National Assessment
Governing Board

# Developing Achievement Levels on the National Assessment of Educational Progress in Writing Grades 8 and 12 in 2011

Submitted to:

National Assessment Governing Board
800 North Capitol Street, NW, Suite 825
Washington, DC 02002-4233
Phone: 202.357.6938

This study was funded by the
National Assessment Governing Board
under Contract ED-NAG-10-C-0003.

Submitted by:
Measured Progress
100 Education Way
Dover, NH 03820
Phone: 603.749.9102

## ALS Writing Process Report

September 2012

**PANELIST NAMES REDACTED BY GOVERNING BOARD**

# *Developing Achievement Levels on the 2011 National Assessment of Educational Progress in Grades 8 and 12 Writing*

# *Process Report*

**Luz Bay**

with

**Chris Clough**

**Jennifer Dunn**

**Wonsuk Kim**

**Leah McGuire**

**Tia Sukin**

**September 2012**

# National Assessment Governing Board

## BOARD MEMBERSHIP
## (2011–2012)

**Honorable David P. Driscoll, Chair**
Former Commissioner of Education
Melrose, Massachusetts

**Mary Frances Taymans, SND, Vice Chair**
Sisters of Notre Dame
National Education Office
Bethesda, Maryland

**Andrés Alonso**
Chief Executive Officer
Baltimore City Public Schools
Baltimore, Maryland

**David J. Alukonis**
Former Chairman
Hudson School Board
Hudson, New Hampshire

**Louis M. Fabrizio**
Data, Research and Federal Policy Director
North Carolina Department of Public Instruction
Raleigh, North Carolina

**Honorable Anitere Flores**
Senator
Florida State Senate
Miami, Florida

**Alan J. Friedman**
Consultant
Museum Development and Science Communication
New York, New York

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

**Doris R. Hicks**
Principal and Chief Executive Officer
Dr. Martin Luther King, Jr. Charter School for Science and Technology
New Orleans, Louisiana

**Honorable Terry Holliday**
Commissioner of Education
Kentucky Department of Education
Lexington, Kentucky

**Richard Brent Houston**
Principal
Shawnee Middle School
Shawnee, Oklahoma

**Hector Ibarra**
Eighth-Grade Teacher
Belin Blank International Center and Talent Development
Iowa City, Iowa

**Honorable Tom Luna**
Idaho Superintendent of Public Instruction
Boise, Idaho

**Honorable Jack Markell**
Governor of Delaware
Wilmington, Delaware

**Tonya Miles**
General Public Representative
Mitchellville, Maryland

**Dale Nowlin**
Twelfth-Grade Teacher
Columbus North High School
Columbus, Indiana

**Honorable Sonny Perdue**
Former Governor of Georgia
Atlanta, Georgia

**Susan Pimentel**
Educational Consultant
Hanover, New Hampshire

**W. James Popham**
Professor Emeritus
University of California, Los Angeles
Wilsonville, Oregon

**Andrew C. Porter**
Dean
Graduate School of Education
University of Pennsylvania
Philadelphia, Pennsylvania

**B. Fielding Rolston**
Chairman
Tennessee State Board of Education
Kingsport, Tennessee

**Cary Sneider**
Associate Research Professor
Portland State University
Portland, Oregon

**Blair Taylor**
President and Chief Executive Officer
Los Angeles Urban League
Los Angeles, California

**Honorable Leticia Van de Putte**
Senator
Texas State Senate
San Antonio, Texas

**Eileen L. Weiser**
General Public Representative
Ann Arbor, Michigan

Ex-officio Member
**John Q. Easton**
Director
Institute of Education Sciences
U.S. Department of Education
Washington, DC

# Table of Contents

# List of Tables

# List of Figures

# Executive Summary

## Overview

This report provides a detailed description of the achievement levels–setting (ALS) process implemented by Measured Progress for the purpose of establishing achievement levels for the 2011 National Assessment of Educational Progress (NAEP) in grades 8 and 12 writing. In addition to providing ALS recommendations in the form of cut scores, this process prompted a revision of the achievement levels descriptions (ALDs) to be used in the establishment of those cut scores, and resulted in the identification of student work examples illustrative of performance at each achievement level and recommendations for the improvement of future ALS implementations. Each of these supporting activities and products is described in detail later in this report.

The National Assessment Governing Board (hereafter referred to as the Governing Board) has been charged by Congress to establish achievement levels for NAEP, showing that the levels are useful, reasonable, and valid. This charge has been operationalized by the Governing Board in the form of three grade-specific achievement levels: Basic, Proficient, and Advanced. The policy of the Governing Board places special emphasis on the setting of achievement levels using a national consensus approach with participation from teachers, other educators, and noneducator members of the general public who are knowledgeable in the specific content area. The contract for this work was awarded on September 23, 2010, and Measured Progress began working with the Governing Board in October 2010 to develop an ALS process consistent with Governing Board's charge and policies.

## Activities Preceding Standard-Setting Studies

### *Technical Advisory Committee on Standard Setting*

As required by the Governing Board, Measured Progress appointed an external Technical Advisory Committee on Standard Setting (TACSS) comprising six experts with national or international reputations for measurement and standard setting. The TACSS met nine times over the course of the project, providing input that influenced the implementation of the standard-setting method: the design of field trials and special studies; the conduct of the ALS meeting; the data analysis procedures, including computation of cut scores; and the formulation of conclusions and recommendations presented to the Governing Board.

### *Body of Work Standard-Setting Method*

Measured Progress implemented the Body of Work (BoW) method for setting the NAEP writing achievement levels for grades 8 and 12. The BoW method (Kingston, Kahl, Sweeney, & Bay, 2001) is the flagship standard-setting method for Measured Progress, and it is the method deemed most appropriate for writing assessments because it was developed specifically for use with performance assessments that are designed to allow for a range of student response scores (Kahl, Crockett, DePascale, & Rindfleisch, 1995). As planned, Measured Progress implemented a technologically enhanced version of the BoW method with the use of the Body of Work Technological Integration and Enhancements (BoWTIE) software, which allowed for a computer-based standard setting. The development and use of BoWTIE is a major advance in the NAEP ALS process. The reasons for developing BoWTIE include the following:

- to overcome the logistical difficulties in materials preparation
- to enhance security of materials
- to promote “green” procedures, minimizing the need for hard-copy materials
- to enhance the overall efficiency and effectiveness of the process

Computerization for efficiency was recommended by the Governing Board in the statement of work.

### Standard-Setting Studies

The ALS project comprised several standard-setting meetings; the outcomes of each meeting contributed to modifications implemented in subsequent meetings. A list of these meetings and their dates follows:

- Field trial, September 22–23, 2011
- Pilot study, November 15–18, 2011
- Special study, November 18–19, 2012
- Field trial 2, January 27, 2012
- Operational ALS meeting, February 7–10, 2012
- Special study 2, February 10–11, 2012

The field trial was used to test the logistics of implementing the process. The pilot study was used to test the process. The special study was used to explore the relationship between performance on the 2011 assessment, based on the new writing framework, and performance on the 2007 assessment, based on the writing framework first implemented for the 1998 NAEP. Results of this special study revealed apparently large changes in the way the achievement levels were understood in the 2011 as compared to the 1998 standard setting. As a result, achievement level descriptions

(ALDs) were modified, which necessitated field trial 2 to test these modified ALDs. Field trial 2 supported use of these modified ALDs for the operational ALS meeting. The achievement levels recommended to the Governing Board were those resulting from the operational ALS meeting. Special study 2 again compared the 2007 and 2011 NEAP writing assessment results.

### ***Panelists***

Process design for the operational ALS meeting called for 30 panelists per grade-level panel, distributed by panelist types as follows: 55% teachers, 15% nonteacher educators, and 30% general public. A total of 55 panelists, 27 for grade 8 and 28 for grade 12, were recruited for the operational ALS meeting. The panelists were identified through a four-stage process:

- Stage 1: Select districts and identify nominators
- Stage 2: Contact nominators and request nominations
- Stage 3: Notify nominees and request acceptance of nomination
- Stage 4: Select and recruit panelists

Much of the initial communication was handled through a recruitment module built into the BoWTIE system. Personal contacts were made as needed to improve the yield of nominees. Key panel demographics are found in Table 1.

*Table 1: Operational ALS Meeting Panel Composition*

| Demographic Variable | Attribute | Grade 8 | | Grade 12 | | All | | Goal |
|---|---|---|---|---|---|---|---|---|
| | | *n* | % | *n* | % | *n* | % | % |
| Panelist Type | Teachers | 16 | 59 | 15 | 54 | 31 | 56 | 55 |
| | Nonteacher Educators | 5 | 19 | 5 | 18 | 10 | 18 | 15 |
| | General Public | 6 | 22 | 8 | 29 | 14 | 25 | 30 |
| Gender | Female | 22 | 81 | 19 | 68 | 41 | 75 | 50 |
| | Male | 5 | 19 | 9 | 32 | 14 | 25 | 50 |
| Race/Ethnicity* | Caucasian | 23 | 85 | 25 | 96 | 48 | 91 | 80 |
| | Non-Caucasian | 4 | 15 | 1 | 4 | 5 | 9 | 20 |
| NAEP Region | Midwest | 6 | 22 | 8 | 29 | 14 | 25 | 35 |
| | Northeast | 5 | 19 | 4 | 14 | 9 | 16 | 20 |
| | South | 6 | 22 | 6 | 21 | 12 | 22 | 25 |
| | West | 10 | 37 | 10 | 36 | 20 | 36 | 20 |

*Two panelists in grade 12 elected not to identify their ethnicity.

## Standard-Setting Process

As mentioned earlier, the standard-setting process was an implementation of the BoW method. The process was implemented in a technologically-assisted form using the BoWTIE software.

### Panelist Training

Advanced materials were sent to the panelists. These materials included the 2011 NAEP Writing Framework, ALDs, meeting agenda, and briefing booklet. The briefing booklet described step-by-step the BoW process as it related to the NAEP writing assessment. The panelists also received on-site training during the operational ALS meeting. This training was consistent with the training provided to panelists for the 1998 NAEP ALS meetings, with modifications made to provide training specific to the BoW standard-setting method. BoWTIE training was included in each aspect of the implementation process. Panelist training occurred in both whole-group sessions and grade-group sessions.

