test Form Review Spreadsheet

Item Review

Item review is an important activity in which the panelists study the items in order to make informed judgments about the knowledge and skills necessary for a correct response when placing their bookmarks during the standard setting process. While important, this activity can be lengthy and tiring for panelists. In order to reduce the potential for fatigue and to reduce the amount of time required for this task, panelists were asked to write knowledge and skills statements for only some of the items in their assigned OIL. The number of items in each OIL were divided as evenly as possible given that some scenarios contained more items than others yet needed to be assigned intact. Generally speaking, each panelist was assigned about 60% of

the items to review, including approximately five scenarios and 52 to 55 items total. Each item from an OIL was assigned to two or three panelists at the same table. In addition, all panelists reviewed all the items from the common item subset after taking the sample assessment.

Panelists wrote statements describing the key knowledge and skills targeted by the items. These notes were recorded and stored in the panelist's copy of the assigned OIL. The OIL associated these notes with the item across worksheets within the workbook such that each panelist's notes were available whenever the panelist viewed the item or an item score point. Figure 3 shows a sample of the item review spreadsheet. The columns to the far right labeled *P1* through *P5* indicated the items for which each panelist was to write knowledge and skills statements during the item review process. Panelists filled in knowledge and skills statements for items they did not review during table discussions of the knowledge and skills statements.



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 3: Segment of the Item Review Spreadsheet

Ordered Item Lists

Item mapping is a judgmental decision-making process utilizing an ordered item list (OIL) with items and item score points arranged in order from the easiest to the most difficult, based on student performance on the NAEP TEL assessment in 2014. As previously described, both panelists and the TEL item set were divided into three groups for the ALS meeting, pairing each panelist rating group with a particular item set. Each item set included a subset of common items. The OILs were implemented as Excel spreadsheets containing item identifiers, RP67 scale score locations, answer keys, scoring rubrics, and other content metadata for the items and score categories. Figure 4 shows a sample of the OIL spreadsheet. The term 'Score Code' was used to refer to the score categories of the items rather than the term 'Score Point' because the scoring rubrics and sample student responses available to the panelists labeled the score categories as

score codes ranging from 1 to the number of categories rather than as score points starting at zero.

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 4: Segment of the Item Review Spreadsheet

Item Map

An Item Map was distributed to panelists with the Round 2 results. Colors were used in the item map to identify the different OILs assigned to the different panelist subgroups. On the item map, items were ordered from easiest at the bottom to hardest at the top. Polytomous items were displayed once for each score point above the minimum score. The pseudo-NAEP scale score at which the item had a 0.67 probability of a correct response was used to map the items and score codes.

An illustration of a section of the item map used in the Operational ALS is shown in Figure 5. The items are separated into assessment area columns. Each item is represented on the map by a unique identifier. Extended constructed-response items include an underscore "_" followed by the score code. Short constructed-response (or dichotomous) items only have one score code that receives credit so their unique identifier does not include an underscore and number. The color of an item unique identifier on the map indicates whether the item is in OIL 1 only (blue), OIL 2 only (green), OIL 3 only (pink), or in common across all three OILs (orange).