### *Body of Work Classification and Feedback*

The BoW method belongs to the holistic family of standard-setting methods in which the panelist task consists of reviewing a series of examinee work samples, or bodies of work (BoWs), and assigning each sample to one of several performance categories (Hambleton & Pitoniak, 2006). The BoW method (Kingston, Kahl, Sweeney, & Bay, 2001) is the method generally deemed most appropriate for writing assessments, as it was developed specifically for use with performance assessments that are designed to measure student achievement using open-response items (Kahl, Crockett, DePascale, & Rindfleisch, 1995). In a traditional implementation, there are three rounds in which panelists classify BoWs. Each round is followed by a process evaluation and presentation of feedback based on the classification round. For the classification tasks, each panelist assigns each BoW to an achievement level based on his or her understanding of the ALDs and the knowledge, skills, and abilities (KSAs) being demonstrated. Cut scores are calculated from these panelist classifications. Two panels were formed for each grade level so that the degree of consistency across panels could be determined for evaluation purposes. Details about the calculation of cut scores are contained in later sections of this report as well as in the Technical Report[1].

[1] *Developing Achievement Levels on the National Assessment of Educational Progress for Writing Grades 8 and 12 in 2011: Technical Report* (Bay, 2012)

### Round 1

During Round 1, panelists examined 50 BoWs that were distributed across the full score range. Each panelist's classifications of all 50 BoWs were used to compute that panelist's Basic, Proficient, and Advanced cut scores. The BoWs were presented to panelists in order of highest to lowest performance based on their *expected a posteriori* (EAP) scores. Details on the computation of EAP scores are provided in the Technical Report. Presentation of the student BoWs, note-taking by panelists, and the recording of classifications were all facilitated through the BoWTIE system.

Feedback was presented to panelists at the end of each round. Each panel was notified of the results of the other grade-level panel. The purpose of the feedback was to inform the panelists' second round of classifications. Additionally, BoWs with a diversity of ratings were selected for whole-group discussion to help promote a common understanding for the application of ALDs to BoW classifications.

### Round 2

During Round 2 of classifications, panelists were presented with the same set of student work, along with the classifications and comments they provided in Round 1. Their task was to provide an achievement level classification for each BoW in light of feedback from the first round of classifications. Panelists were told that they could change all, some, or none of their Round 1 classifications of student BoWs. They were reminded that their classifications should ultimately be based on the match between the ALDs and the KSAs demonstrated in each BoW.

Panelist classifications from Round 2 yielded new cut scores, which were used to produce new feedback that was again presented to the whole group and discussed.

Round 3

To set the final cut scores, panelists classified a new set of 50 BoWs in Round 3. This new sample was selected using the same methodology employed to select the first set of BoWs. Although Round 3 in a BoW standard setting is typically designed to be a “pinpointing” round, results from the pilot study indicated that a third “rangefinding” round would be more appropriate. Details about this can be found in Chapter 4. Results from this process were then used in calculating the final cut scores.

### *Selection of Exemplar Performance*

Exemplar performances were identified and delivered as one of the products of the ALS process. BoWs were selected from the NAEP writing form consisting of two tasks that had been identified for public release, and these were presented to the panelists for evaluation in BoWTIE. The panelists rated each BoW with regard to its representation of what students know and can do at the achievement level to which its score corresponded. One BoW was selected for each grade and for each achievement level based on the panelists’ ratings and comments.

### *Process Evaluations*

At the end of the first day and after each round of classifications, panelists were administered an evaluation form designed to assess their understanding of instructions, tasks, and materials. These questionnaires were delivered through BoWTIE and responses were saved directly to a database. The evaluations were reviewed daily, and any sources of confusion, dissatisfaction, or other concerns were then addressed with individual panelists or the panel as a whole. These responses were further used as evidence of procedural validity.

### Standard-Setting Outcomes

There are three major components to NAEP achievement levels: (a) the ALDs, (b) cut scores for the achievement levels, and (c) exemplar student responses considered illustrative of performance at each achievement level. Each component is described in the following subsections.

#### Achievement Levels Descriptions

ALDs had been established for the Governing Board by another contractor, and these were given provisional approval by the Committee on Standards, Design and Methodology for use by Measured Progress in the ALS process. Although these ALDs had been provisionally approved by the Governing Board, the results of the pilot study and the special study, along with panelist debriefings, led to a decision to revise the ALDs. The revised ALDs were then evaluated as part of field trial 2 and accepted for use in the operational ALS meeting.

#### Cut Scores

Cut scores, provided on the NAEP scale, and the percentage of students scoring at or above each achievement level are provided for each round in Table 2. The Round 3 values represent the recommendations presented to and approved by the Governing Board. Between 81% and 93% of panelists indicated that the percentages at or above each achievement level reflected their expectations. Less than a third indicated that they would change one or more cut scores if they could.

*Table 2: Cut Scores and Percentages of Students Scoring At or Above Each*

| Grade | Achievement Level | Round 1 | | Round 2 | | Round 3 | |
|---|---|---|---|---|---|---|---|
| | | Cut Score | % At or Above | Cut Score | % At or Above | Cut Score | % At or Above |
| 8 | Basic | 120 | 80.36 | 120 | 80.60 | 120 | 80.37 |
| | Proficient | 171 | 28.30 | 174 | 25.60 | 173 | 26.77 |
| | Advanced | 216 | 1.89 | 220 | 1.34 | 211 | 3.01 |
| 12 | Basic | 120 | 80.26 | 122 | 79.06 | 122 | 79.05 |
| | Proficient | 170 | 29.81 | 167 | 32.73 | 173 | 26.83 |
| | Advanced | 214 | 2.31 | 213 | 2.59 | 210 | 3.24 |

### Exemplar Responses

Exemplar BoWs were selected after the third round of classifications. Based on the Round 3 cut scores, 16 student BoWs (eight from the set of BoWs used for Rounds 1 and 2, and eight from the set of BoWs used for Round 3) from the NAEP form with two marked-for-release tasks were classified into achievement levels. For each grade, two BoWs were classified as potential exemplar responses for Advanced, four for Proficient, and six for Basic. Panelists were asked to judge whether each BoW was illustrative of performance at the achievement level in which it was classified. They were asked to rate each BoW as “Very Good,” “Okay,” or “Do Not Use.” They were also asked to comment on their judgments, especially if they rated a BoW “Do Not Use.” Based on discussion with the TACSS, the following three criteria were used in the selection of one exemplar BoW for each achievement level at each grade:

- At least 50% of the panelists rated it as “Very Good.”
- Not more than three panelists rated it as “Do Not Use.”

- Amount of support or opposition evidenced in panelist comments on the BoW.

### Procedural and Internal Validity

Measured Progress performed tasks to support the procedural validity of the ALS. These tasks fell into three major categories:

- providing documentation of and orientation to the ALS procedure
- providing support to help panelists understand their tasks
- evaluating whether procedures were executed as intended

A series of five process evaluations were conducted during the operational ALS meeting to gather evidence to support procedural validity. These process evaluations were aimed at describing panelists' understanding of process activities, materials, and instructions. In general the results from the surveys confirm the procedural validity, with average scores on the Likert-type scales being above the mid-point (i.e., greater than 3 on a 5-point scale).

The design of the operational ALS meeting also allowed examination of the internal validity of the cut scores. If the three rounds and feedback were functioning as intended, cut score variability should have decreased across rounds (Reckase, 2012). That variability did decrease from Round 1 to Round 3. This result supports internal validity of the cut scores by showing that the process resulted in less variability among panelists' cut scores by Round 3.

### Recommendations

With the support of the TACSS, the achievement levels determined by the ALS process were recommended to the Governing Board for reporting the results of the 2011 NAEP writing in grades 8 and 12. The recommended achievement levels have three parts: (a) the ALDs, (b) the cut scores, and (c) the exemplar responses. Recommendations are also being made in two additional areas to help improve future NAEP standard settings.

### *Recruiting Procedures*

Recruitment may be improved if the Governing Board adopts the following recommendations. First, allow the use of multiple panelists from a single nominator, provided the qualifications and NAEP diversity criteria are met. Second, allow use of at-large nominations, provided the qualifications and NAEP diversity criteria are met.

### *Achievement Levels–Setting Procedures*

Two procedural recommendations are being made. One is a global recommendation, while the other is smaller in scope. Because the current standard setting showed the increases in efficiency gained through the use of BoWTIE, the first recommendation is to use similar technology-assisted approaches in future standard settings.

The second recommendation, which is specific to the BoW methodology, is to consider the use of other, more appropriate approaches to computing cut scores. The Governing Board may wish to request that future standard settings use a generalized linear mixed model to obtain the aggregate cut scores from the panelists' individual classifications.

# Chapter 1—Introduction

The Governing Board sets policy regarding the NAEP. The activity for which the Governing Board is perhaps best known is that of setting achievement levels for NAEP. Congress charged the Governing Board with this role and with the responsibility of showing the achievement levels to be useful, reasonable, and valid. The Governing Board states that:

The purpose for developing student performance levels on NAEP is to clarify for all readers and users of NAEP data that these are expectations which stipulate *what students should know and should be able to do* at each grade level and in each content area measured by NAEP.

Governing Board policy[2] specifies the establishment of three achievement levels:

The levels-setting process shall produce three threshold points for each content area and at each grade level assessed, demarcating entry into three categories: Basic, Proficient, and Advanced. These levels are defined as:

**Basic**: This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.

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

**Advanced:** This level signifies superior performance beyond proficient.

The policy also specifies the following:

[2] National Assessment Governing Board (1990). *Developing Student Performance Levels on the National Assessment of Educational Progress* (amended March 4, 1995, http://www.nagb.org/policies/plindex.htm )

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

## 1.1 The Governing Board and the NAEP Achievement Levels for Writing

The Governing Board has completed a comprehensive process for setting achievement levels on the 1992 NAEP assessments in mathematics and reading, the 1994 assessments in U.S. history and geography, the 1996 assessment in science, the 1998 assessments in civics and writing, the 2005 assessment in 12th-grade mathematics, the 2006 assessment in 12th-grade economics, and the 2009 assessment in science. These levels have been used in reporting NAEP results for the subjects in subsequent assessment years (i.e., in 1994, 1996, 2000, 2004, 2006, 2007, 2010 Report Cards) (National Assessment Governing Board, 2010).

The contract for this work was awarded on September 23, 2010, and Measured Progress began working with the Governing Board in October 2010 to develop an ALS process consistent with the Governing Board's charge and policies. The design and implementation of the achievement levels–setting (ALS) process followed the basic tenets and procedures in the Governing Board's policy on achievement levels. Additionally, the standard-setting method selected and the computer-based implementation of the ALS process reflected two important characteristics of the writing assessment: (a) the assessment consisted entirely of constructed-response items and (b) it was the first computer-based NAEP assessment ever implemented.

Measured Progress implemented the Body of Work (BoW) method for setting the NAEP writing achievement levels. The BoW method (Kingston, Kahl, Sweeney, & Bay, 2001) is the flagship standard-setting method for Measured Progress, and this method is deemed most appropriate for writing assessments because it was developed specifically for use with performance assessments that are designed to allow for a range of student responses. Using the BoW method, ALS panelists examined student responses to two writing prompts. Panelists set achievement level cut scores by examining a student's writing holistically and classifying it into one of the performance levels—Basic, Proficient, or Advanced—or, below the Basic level. As planned, Measured Progress implemented a technologically enhanced version of the BoW method with the use of the Body of Work Technological Integration and Enhancements (BoWTIE) software. BoWTIE, which allowed a fully computer-based implementation of the BoW method, was developed for the following reasons:

- to overcome the logistical difficulties in materials preparation
- to enhance the security of materials
- to promote "green" procedures, minimizing the need for hard-copy materials
- to enhance the overall efficiency and effectiveness of the process

Computerization for efficiency was recommended by the Governing Board in the statement of work.

During the ALS process, each panelist used two laptop computers provided through the NAEP program: (a) a NAEP laptop computer for viewing the NAEP writing tasks, which was intended to replicate the students' experience of taking the NAEP, and (b) a NAEP laptop configured for BoWTIE and used for all other aspects of the process. A small scale field trial was implemented for the purpose of testing the logistics of using an entirely computer-based system by implementing selected parts of the process designed for the operational ALS meeting. Prior to the ALS meetings, a complete pilot study was implemented to analyze procedures intended for operational implementation. Because the pilot study resulted in changes to the ALDs, a second field trial was required before the operational ALS meeting took place. Two special studies were implemented—one at the end of the pilot study and one at the end of the operational ALS meeting—to provide information on the relationship between performance on the new 2011 writing NAEP and performance on the 2007 writing NAEP for grade 8.

Consistent with its *Policy Statement*[3], the Governing Board shall establish the achievement levels to define what students should know and be able to do in writing at grades 8 and 12. The achievement levels will be used for reporting NAEP results to the American public. In arriving at a policy decision on writing achievement levels, the Governing Board will be informed by the recommendations for the achievement levels produced from the work described herein.

The ALS project involved several standard-setting meetings, each contributing to modifications implemented in subsequent meetings. The first meeting, the field trial, was implemented to test the logistics of requiring panelists to use two laptop computers. After the field trial, the process intended to be used to set achievement levels was tested in a pilot study. The operational ALS meeting yielded the results that are being recommended to the Governing Board. A special study was implemented at the end of the pilot study and at the end of the operational ALS meeting to explore the relationship between performances on the 2011 assessment, based on the new writing

[3] *Ibid.*

framework, and performance on the 2007 assessment, based on the writing framework first implemented in the 1998 assessment. Table 3 summarizes the purpose of each meeting. A copy of the agenda for each of these meetings is included in Appendix A. The agenda for each special study implementation is included in the pilot study or operational ALS meeting agenda.

*Table 3: Achievement Levels–Setting (ALS) Meetings*

| **Meeting** | **Primary Purpose** | **Date** | **Venue** |
|---|---|---|---|
| Field Trial | To test the logistics involved in using two laptop computers | September 22–23, 2011 | Portsmouth, NH |
| Pilot Study | To implement t the intended process for the operational meeting | November 15–18, 2011 | St. Louis, MO |
| Special Study | To compare performance on the 2007 and 2011 assessments | November 18–19, 2012 | St. Louis, MO |
| Field Trial 2 | To test the implementation of modifications[4] based on pilot study findings | January 27, 2012 | Dover, NH |
| Operational ALS Meeting | To set achievement levels that will be recommended for consideration of the Governing Board | February 7–10, 2012 | St. Louis, MO |
| Special Study 2 | To compare performance on the 2007 and 2011 assessments | February 10–11, 2012 | St. Louis, MO |

## 1.2 Purpose of This Document

This report provides a detailed description of the ALS process and the outcomes of a meeting held February 7–10, 2012, to set achievement levels for the 2011 NAEP writing in grades 8 and 12. It also summarizes project activities that preceded the ALS meeting, including the adaptation of the BoW standard-setting method based on NAEP ALS tradition and the development of the BoWTIE software for a fully computerized ALS process. The operational ALS meeting was preceded by a pilot study held November 15–18, 2011, which was preceded by a field trial on September 23–24, 2011.

[4] The two modifications that were implemented in Field Trial 2 were the revised ALDs and the inclusion of the response classification exercise in the panelist training.

Additionally, findings from the pilot study prompted a second field trial, held on January 27, 2012.

This document serves as the primary source of information for all aspects of the ALS process implemented by Measured Progress that resulted in a recommendation to the Governing Board of cut scores used to differentiate student performances at the Basic, Proficient, and Advanced levels of achievement in the 2011 NAEP writing for grades 8 and 12. Each activity leading to the operational ALS meeting is described in detail, including intermediate results leading to process modification relative to what was described in the Design Document (Measured Progress, 2011). The recruitment process, which ensures broad participation of different types of panelists as prescribed by the Governing Board, is detailed in this document. This document also reports the recommended achievement levels—composed of the ALDs, the cut scores delineating the achievement levels on the score scale, and exemplar student BoWs illustrative of performance at each level of achievement—as well as supporting details intended to provide evidence that the achievement levels for the 2011 NAEP grades 8 and 12 writing are useful, reasonable, and valid.

## 1.3 Organization of This Document

The following six chapters represent the main body of this report. Chapter 2, “Project Overview,” discusses the people involved in the project and presents a detailed description of the ALS process, including recruitment of panelists and the development of the ALDs. Section 2.9, Achievement Levels–Setting (ALS) Process, describes the process used in the operational ALS meeting, which was based on findings from the meetings that led up to the operational meeting. Differences in meeting implementation are addressed, as appropriate, in either section 2.9 or in the specific chapters for the different meetings.

Chapters 3 through 5 discuss each of the meetings leading up to the operational ALS meeting. The process implementation in each of the meetings builds upon findings from the previous meetings. Chapter 6, “Operational Achievement Levels-Setting Meeting,” is organized into six sections and discusses the selection of exemplar responses and validity evidence in addition to panelists, process, and results.

Chapter 7 presents recommendations to the Governing Board. This chapter includes the recommended achievement levels and other recommendations on recruitment, ALS procedures, and investigation of validity evidence.

# Chapter 2—Project Overview

This chapter provides an overview of the project that includes a description of the process that was implemented operationally. Measured Progress conducted an ALS process which produced a set of recommendations for the Governing Board to consider when establishing achievement levels on the 2011 NAEP for writing grades 8 and 12. All aspects of the ALS process were established with guidance from the Governing Board's Contracting Officer's Representative (COR), Dr. Susan Cooper Loomis and the Committee on Standards, Design and Methodology, advice from the Technical Advisory Committee on Standard Setting (TACSS), and input from relevant stakeholders.

## 2.1 Technical Advisory Committee on Standard Setting

Per the Governing Board's requirements, Measured Progress appointed an external Technical Advisory Committee on Standard Setting (TACSS), which was consulted in all aspects of the project. The TACSS is a six-member group of well-respected measurement experts with national or international reputations. Collectively, TACSS members have expertise in large-scale assessment standard setting with prior experience in NAEP achievement levels setting. At least one member of the TACSS was a state testing director, and one seat on the TACSS was designated for a representative from the Design, Analysis, and Reporting contractor for NAEP, the Educational Testing Service (ETS). Below are the names and affiliation of TACSS members.

- Dr. Bill Auty
  Consultant (Former Assistant Superintendent, Oregon Department of Education)
- Dr. Wayne Camara
  Executive Vice President, Research and Development, The College Board

- Dr. Barbara Dodd
  Professor of Educational Psychology and Quantitative Methods, University of Texas – Austin
- Dr. Matthew Johnson
  Associate Professor of Statistics and Education, Teachers College, Columbia University
- Dr. Mary Pitoniak
  Strategic Advisor for Statistical Analysis, Data Analysis and Psychometric Research, ETS representative
- Dr. Mark Reckase
  University Distinguished Professor of Measurement and Quantitative Methods, Michigan State University

The TACSS met nine times over the course of the project and provided input on key components of the project, including modifications to the Body of Work (BoW) standard-setting method; the design of field trials and special studies; the conduct of the ALS meeting; data analysis procedures, including computation of cut scores; and the formulation of conclusions and recommendations presented to the Governing Board. Seven TACSS meetings were held in-person while two were held online via WebEx. Summaries of the TACSS meetings, which include details of their recommendations as well as discussions that led to the recommendations, are included as an appendix to the Technical Report[5]. In addition to consultation, some members of the TACSS attended the field trials, the pilot study, and the operational ALS meeting as observers. A representative from the National Center for Education Statistics (NCES) who attended most of the TACSS meetings also attended the pilot study as an observer.

[5] *Developing Achievement Levels on the National Assessment of Educational Progress for Writing Grades 8 and 12 in 2011: Technical Report* (Bay, 2012)

## 2.2 Technical Assistance from the NAEP Alliance

Some materials, data, and equipment necessary for the implementation of the ALS process were provided by the National Center of Education Statistics (NCES). The NAEP Alliance member companies, and the assistance that each provided, are listed below:

- Educational Testing Service
  - student-level data, including raw scores and plausible values
  - frequency distribution of the first plausible values
  - task-level data, including item response theory (IRT) parameters
  - a representative to TACSS to provide on-going technical advice
- Pearson
  - PDF copies of student responses
  - ancillary materials
  - scoring guide
- Westat
  - computer-based assessment (CBA) laptops
  - ALS laptops
- Fulcrum IT
  - software modifications for administering NAEP to panelists
  - software modifications for accessing panelists' responses
  - software modifications for reviewing all writing tasks

The equipment and software modifications provided by Westat and Fulcrum IT, respectively, are described in detail in various sections of this report[6].