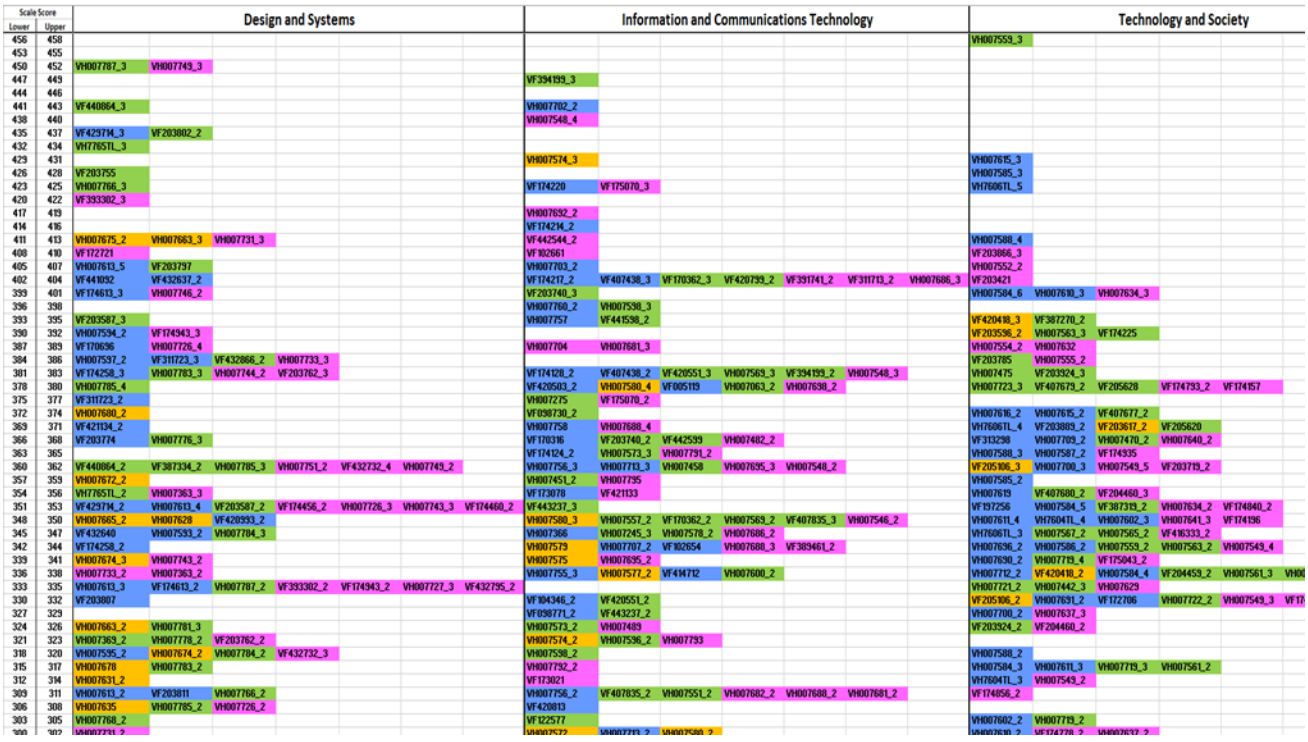


Figure 5: An Illustration of the Operational ALS Item Map.

Computation of Cut Scores

Panelists indicated their bookmark for each achievement level in columns four through six of the ordered item list (See Figure 4). The panelists were told to place their bookmark on the last item where they believed a student with borderline performance would have a 2/3 or better chance of getting the item correct. After they had decided on their three bookmarks for the round, they entered them into an online survey tool. A questionnaire was administered after each round of ratings. 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 panelist's entries against the item marked with a bookmark in the OIL for quality control purposes. Figure 6 shows a screenshot of the online survey question used to collect the panelists' bookmarks. The panelist cut score for an achievement level was then computed as the scale score value that fell at the midpoint between the scale score of the item bookmarked by the panelist and the next item in the OIL.