[6] Sections 2.9.8.2, 2.8.9.3, 3.2.1, and 3.3.1.

Requests for materials and equipment necessary for the implementation of the ALS meetings were discussed during a monthly online meeting with Alliance members. The goals of the meetings included the following:

- continuing conversations regarding requests and enhancing Measured Progress's understanding of NAEP
- following up on requests made during the last meeting
- updating request time lines regarding deliverables
- clarifying and confirming Measured Progress's understanding of NAEP data

Formal requests to NCES for equipment and materials were made through the Governing Board's COR; no requests were made directly by Measured Progress to the Alliance partners. Interim meetings were scheduled as needed.

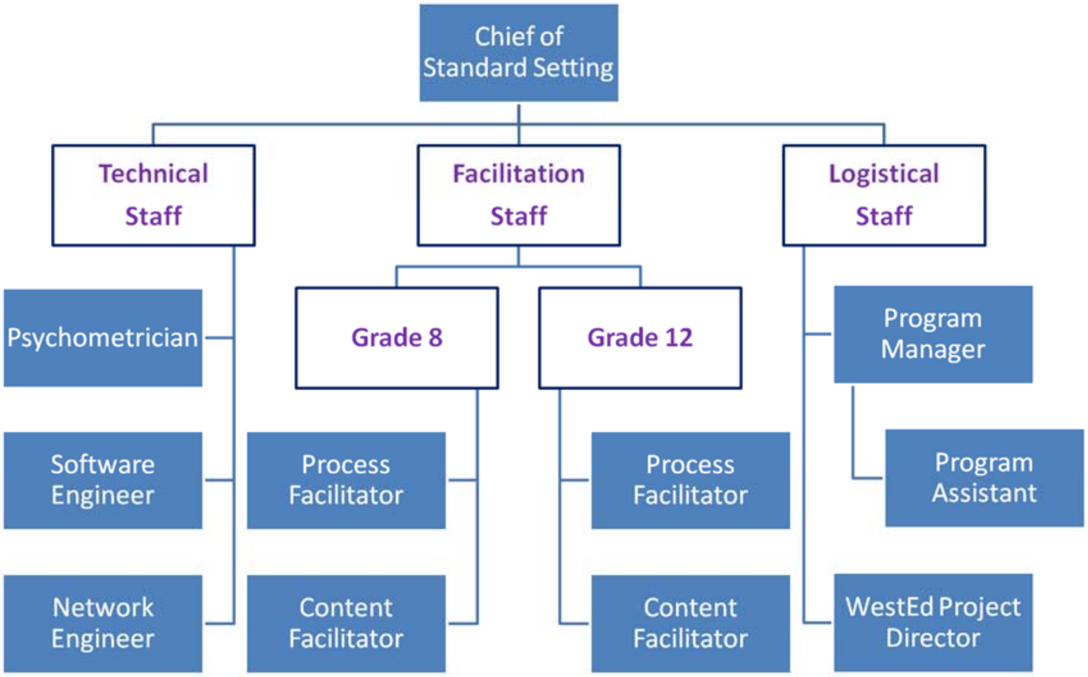
## 2.3 Project Staff

Setting standards for our nation's youth is an extremely involved endeavor that requires staffing to match the importance and high-profile nature of the project. As the Project Director, Dr. Luz Bay provided intellectual leadership to the project and she led the Measured Progress team of program management staff, psychometricians, and data analysts in all aspects of preparation for standard-setting meetings, meeting logistics and implementation, and reporting results. In addition to consulting with the TACSS, Dr. Bay also consulted with an internal Technical Advisory Group (TAG), which includes two of the original authors of the BoW method. Measured Progress named WestEd as the subcontractor to collect public comments and coordinate hotel logistics for the panel meetings. For the development of the Body of Work Technological Integration and Enhancements (BoWTIE) software, Dr. Bay served as the Product Owner. As such, Dr. Bay worked closely with the information technology staff to assure that expertise in the proposed process was integrated into the development of the software and its delivery.

## 2.4 Achievement Levels–Setting Staff

ALS meetings require staffing to cover three different aspects of meeting implementation: technical, facilitation, and logistics. Figure 1 presents the organizational structure for meeting staffing. The Measured Progress staff members who worked on the project in terms of preparation for the meetings also travelled to St. Louis for the ALS meetings. The ALS project director, Dr. Luz Bay, served as the Chief of Standard Setting (CoSS) and had overall responsibility for all aspects of process implementation in all ALS meetings. The support staff included a psychometrician, software and hardware support, several persons to assist with logistics operations at each meeting, and the WestEd project director, who served as site liaison for the pilot study and operational ALS meeting.

*Figure 1: ALS Meeting Organizational Chart*



### *2.4.1 Process Facilitators*

"The role of the facilitator is to structure the conversations, monitor the discussions, and generally make certain that the intended methods are being followed" (Skorupski, 2012). Although the CoSS was in charge of the overall facilitation of ALS meetings, the process implementation occurred mostly in the grade-level breakout rooms. This structure means that the role of the process facilitator, the person primarily in charge of providing information, giving direction, ensuring that all panelists understand what to do and how to do it, and maintaining the agreed-upon schedule is vital to the success of any standard-setting process. Most importantly, the process facilitators were responsible for ensuring that the ALS process was executed according to the study design. Dr. Joseph St. George, a program manager at Measured

Progress, served as the process facilitator for grade 8. Dr. Phil Robakiewicz, director of Client Services at Measured Progress, served as the process facilitator for grade 12.

### 2.4.2 Content Facilitators

The content facilitator takes the lead for parts of the process for which subject-matter knowledge and understanding of the content of the assessment is imperative. For the NAEP ALS project, subject matter knowledge of writing was not sufficient for the role of content facilitator. In-depth knowledge of the NAEP writing assessment was essential. Dr. Carol Jago served as the content facilitator for the first field trial. Ms. Pat Porter and Dr. George Kamberelis served as the content facilitators for the remaining panel meetings in the project. All three content facilitators were members of the 2011 NAEP Writing Framework Steering Committee. Their membership in the Framework Steering Committee provided the basis for their deep level of familiarity with the NAEP writing assessment and the rationale behind different aspects of the assessment. Ms. Porter and Dr. Kamberelis assisted in finalization of the achievement levels descriptions (ALDs), and Dr. Jago provided input on that process.

## 2.5 Panelist Nomination and Selection

The ALS process was conducted by using the informed judgments of well-qualified and broadly representative panels to recommend achievement level cut scores consistent with the Basic, Proficient, and Advanced ALDs and to identify exemplar performance for each level. The following section describes how panelist nomination and selection was accomplished.

### 2.5.1 Multistage Selection Process

As specified in the Design Document, Measured Progress implemented a multistage process for the recruitment of panelists for the ALS process. This process consisted of four stages:

- Stage 1: Select districts and identify nominators
- Stage 2: Contact nominators and request nominations
- Stage 3: Notify nominees and request acceptance of nomination
- Stage 4: Select and recruit panelists

In brief, this selection process commenced with a sampling of school districts from which nominators were selected and invited to submit up to four nominations for panelists. The nominees were then notified of their nomination and asked to submit their credentials for consideration. Based on their qualifications and the demographic criteria set forth by the Governing Board, panelists were selected from this pool of nominees. The databases, sampling variables, nominator types, and panelist classification targets are described in the sections that follow.

#### *2.5.1.1 Stage 1: Select Districts and Identify Nominators*

Districts were the primary sampling unit for the sampling design. Identified using the 2008–2009 Common Core of Data (CCD), only districts with students at grades 8 and above (15,468 of the 18,350) were considered (U.S. Department of Education [U.S. DOE], National Center for Education Statistics [NCES], & Institute of Education Sciences [IES], 2010a). Recruiting 91 panelists for the pilot study and operational meeting required identifying a much larger sample of nominators to account for nonresponsive nominators, unqualified nominees, and nominees who either did not accept their nominations or had to withdraw after accepting their nominations. Based on response rates reported in previous panelist recruitment efforts for the NAEP (ACT, 2007), and particularly the recent trend of decreasing response rates for nominated teachers, the anticipated response rates used in the sampling plan varied according to the type of panelist being recruited. For all panelist types, it was assumed that 30% of the nominators contacted would respond with at least one nomination (a conservative estimate of only one nominee was expected). Given the historical difficulty in recruiting panelists from private institutions, an 80:20 ratio of public school districts to private schools was used for selecting nominators in pursuit of the goal of a 90:10 ratio for panelists.

In light of these assumptions and in pursuit of the original goal of 100 panelists to produce an adequate overage, 2,474 nominators were selected with the goal of yielding 110 panelists. District sampling without replacement was performed for the three panelist types: teacher, nonteacher educator, and general public. Included in the district sampling was a sampling of private schools (through Private School Universe Survey) and postsecondary institutions (through College Navigator), which was performed as the first step before identifying the respective nominators. See Table 4 for a description of the nominators.

*Table 4: Description of Nominators for Each Panelist Type*

| Panelist Type | Sampling Unit | Nominators |
|---|---|---|
| Teacher | Public School District | Superintendent |
| | | Principal |
| | | School Board President |
| | | Head of Teacher Organization |
| | | President of Parent Teacher Organization (PTO) |
| | Private School | Principal |
| Nonteacher Educator | Public School District | Superintendent |
| | | Principal |
| | | School Board President |
| | | State Curriculum Specialist |
| | Private School | Principal |
| | Postsecondary Institution | Chair of Appropriate Academic Departments (e.g., School of Journalism, Interdisciplinary Studies in Writing) |
| | | Director of Writing Center, Writing Fellows Program |
| General Public | Public School District | Mayor |
| | | City or Town Manager |
| | | Education Committee Chair of the Chamber of Commerce |
| | | Editor-in-Chief of the Local Newspaper |
| | | Librarian |
| | | Member of the School Board |
| | | Department of Human Resources and Directors for Corporations |
| | | Author |

Public school districts were selected from the CCD to proportionately represent the four NAEP regions, socioeconomic status (SES), and urban and rural demographics, as set forth by the Governing Board. The representation by NAEP region is based on the number of districts in each of the four regions. The selected districts included a representative sample of districts with low SES as indicated by the percentage of students who participated in the National School Lunch Program in the district (U.S. DOE, NCES, & IES, 2010b, 2010c). This demographic classification of selected public school districts is presented below in Table 5.

*Table 5: Demographic Classification of Selected Public School Districts From CCD*

| **Demographic Variable** | **Attribute** | **Percentage** |
|---|---|---|
| NAEP Region[1] | Midwest | 37 |
| | Northeast | 19 |
| | South | 23 |
| | West | 21 |
| Socioeconomic Status (SES)[2] | Low SES | 22 |
| | Not Low SES | 78 |
| Urbanicity[3] | Large City | 17 |
| | Large Suburb | 32 |
| | Rural | 20 |
| | Other | 30 |

[1] U.S. DOE, NCES, & IES, 2011
[2] U.S. Census Bureau, 2011
[3] U.S. DOE, NCES, & IES, 2010b

The top 22% of districts, ranked from high to low, based on the percentage of students enrolled in the National School Lunch Program, were taken as the target for districts with low SES. The sampling of districts also targeted 17% of large city districts to ensure their representation in the sample. In addition, 32% of the districts were selected to represent the proportion of students educated in districts categorized as large suburbs, 20% to represent the percentage of students schooled in rural districts, and 30% to represent all other urban areas.

### Identification of Nominators Using District Sample

Nominators were identified from public school districts and private schools to solicit teacher nominees from which 55% of the panel would be selected. Multiple teacher nominators were identified whenever possible in each of the 525 public school districts. Private school teacher nominators were identified from a sampling of 444 private schools in the Private School Universe Survey. The principals (also titled chancellors or headmasters) for each of these schools were identified as teacher nominators. Nominator names and contact information were then gathered through Internet and telephone research, based on the district sample and predefined qualifications. This resulted in a total of 1,612 teacher nominators. The ratio of teacher nominators from public districts to those from private schools was 72:28, representing a higher proportion of private school nominators than the originally designed ratio of 80:20. This design was established in order to ultimately achieve a 90:10 public-to-private publicity ratio among teacher panelists.

A total of 404 nominators were identified through Internet and telephone research to meet the target of selecting 15% of panelists who are nonteacher educators. These nominators were made up of (a) 133 nominators, including superintendents, principals, school board presidents, heads of teacher organizations, and presidents of PTOs, from 34 public school districts; (b) 18 nominators (principals, chancellors, and heads of schools) from 22 private schools: and (c) 253 nominators (e.g., department chairs) from 229 postsecondary institutions.

A total of 456 districts were identified to meet the goal of selecting 30% of the panelists who are not educators (“general public”). Internet and telephone research was conducted from towns in the same locality to identify nominators from the chairs of the education committee of the Chamber of Commerce, mayors and/or city managers, editors-in-chief of the local newspapers, librarians, members of school boards, and directors of human resources departments or divisional directors at large corporations.

As nominator identification began, it became evident that a larger sample of districts would be required to meet the goal number of nominators. A supplemental sampling of 346 districts (total) was added to the original sample of 1,364 districts (total), using identical criteria. In total, these efforts yielded 2,474 nominators from a total of 1,710 districts.

The nominators' names, contact information, and demographic data were then uploaded to BoWTIE's recruitment module, through which the nominators were contacted, their activity tracked, and their submissions gathered. The recruitment module maintained the relationship between each nominator and his or her nominees.

#### *2.5.1.2 Stage 2: Contact Nominators and Request Nominations*

Having identified nominators for each panelist type—teacher, nonteacher educator, and general public—program management utilized BoWTIE's recruitment module to request via e-mail that each one nominate up to four persons of the appropriate panelist type for the appropriate grade level(s). The text and links in the e-mail provided the qualifications for nominees and a preliminary description of the panelists' task, travel, and reimbursement. The nominators were instructed to submit their nominations using an electronic form that they could access by clicking a hyperlink in the e-mail. The hyperlink took the nominators to a login page for the recruitment module, which they accessed using the personalized login credentials also provided in the e-mail. A sample of the teacher nominator version of this e-mail is contained in Appendix B; this e-mail was modified as needed for nominations of nonteacher educators and the general public. All information and materials used in panelist recruitment are in Appendix B.

The recruitment module then presented the nominators with the option to submit their nominations and/or to nominate themselves (see screenshot in Appendix B). While submitting their nominations, nominators could also submit basic information about their nominees, including contact information and qualifications. The recruitment module's nomination form may also be seen in Appendix B.

Nonresponsive nominators were sent scheduled follow-up e-mails through the recruitment module, including notification of the extension of the response window to encourage more nominator response. Ultimately, phone calls were made to follow up with nonresponsive nominators. When a nominator was successfully contacted, the phone call generally tended to result in nominations.

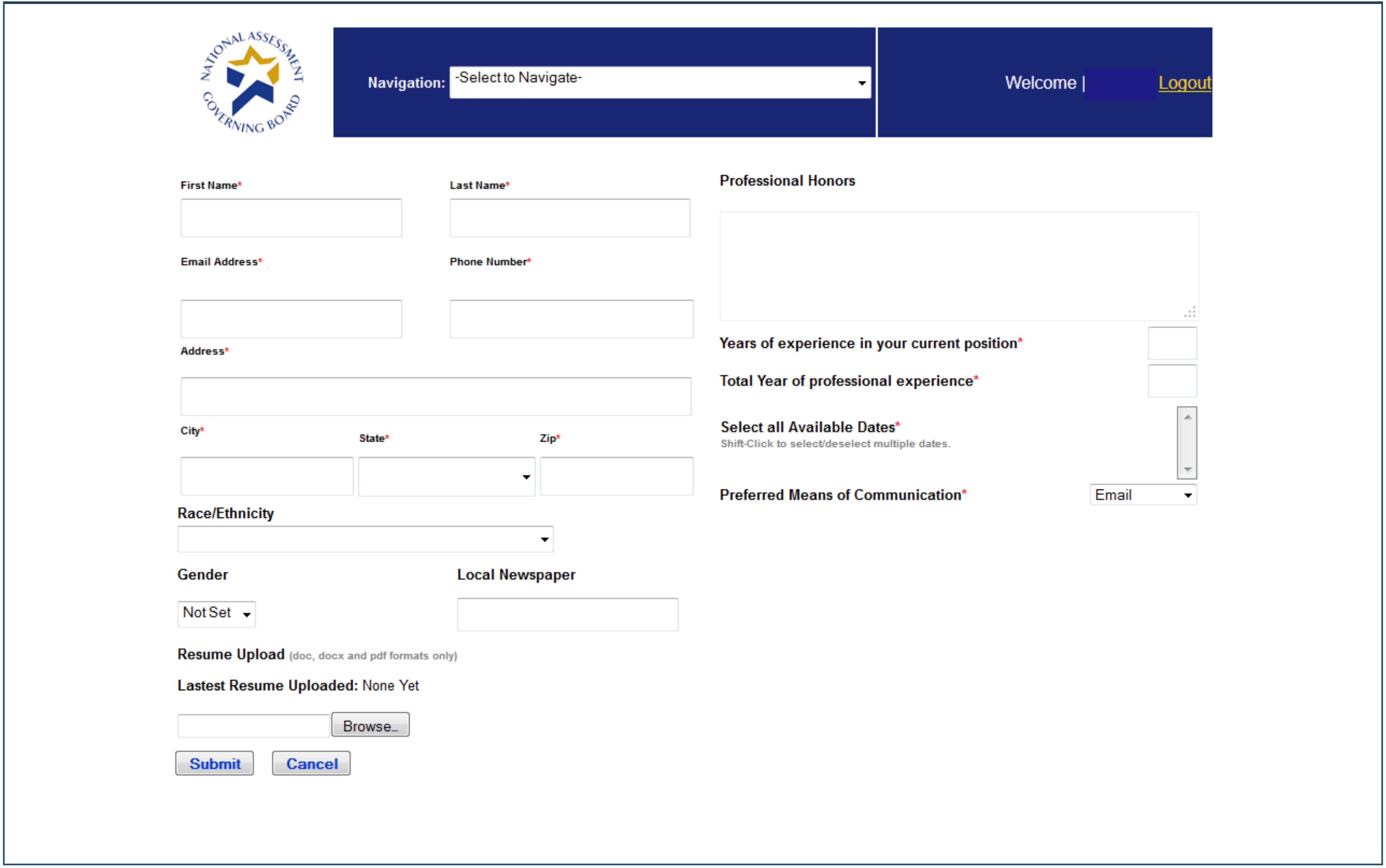
In addition to the problem of nonresponsive nominators, the difficulty of out-of-date or erroneous e-mail addresses occasionally arose. These were tracked and removed from the contact list if an updated e-mail address could not be located.

Despite these efforts to solicit nominations from the anticipated 30% of nominators, only 4% (95 of 2,474) responded with nominations for the pilot study or the operational ALS meeting. In total, these 95 nominators supplied a list of 141 nominees, who were, in turn, contacted to request further information and acknowledgment of their nomination.

#### *2.5.1.3 Stage 3: Notify Nominees and Request Acceptance of Nomination*

Once nominees were identified for each of the panelist types, a personalized e-mail was sent to each one through BoWTIE's recruitment module. This e-mail and its attachments informed them of their nomination, their nominator, and their basic role as a panelist. Additionally, the e-mail contained essential information about the accommodations, reimbursements, and netbook computer incentive for participation. The e-mail contained hyperlinks that provided the nominees the opportunity either to send an e-mail reply declining their nomination or to log in to the recruitment module and accept the nomination. Once nominees logged in with the username and password provided in the e-mail, they were presented with welcome and profile screens (Figures 2 and 3), where they were able to update their profile and credentials.