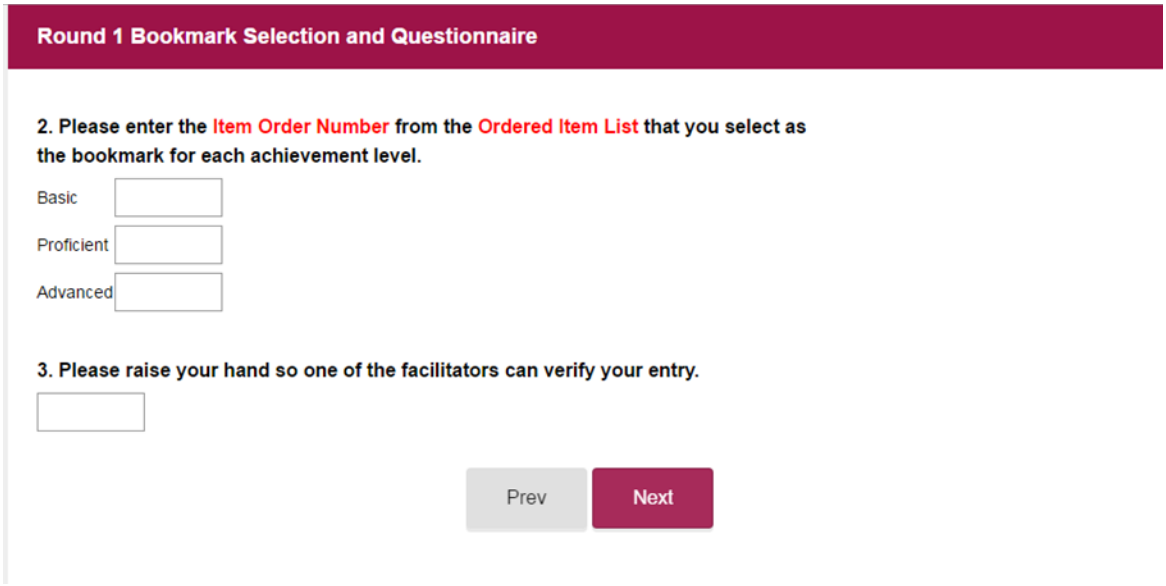


Figure 6: Bookmark Recording Form

Cut Score Feedback Provided After Each Round

For a given ALS round, the recommended cut score for each achievement level was computed as the median value across the distribution of panelists' cut score recommendations. Panelists were provided with a table that showed the whole group cut score as well as the median cut score for their table and their individual cut scores resulting from the round. A sample is shown in Figure 7. As noted in the section above describing the psychometric procedures, the table median was not provided to panelists after Round 3.

	Item Order Number			Scale Score		
Panelist ID	**Basic**	**Proficient**	**Advanced**	**Basic**	**Proficient**	**Advanced**
MMH05WI2	34	70	108	125	155	193
SAE05MD4	19	54	105	112	145	192
SCC05OK5	30	73	90	123	159	178
SKC05AR1	7	39	91	95	133	179
WKE05MT3	22	66	96	116	151	181
Table Median				116	151	181
Group Median				116	150	193

Figure 7. Sample Panelist Cut Score Feedback

Rater Location Chart for Each Round

As part of the feedback after the first and second rounds of item mapping, panelists received two rater location charts that displayed the distribution of cut scores at a table level and for all panelists. After Round 3, only the whole group rater location chart was provided. The rater location charts also displayed the median cut score (represented by a vertical dashed line) at each achievement level for the table or whole group, depending on the chart level. An example of a rater location chart is shown in Figure 8. The rater location chart was color coded so that panelists could identify the three achievement levels.