*Figure 2: Recruitment Profile Screen*



Navigation: -Select to Navigate-

Welcome | [redacted]

First Name*

Last Name*

Email Address*

Phone Number*

Address*

City*

State*

Zip*

Race/Ethnicity

Gender: Not Set

Local Newspaper

Resume Upload (doc, docx and pdf formats only)

Lastest Resume Uploaded: None Yet

Professional Honors

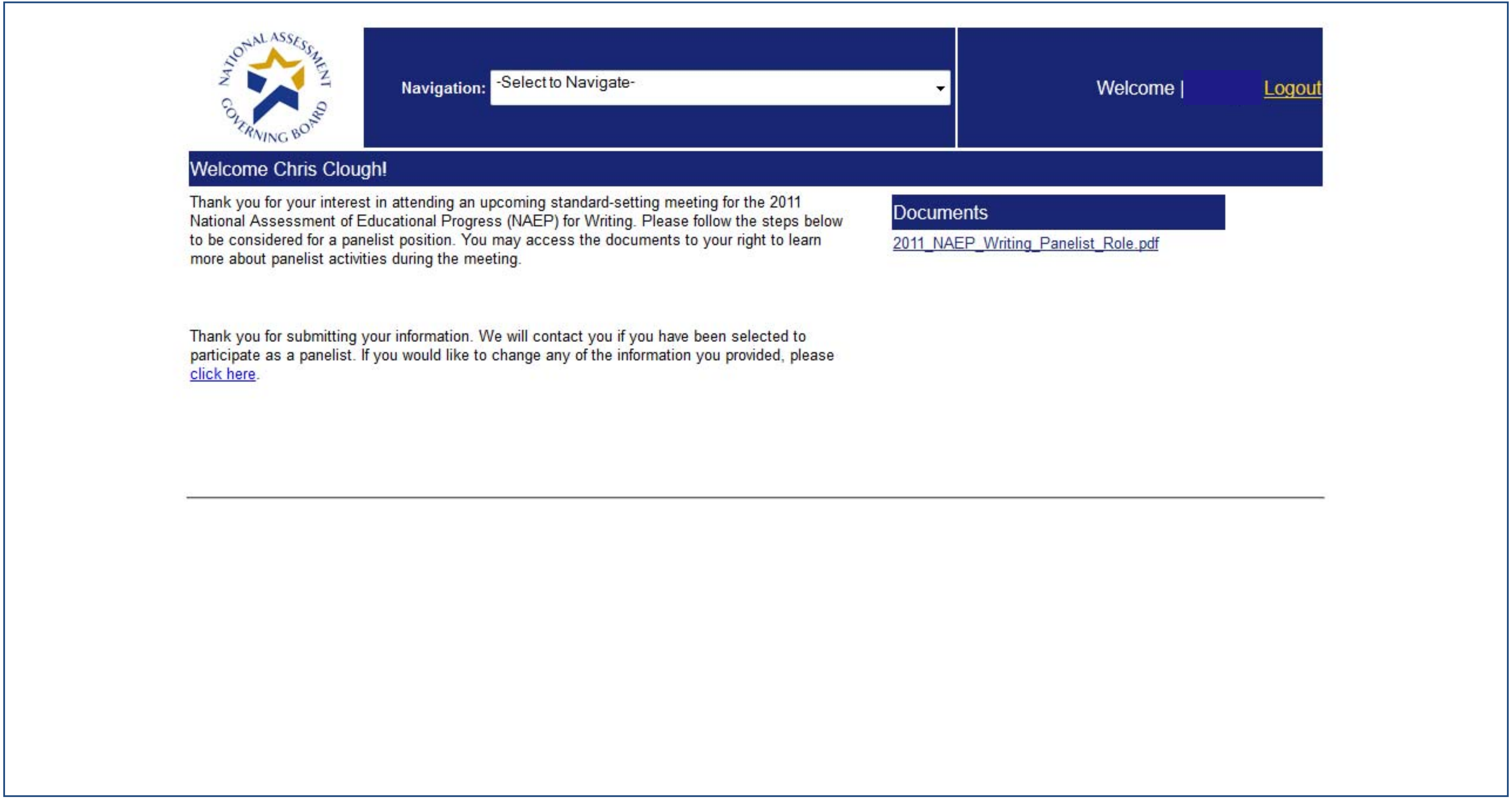
Years of experience in your current position*

Total Year of professional experience*

Select all Available Dates*
Shift-Click to select/deselect multiple dates.

Preferred Means of Communication*: Email

*Figure 3: Recruitment Welcome Screen*



Welcome Chris Clough!

Thank you for your interest in attending an upcoming standard-setting meeting for the 2011 National Assessment of Educational Progress (NAEP) for Writing. Please follow the steps below to be considered for a panelist position. You may access the documents to your right to learn more about panelist activities during the meeting.

Thank you for submitting your information. We will contact you if you have been selected to participate as a panelist. If you would like to change any of the information you provided, please click here.

Documents

2011_NAEP_Writing_Panelist_Role.pdf

#### *2.5.1.4 Stage 4: Select and Recruit Panelists*

Nominees that registered in the recruitment module were included in a recruitment report that was used to aid in the selection of panelists for each meeting. Program management reviewed the nominees' qualifications and assigned a qualification score between 1 and 5, depending on how well each nominee met the panelist eligibility criteria. Panelist selection began with those who were most qualified and who had affirmed their availability for the upcoming meeting.  Panel composition was continually monitored in order to attain the proper percentages of panelist types and demographic distribution of NAEP regions, gender, publicity, and urbanicity. While the qualifications of the nominees remained the highest priority, adjustments were made to the roster until the best possible balance of qualifications and demographic criteria fit was achieved. This means that, in some cases, qualified nominees were overlooked in favor of similarly qualified nominees who better filled the demographic distribution criteria. Finally, a suggested panel was submitted to the client for review and recommendation, which resulted in several requests for additional information from the preselected nominees or outright replacements of the nominees.

Once a satisfactory panel was proposed, program management began to recruit panelists from among the preselected nominees. These nominees were contacted with a Panelist Notification e-mail, which informed them that they had been selected as one of panelists for the upcoming meeting, and requested that they e-mail their participation confirmation to program management as soon as possible. The e-mail also provided links to the Nation's Report Card and the hotel serving as the meeting site, and explained that once confirmation was received additional information would be sent. A sample of this e-mail may be viewed in Appendix B.

In the case of the pilot study panelists, a series of e-mails including advanced materials and logistics instructions were sent in response to the panelists' confirmations. In the case of the panelists selected for the operational ALS meeting, an e-mail with advanced materials attached and a single all-inclusive logistics e-mail were sent after the panelist confirmations were received. The details of these communications are included in the Panelists section of each ALS meeting, later in this document.

Occasionally, a nominee was unresponsive to the Panelist Notification e-mail or experienced a change in availability. In these cases a suitable replacement was found to duplicate the qualifications and demographic criteria. The replacement process for the pilot study was substantially successful, while greater difficulty was experienced with the operational ALS meeting, as described later in this section.

In addition to the standard recruitment described above, the Ohio NAEP State Coordinator provided a large number of well-qualified nominations from Ohio, and this resulted in a high number of pilot study panelists from Ohio. Furthermore, content and process facilitators were asked to recommend additional general public nominations to increase the general public representation on the panel. Recruiting from the general public presented the greatest difficulty of the three panelist types throughout the recruitment process. Due to limited lead time, as well as limited contact information, these nominees were pursued via phone as well as e-mail when possible. One of these nominees was approved and, just a few days before the pilot study, agreed to participate as a panelist. Nevertheless, the goal for general public panelists was not met, and it was clear that general public recruitment efforts would need to be increased for the operational ALS meeting.

Panelist recruitment for the operational ALS meeting proved to be more challenging. Among the contributing factors were (a) low nominator response rate, (b) reduction of the nominee pool due to panelist replacement for the pilot study, (c) prevalence of availability changes (perhaps due to late notification of selection), and (d) difficulty recruiting nominees from the general public. An additional follow-up was attempted via phone with more than 500 nonresponsive nominators from the original samples. Most of these calls concluded with voicemail messages and generated very few additional nominees. As a result, it became evident that additional recruitment efforts would be required for the operational ALS meeting.

With the COR's approval to allow at-large nominations, and based on recommendations from the TACSS, a variety of additional recruitment efforts were undertaken. It should be noted that due to the unique nature of these efforts and the desire for very rapid turnaround of information, tracking of these additional campaigns was carried out in spreadsheets instead of through the recruitment module.

An effort was made to recruit from among the National Council of Teachers of English (NCTE), for which Measured Progress rented two NCTE mailing lists, Writing and Assessment, and matched them to the zip codes in the original sample. This resulted in 349 teacher nominees via at-large nomination. Because the NCTE mailing lists provide mailing addresses, not e-mail addresses, a hard-copy mailing was sent that indicated these teachers had been nominated due to their extensive knowledge of and experience with the subject area. Additional introductory details were provided in the letter (see Appendix B) and an enclosed information sheet. Finally, a hard-copy nominee information form was provided that the nominees could fill in and return to Measured Progress via mail, fax, or, if scanned, e-mail (see Appendix B). It may be worth noting that, to comply with NCTE list rental conditions, the letter itself could not reference NCTE, and a copy of the letter, along with the anticipated mail date, was submitted to NCTE in advance. This effort resulted in very few nominee responses.

Additional efforts were made to work with the State Education Editors group and the Education Writers Association. Neither effort generated responses, as the State Education Editors communicated that sponsor information is not shared publicly and, after an initial response of interest, there was no further collaboration with the Education Writers Association.

E-mails were sent to state curriculum specialists and other state education department contacts, as well as NAEP state coordinators, inviting them to nominate teachers and nonteacher educators for the ALS process. This produced approximately 10 nominees, who were then contacted as described earlier in this section.

Program management also commenced Internet and phone research to distill a list of 74 authors with the appropriate credentials to invite to participate as general public nominees. This effort relied upon state writers associations, publishing houses, and often the authors' own websites to provide the necessary information. The authors were initially sent a personalized e-mail similar to the nominee e-mail referred to in Stage 3, earlier, but customized to bypass use of the recruitment module. The PDF version of the nominee form (Appendix B) attached to the e-mail was designed as an electronic form that could be filled out by the nominees and e-mailed back to program management. A second attachment, the standard information form for general public nominees, detailing the panelist qualifications, participation specifics, billing, reimbursements, and the incentive program, was also included. Within days of sending this e-mail, program management followed up with phone calls to the authors. These calls proved fruitful, and several of these nominees were approved and selected as panelists, at which point they were informed via e-mail of their selection and asked to reply to confirm their acceptance. They were also asked at this point to indicate whether they would be willing to participate in special study 2, which was described as being likely to occur after the operational ALS meeting. After their confirmations were received, these panelists were sent the Panelist Details e-mail in Appendix B. The NAEP Panelist Details e-mail is described in detail in the Advanced Materials section of this chapter.

Finally, the Director of National Programs and Site Development at National Writing Project was contacted for help in the nomination process. Dr. Elyse Eidman-Aadahl is familiar with the NAEP ALS process because she was a grade 12 content facilitator for the 1998 NAEP ALS process for writing, and a member of the planning committee for the 2011 NAEP Writing Framework. Dr. Eidman-Aadahl asked Dr. Linda Friedrich, Director of Research and Evaluation for the National Writing Project, to nominate well-qualified teachers from various states. Dr. Friedrich submitted 42 teacher nominees, who were then contacted and invited to submit their credentials. This was initiated through e-mail, with follow-up phone calls to the nominees. Ultimately 15 of these nominees were selected as panelists, invited to confirm their acceptance of the selection through e-mail, and finally sent the Panelist Details e-mail in Appendix B. Before these 15 nominees were notified, approval was received from the COR to utilize this many nominees from a single nominator. Due to the nature of her work, Dr. Friedrich was able to nominate very qualified teachers from across the nation, thus providing assurance that the panels would still be broadly representative.

The panelist roster continued to evolve as the operational ALS meeting approached, due to these ongoing recruitment efforts, changes of availability on the part of selected panelists, and approximately six panelist cancellations received subsequent to their confirmations (due to health and personal reasons). Overall, direct telephone communication seemed to be a key aspect of eliciting responses and securing a final roster of panelists. Concomitant with the ongoing nature of panelist recruitment and replacement was the inability to provide a status notification prior to the start of the meeting to those nominees (or their nominators) that were not selected as panelists. The fluid nature of the panel composition discouraged eliminating any from the nominee pool until the final roster was confirmed. This is regretful, particularly as it communicated to some nominees that their time and responsibilities were not highly valued by program management. E-mails were sent near the end of the operational ALS meeting to the nonselected nominees and their nominators with the intent to provide a status update, to acknowledge and explain the delay in response, to extend appreciation and high regard for them and their accomplishments, and to express gratitude for their interest and flexibility.

### *2.5.2 Recruitment for Field Trial*

Panelist recruitment for the first field trial, which was the first recruitment effort for this contract, was essentially an abbreviated version of the full recruitment for the pilot study and operational ALS meeting. Notable exceptions included a limited geographic demographic of a 50-mile radius from the standard-setting site in Portsmouth, New Hampshire, and the use of only three variables for panelist selection: (a) panelist type (teacher, nonteacher educator, and members of the general public), (b) type of educational institution (public school district or private school), and (c) the qualifications of panelists. This recruitment employed the multistage process described earlier in this report.

A total of 263 nominators were selected within the 50-mile radius to recruit 20 panelists for the single grade 12 panel. These nominators included 143 teacher nominators (80 from public districts and 63 from private schools), 44 nonteacher educator nominators, and 76 general public nominators. These nominators were contacted via e-mail through the BoWTIE recruitment module; the e-mail included background information and instructions, specifying that nominators may nominate up to four qualified nominees (including themselves) per grade level for the grade(s) the nominator represented. Nominators were to make their nominations through an online form to which a hyperlink was provided, along with a unique username and password combination that gave them access to the form and personally identified them within our recruitment system. Examples of this nominator e-mail and the online nomination form are included in Appendix B.

Although 30% of the nominators were estimated to respond to the request for nominations by submitting at least one nominee for consideration, 10% (26) actually responded, resulting in a total of 46 nominees. The nominees were sent an e-mail through the recruitment module that contained an explanation of the program, the role of the panelists, and basic logistical information. Through this e-mail, the nominees were invited to submit their credentials by logging in to the BoWTIE recruitment module, utilizing the unique username/password combination provided in the e-mail. Thirty-five of these nominees accepted their nominations by responding and providing further information through the online form contained in Appendix B. Once the qualifications and demographic data of these nominees were reviewed, selections were made to fill the panel. Finally, the panelists were recruited through an e-mail that informed them of their selection, requested an e-mail reply to confirm the panelists' availability, and provided links to the Nation's Report Card for writing, the field trial site (Sheraton Portsmouth Harborside Hotel in Portsmouth, NH), and driving directions.

### 2.5.3 Recruitment for Field Trial 2

Panelist recruitment for field trial 2 was unique among recruitment efforts in this project. Because field trial 2 materialized in response to the findings of the pilot study, and because it was determined that panelists from the original field trial should not participate, to avoid any contamination in the study, a new recruitment was required within a relatively short 3-week timeframe. With the Governing Board's approval, Measured Progress performed a sampling of convenience in conjunction with a temporary employment agency, Kelly Services, to recruit 40 panelists from within a 50-mile radius of Dover, New Hampshire, to fill the two grade-level panels. Measured Progress provided Kelly Services with information relating to panelist participation, such as the panelist types and qualifications, meeting dates, and reimbursement details. Kelly Services then recruited the panelists from among their viable contacts. A draft of the Kelly Services recruitment flyer may be seen in Appendix B.

## 2.6 Advanced Materials

Advanced materials were sent to panelists one to three weeks prior to each meeting for the purpose of familiarizing them with the NAEP, the details of the meeting, and their roles in general. The various materials were sent electronically and/or on hard copy. The following were included in the advanced materials: a list of panelist qualifications; a description of panelist tasks; information about reimbursements, the Governing Board, the NAEP, and the standard-setting hotel site; a meeting agenda; the 2011 NAEP achievement levels descriptions (ALDs) for writing for all grades, and the Writing Framework for the 2011 NAEP. The panelists were also informed about the incentive program, as described later in this report.

### 2.6.1 Advanced Materials for Pilot Study

In addition to the advanced materials common to all the meetings, additional logistical information was provided to the pilot study and the operational ALS meeting, both conducted in St. Louis. For the pilot study, an e-mail was sent to the panelists approximately three weeks before the meeting containing information about flight arrangements, hotel accommodations, and on-site registration. This e-mail also contained links to the 2011 NAEP Writing Framework and the websites for the Nation's Report Card, the Governing Board, and the airport. Attached to this e-mail were the ALDs for writing, preliminary meeting agenda, confidentiality agreement, press release form, and hotel map (or floor plan).

This was the first in a series of e-mails designed to provide logistics information and enhance the panelists' sense of ownership of their roles in the ALS activities. When clarification was needed by a panelist on some aspect of the advanced materials, additional e-mail and phone communications took place to answer questions and ensure understanding.

The hard-copy advanced materials were shipped to the panelists approximately one week before the pilot study. In addition to the common advanced materials, this shipment included a briefing booklet, confidentiality agreement, press release form, hotel map, and a copy of the Nation's Report Card for Grade 12 Reading and Mathematics for 2009. Shipping was monitored via online tracking.

An additional e-mail contained information related to dress code, ground transportation arrangements, hotel check-in, early registration, substitute teacher reimbursement, and important phone numbers.

### 2.6.2 Advanced Materials for Field Trial 2

The advanced materials for field trial 2 were sent to the panelists electronically, and included not only the materials common to all the meetings, but also an updated version of the ALDs for all grades. Additionally, the panelists were provided details of an increased honorarium that reflected their participation in the exceptionally long, 10-hour day.

### 2.6.3 Advanced Materials for Operational ALS Meeting

As before, advanced materials were sent to the operational ALS meeting panelists electronically and on hard copy to familiarize them with the writing assessment and the goals for the ALS meeting. Although these materials were predominantly the same as previous meetings, the organization and method of delivery evolved and improved enough to warrant a full description, provided below.

Once a nominee was selected, an e-mail was sent requesting they confirm their selection. This e-mail included hyperlinks to the Nation's Report Card Writing page and the hotel, as well as basic information about dates, times, accommodations and reimbursements. After confirmation was received from the panelists, advanced materials were sent on hard copy and electronically.

A packet of hard-copy materials was shipped to the panelists one week before the operational ALS meeting. This packet included the 2011 Writing Framework, ALDs for all grades, tentative meeting agenda, hotel floor plan, a public transportation map, confidentiality agreement, press release form, and cover letter. The panelists were invited to bring these hard-copy materials, but were also informed that copies would be available on-site. Panelists who were confirmed after the original shipment date of the hard-copy materials were sent the materials via overnight delivery; thus, all the panelists received hard-copy materials in advance of the meeting.

In addition to the hard-copy advanced materials mailing, electronic materials were sent via e-mail, including some of the advanced materials and logistical information; however, instead of the series of e-mails sent to the pilot study panelists, a single, detail-rich NAEP Panelist Details e-mail was sent to the operational ALS meeting panelists a week prior to the meeting. The single-e-mail approach was selected to ensure consistency and facility in communicating logistics to all the panelists, regardless of their selection date. Additionally, the logistical details were intentionally presented in the clearest, most-easily consumable and usable format possible because the confirmation of some panelists occurred very near to the onset of the operational ALS meeting. Presenting these details in a singular, clear, and exhaustive e-mail was later reflected upon favorably by a number of the panelists.

The body of this e-mail thanked panelists for agreeing to serve on one of the panels, and included information about travel and hotel accommodations, reimbursements, netbook shipment (see Incentives and Reimbursements, below), dress code, pertinent contact information, and an early registration table the night before the meeting began. The e-mail also requested panelist input about dietary restrictions and special study participation.

Attached to this e-mail were electronic copies of the confidentiality agreement, press release form, tentative meeting agenda, hotel floor plan, and public transportation information. Embedded in the text of the e-mail were hyperlinks to the 2011 NAEP Writing Framework and the websites for the Nation's Report Card, the Governing Board, the airport, and the hotel, rounding out the materials provided to familiarize the panelists with their role in advance of the meeting. Panelists confirmed after the initial distribution of this e-mail were given copies immediately upon confirmation. A copy of this "details" e-mail is included in Appendix B.