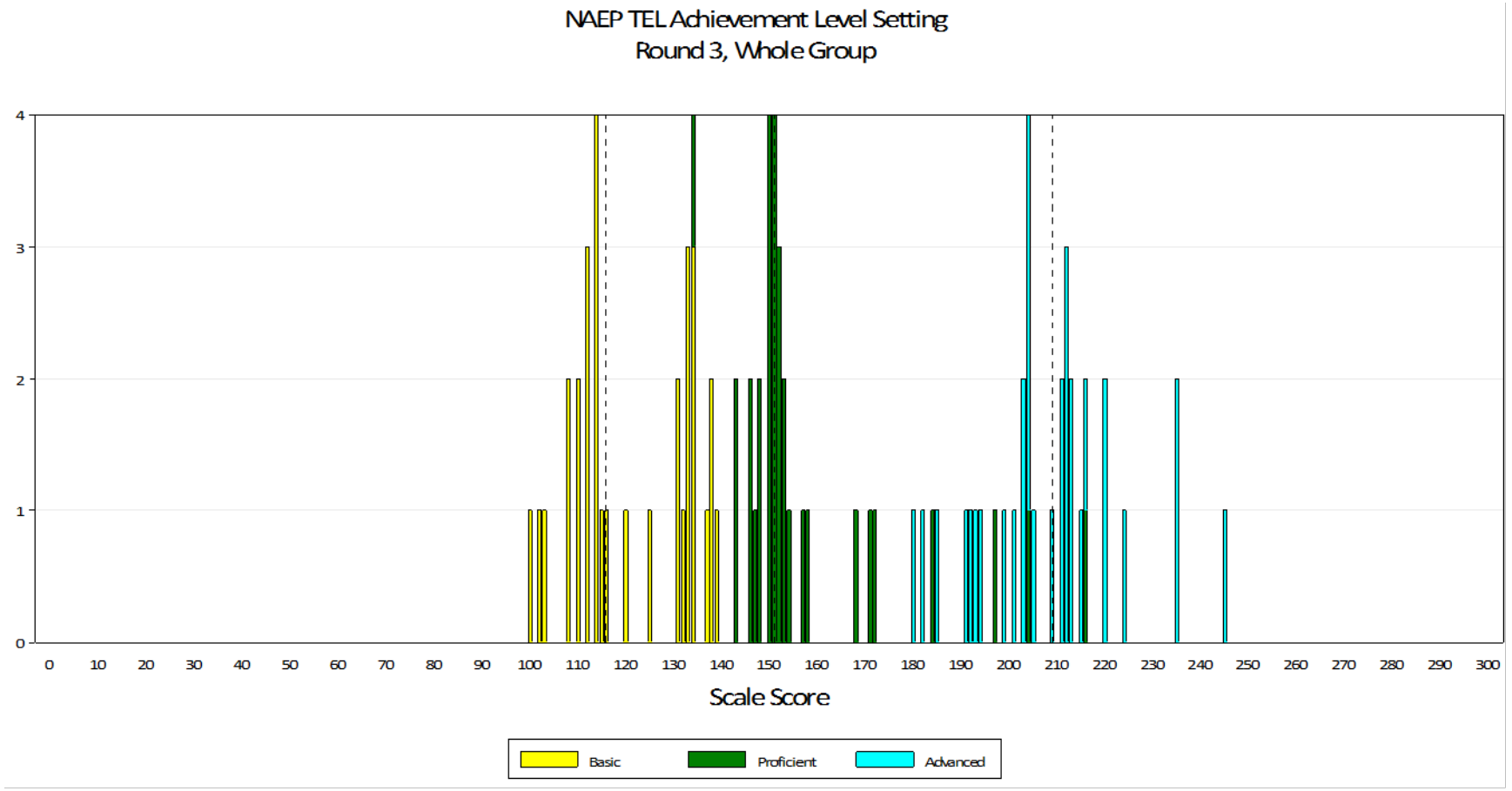
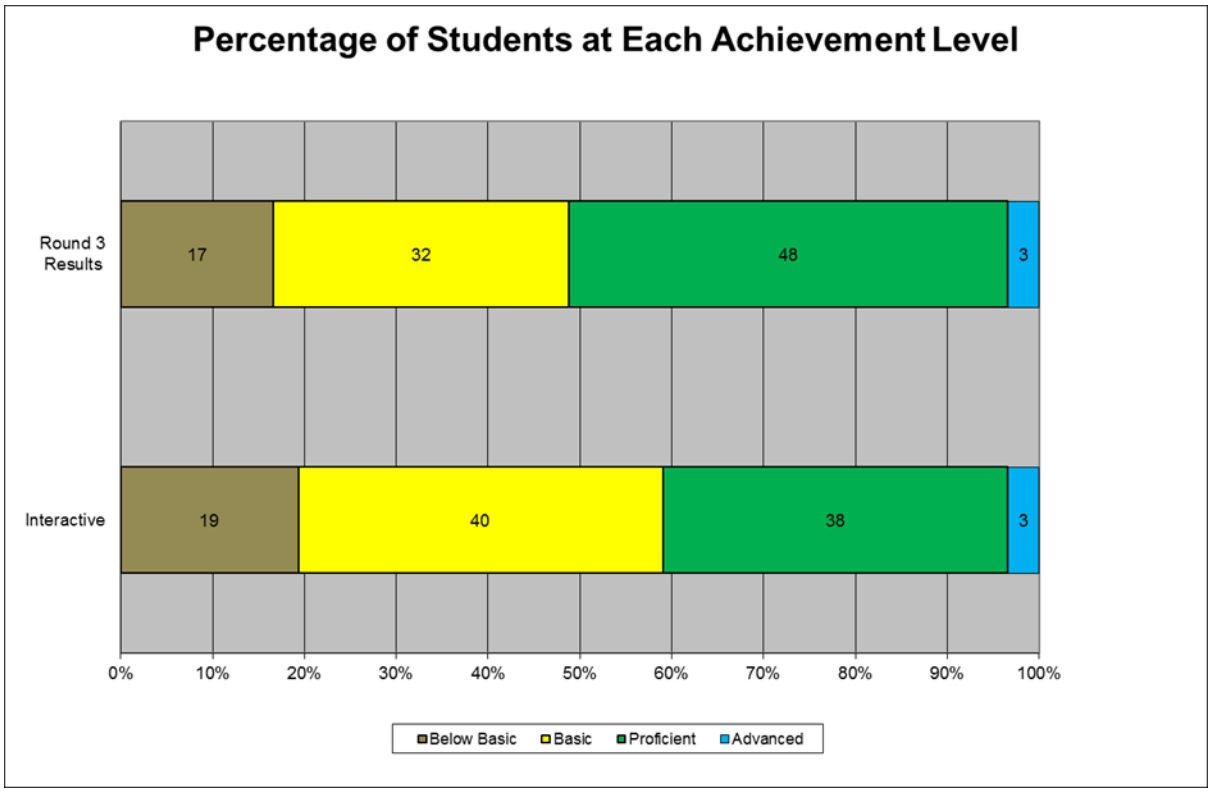


Figure 8: An Illustration of a Panel-Level Rater Location Chart.

Consequences Feedback and Questionnaire

The feedback after Rounds 2 and 3 included a chart of the percentage of students within each achievement level. This feedback has been called consequences data in past NAEP ALS meetings, and it has also been called impact data. These percentages were based on the performance distribution of the grade 8 NAEP TEL assessment operational results and the median group cut scores for the relevant round.

After reviewing and discussing the results of Round 3, including the consequences data, panelists completed the consequences review during which they considered the group's recommendations, provided individual feedback regarding the recommended cut scores, and used an interactive tool to understand the impact of adjusting the Round 3 cut score recommendations. The tool allowed panelists to explore the implications for consequences data of different cut score recommendations by adjusting the placement of the cut scores and seeing what percentage of students would fall into the achievement levels using the new cut scores. Figure 9 shows the interactive consequences tool. 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 percentage of students in each level. The Basic and Proficient cut scores have been changed in the figure and the effect can be seen in the lower bar labeled *Interactive*.



	Basic	Proficient	Advanced
Round 3 Cut Scores	116	151	209
New Scale Score Value	120	158	209

Figure 9: Interactive Consequences Tool

Exemplar Item Rating Form

Potential exemplar items were selected from a set of items that NCES recommended for public release and that also fell into one of the achievement levels identified by the Round 3 cut scores. Within each OIL, all potential exemplar items that had a response probability of 0.67 at a score code within each achievement level range were presented to panelists to judge their appropriateness as exemplar items. Each panelist reviewed only the potential exemplar items that had been included in their item rating set; consequently, some items were viewed by only one-third of the panelists and some items were viewed by all panelists. The items were presented in an Exemplar Item List as an Excel file that indicated the position of the item or score code in the panelist's OIL, the scale score location (which is the pseudo-NAEP scale score at which the item or score point has a 0.67 response probability), the item ID with a link to the item image, the

item label, score code, maximum score code, the answer key for a multiple-choice item or a link to the scoring rubric, and comments the panelists had entered in the OIL on the Item Review spreadsheet. A section of the spreadsheet that was used to present this information to the panelists is shown in Figure 10.

Item Order	Location	Item ID	Label	Score Code	Max Score Code	MC Key or Rubric	K&S Comments Entered during Item Review
23	317	VH007611	Chicago - Item 2 Cluster	3	4	Rubric	understand relationship between beliefs and synthesize data and points of view
25	319	VH007674	Iguana Home - Item 7	2	3	Rubric	Full credit for a students NO answer and their ability to analyze a behavior and use evidence to compare it to an expected result to determine if problem solving criteria have been met. Partial for NO answer, valid observation but no references to evidence.
27	321	VH007574	Recreation Center - Item 2 Cluster	2	3	Rubric	Full Credit given to students that can properly organize a message to be the most effective for certain audience
28	324	VH007663	Iguana Home - Item 4	2	3	Rubric	Full credit given as students answer NO and are able to successfully predict all four factors necessary to provide a fit habitat. Partial credit for identifying 2 or 3 factors.
33	331	VF205106	Citizen Journalism Cluster Item	2	3	Rubric	To receive full credit students need to understand individual privacy issues, how technology and society influence each other the positive and negative effects of technology and how to cite the work of others.
36	333	VH007613	Chicago - Item 3 Cluster	3	5	Rubric	analyze correct clip for animation, design function, sequencing,
39	336	VH007577	Recreation Center - Item 5 Cluster	2	2	Rubric	Full credit for identifying and selecting the most effective elements of a message based upon a criteria.
41	337	VF420418	E-Books Cluster Item	2	3	Rubric	Full credit correctly students need to recognize and understand benefits, analize and compare the advancements and benefits of an evolving technology (evaluate tradeoffs) explain cost benefits
44	340	VH007674	Iguana Home - Item 7	3	3	Rubric	Full credit for a students NO answer and their ability to analyze a behavior and use evidence to compare it to an expected result to determine if problem solving criteria have been met. Partial for NO answer, valid observation but no references to evidence.
46	341	VH007575	Recreation Center - Item 3	2	2	D	Students are able to evaluate a resource and select most convincing one based upon criteria.
48	342	VH007579	Recreation Center - Item 6	2	2	B	Students demonstrate the ability to select an appropriate resource based upon a criteria.
51	343	VF174258	Design Steps - Set Member 1 of 2 Cluster Item	2	3	Rubric	Full credit given as students correctly analyze and order all proper steps in construction. Partial credit given for three or more correctly ordered steps.

Figure 10: Sample Exemplar Item Selection List Part 1

The average response probability within the assigned achievement level was presented to the panelists along with the response probability at each of the three scale score cut points in the spreadsheet columns to the right of those shown in Figure 10. This section of the spreadsheet is shown below as Figure 11.

Level	Avg Prob in Level	Prob at Basic Cut	Prob at Prof Cut	Prob at Adv Cut	Exemplar Rating
B	0.83	0.66	0.92	1.00	Should not be Used
B	0.78	0.64	0.88	0.99	Might be Used
B	0.74	0.64	0.82	0.97	Should not be Used
B	0.72	0.63	0.79	0.94	Might be Used
B	0.70	0.53	0.82	0.98	Might be Used
B	0.68	0.48	0.83	0.99	Might be Used
B	0.66	0.50	0.78	0.96	Should not be Used
B	0.65	0.52	0.76	0.96	Might be Used
B	0.62	0.46	0.74	0.94	Should be Used
B	0.64	0.57	0.71	0.88	Should not be Used
B	0.63	0.53	0.71	0.91	Should not be Used
B	0.60	0.44	0.73	0.97	Should be Used

Figure 11: Sample Exemplar Item Selection List Part 2

Panelists reviewed each item and the data for the item and then independently identified potential exemplar items for each achievement level. The last column in Figure 10 contained a drop down box for each item or score point so the panelists could rate the item as "Should be Used," "Might be Used," or "Should not be Used" as an exemplar for the achievement level.

Results of the exemplar rating task are provided in Appendix R of the Process Report (Pearson, 2016). Once the cut scores from the ALS process were available, Pearson reviewed the criteria for selecting exemplar items originally specified in the Design Document in light of the eligible items 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, 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. These are provided in Appendix S of the Process Report (Pearson, 2016). Governing Board staff then reviewed the recommended items and identified the items that would be recommended to the Board to serve as exemplars. They added the restrictions that cluster items (where multiple items were scored as a single set) were not eligible and that only the highest score code for polytomous items was eligible for selection.

References

ACT, Inc. (2010). *Developing achievement levels on the 2009 National Assessment of Educational Progress in science for grades four, eight, and twelve: Process report*. Iowa City, IA: Author.

ACT, Inc. (2005). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade twelve mathematics: Process report*. Iowa City, IA: Author.

Beimers, J. N., Way, W. D., McClarty, K. L., & Miles, J. A. (2012). Evidence based standard setting: Establishing cut scores by integrating research evidence with expert content judgments. *Research Bulletin 21*, Pearson Education, Inc. www.pearsonassessments.com.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37*(1), 36–48.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433–470). Westport, CT: Praeger.

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. (2012). Selecting and training standard setting participants: State of the art policies and procedures. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods and innovations* (pp. 107-134). New York, NY: Routledge.

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.

National Assessment Governing Board (2013). *Technology and Engineering Literacy Framework for the 2014 National Assessment of Educational Progress*. Washington, D.C.

Pearson (2016). *Developing Achievement Levels on the 2014 National Assessment of Educational Progress in Grade 8 Technology and Engineering Literacy: Process Report.* Austin, TX.

Price, R.M., & Bonett, D.G. (2001). Estimating the variance of the sample median. *Journal of Statistical Computation and Simulation, 68*(3), 295-305.

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.

Appendix A: Minutes from TACSS Meetings