There were additional individual communications via e-mail and phone as necessary to address personal requests and questions raised by the panelists. Additionally, a brief e-mail was sent days after the initial e-mail to remind the panelists to promptly complete their flight arrangements through the travel agency.

## 2.7 Incentives and Reimbursements

During the recruitment process, nominators and nominees were notified that a netbook would be given to those selected as multi-day panelists. The netbooks were to serve as an incentive for participation in the studies, and the netbooks selected for the incentive program had a 10.1-inch screen, 1 GB RAM, 160 GB hard disk drive, and a webcam. After each applicable meeting (i.e., field trial, pilot study, and operational ALS meeting), a netbook was shipped to each panelist, accompanied by instructions for first startup and a cover letter.

The panelists were also reimbursed for mileage and related travel expenses per federal guidelines. Similarly, meal expenses were reimbursed for meals not included as during the standard-setting process. A reimbursement request form was provided to the panelists at the standard-setting site to be filled out and returned to program management at the conclusion of the meeting. Specified on the form were the appropriate *per diem* amounts in conjunction with entry points for additional travel expenses and mileage. Once the forms were returned and the information confirmed for accuracy, the requests were honored, and program management delivered the reimbursement checks, along with a certificate of appreciation and a cover letter (Appendix B), to the panelists.

Finally, the school districts with teachers on the panel were invited to seek reimbursement of costs associated with hiring substitute classroom coverage in the teacher panelists' absence. The average budgeted amount for this reimbursement was $100.00 per day per substitute. The districts were instructed to submit their requests on district letterhead or official invoice, signed by the teacher's supervising principal or superintendent; in one instance, the signature of the district's controller was accepted. Once program management reviewed the information and confirmed the accuracy of the request, a reimbursement check was mailed directly to the district.

### *2.7.1 Incentives and Reimbursements for Pilot Study*

Travel reimbursements were a larger consideration for the pilot study and the operational ALS meeting, although they were reported on the same kind of form used by the other meetings. The major travel expenses were funded by the contract at no cost to the panelists.

The panelists were sent contact information and instructions for setting up their flight to St. Louis through a travel agency that Measured Progress worked with very closely. Billing was designed to allow the contract to cover the fares when the panelists called to make their reservations.

Ground transportation in St. Louis to and from the airport was similarly prearranged by Measured Progress. For travel from the airport to the hotel, a cab company was secured in advance to provide transportation at no charge panelists; expenses were billed to Measured Progress. The panelists were instructed to call a cab at the number provided upon their arrival at the airport. Ground transportation for the return to the airport was provided by charter bus, contracted through a small St. Louis business.

### *2.7.2 Incentives and Reimbursements for Field Trial 2*

As with other meeting panelists, field trial 2 panelists were reimbursed for mileage and travel expenses consistent with federal guidelines. In view of the 10-hour length of the single-day meeting, field trial 2 panelists were given a $200.00 honorarium. Meal reimbursements were quite low because, due to the schedule of the day, Measured Progress provided a working breakfast and working lunch.

### 2.7.3 Incentives and Reimbursements for Operational ALS Meeting

In addition to the incentives as described for other meetings, an additional $100.00 honorarium was granted to those who opted to participate in special study 2, which immediately followed the operational ALS meeting. Participants in special study 2 received a modified version of the form that accounted for the additional day's *per diem* and the $100.00 honoraria. The panelist reimbursement form may be reviewed in Appendix B.

Similar to the pilot study, the panelists were given information about flight arrangements, ground transportation, and hotel accommodations, all provided by the contract at no charge to the panelists. It is worth noting that a number of operational ALS meeting panelists requested amended flight itineraries in order to accommodate plans to visit family or friends after the operational ALS meeting. In these cases, the panelists paid the fare up-front, and Measured Progress reimbursed the amount of the standard fare that had been offered to them initially; the flight amendments were not charged to the contract.

Finally, additional support was provided by Measured Progress's IT group to a panelist whose netbook was malfunctioning by retrieving and replacing the defective equipment.

## 2.8 Registration

On the first day of standard setting, panelists registered with program management and confirmed their contact information, submitted their signed confidentiality agreements and press release forms, and were given personalized materials folders that included the meeting agenda, grade-specific ALDs, policy definitions, briefing booklet, reimbursement request form, and name badges with the panelists' software (BoWTIE) usernames printed on them.

### 2.8.1 Registration for the Pilot Study

In addition to the registration described above for all meetings, early registration for the pilot study was made available on the evening of the travel day. This was announced in the final logistics e-mail to the panelists in advance of the meeting, but was not well attended. Additionally, hard copies of the 2011 Writing Framework and the Nation's Report Card for Grade 12 Reading and Mathematics for 2009 were available at both registration times.

### 2.8.2 Registration for Field Trial 2

Due to the unique recruitment necessitated by field trial 2, the panelists were asked to provide a photo I.D. at registration. Narrative and matrix forms of the updated ALDs were provided in the personalized materials folders in addition to the materials specified above. Hard copies of the 2011 Writing Framework were also made available.

### 2.8.3 Registration for Operational ALS Meeting

As in the pilot study, registration was held prior to the first session on the first day of the meeting, with early registration made available on the evening of the travel day. Contrary to the pilot study, early registration for the operational ALS meeting was very well attended. As a result, the panelists had the opportunity to voice their last-minute logistical questions and mingle with one another in advance of the meeting's commencement. The early registration was promoted in the details e-mail.

The updated ALDs were again included in the personalized panelist folders along with the policy definitions, briefing booklet, confidentiality agreement, press release form, reimbursement request form, hotel map, public transportation information, and personalized name badges, which had the panelists' BoWTIE login information affixed to the back. Hard-copy versions of the 2011 Writing Framework and the Nation's Report Card for Grade 12 Reading and Mathematics for 2009 were available at the registration table upon request.

## 2.9 Achievement Levels–Setting Process

The ALS process refers to all activities through which the three components of the achievement levels are obtained. The NAEP writing achievement levels are composed of the ALDs, the cut scores and percentages of students at or above the cut scores, and student work illustrative of what students performing at each achievement level know and can do. This section describes all activities, using the informed judgments of well-qualified and broadly representative panels, contributing to the establishment of each component.

### *2.9.1 Development of Achievement Levels Descriptions*

The ALDs are the statements of the standards that are translated to the scale through the ALS process. For the current project, they are the operational definition for the 2011 NAEP writing of the policy definitions established by the Governing Board. Historically, starting with the 1998 ALS process for NAEP civics and writing, ALDs have been developed prior to convening the ALS panelists. Having the ALDs reviewed and finalized prior to convening the ALS panels saves time in the process and allows panelists to focus on their understanding of the descriptions (Loomis, 2012).

The ALDs for the 2011 NAEP writing were developed by a contractor to the Governing Board and were provisionally approved by the Governing Board in August of 2011 to be used in the ALS process, with full approval contingent upon the results of the studies. The ALDs were first used in the field trial, which was implemented only for grade 12. They were again used in the pilot study. After consideration of possible influences on the results of the pilot/special study, TACSS recommended that the ALDs be revised. Additionally, the Committee on Standards, Design and Methodology (COSDAM) strongly recommended that if there would be a modification to the ALDs, they should be tested with panelists in a small-scale study.

The content facilitators, George Kamberelis and Pat Porter, worked with Governing Board staff to better align the ALDs with the policy definitions. The ALDs were also revised to make the language more parallel within achievement levels across grades, and within each grade across achievement levels. A matrix version of the ALDs helped in ensuring the parallelism of the descriptions. This version of the ALDs was tested with panelists during a small-scale study referred to as field trial 2, for which panelists had access to both the narrative and matrix formats. More minor modifications were made, based on the panelists' debriefing at the end of field trial 2. The final version of the ALDs for grades 4, 8, and 12[7] is provided in narrative format in Figures 4 through 6, and in matrix format in Tables 6 through 8.

[7] The Governing Board developed ALDs for all three grades to help ensure appropriate calibration and alignment across grades and levels, although grade 4 was not part of the ALS project described in this report.

*Figure 4: Writing Achievement Levels for Grade 4*

**BASIC**

Fourth-grade students writing at the Basic level should be able to address the tasks appropriately and at least partially accomplish their communicative purposes. Texts should be appropriately structured. Many of the ideas in the texts should be developed, and their texts should include supporting details and examples that are relevant to the topic, purpose, and audience. Most sentences should be well structured, and texts may be composed mostly of simple sentences. Many of the words and phrases should be appropriate to the topics, purposes, and audiences. Spelling, grammar, usage, capitalization, and punctuation skills should be sufficiently accurate to convey general meaning, although there may be some errors that detract from meaning.

**PROFICIENT**

Fourth-grade students writing at the Proficient level should be able to address the tasks appropriately and accomplish their communicative purposes. Texts should be appropriately structured and coherent. Most of the ideas in their texts should be developed effectively, and their texts should include supporting details and examples that support the main ideas. Texts should have well structured sentences and a variety of sentence types—simple, compound, and complex. Words and phrases should be thoughtfully selected and appropriate to the topics, purposes, and audiences. Spelling, grammar, usage, capitalization, and punctuation should be sufficiently accurate to communicate clearly with the reader. There may be some errors in the texts, but these errors should not impede meaning.

**ADVANCED**

Fourth-grade students writing at the Advanced level should be able to address the tasks appropriately and accomplish their communicative purposes in effective ways. Texts should be well structured and coherent. The ideas in the texts should be developed fully and effectively. Their texts should include supporting details and examples that are closely related to the topic, purpose, and audience and that enhance communicative effectiveness. Sentences should be well structured, and texts should include a variety of sentence types (simple, compound, and complex) to enhance their communicative effectiveness. Words and phrases should be chosen skillfully, and they should both enrich meaning in the texts and enhance communicative effectiveness. Spelling, grammar, usage, capitalization, and punctuation should be mostly accurate and well developed, and they should be used appropriately. Grammatical, mechanical, and usage choices should contribute to communicative effectiveness. There may be a few errors, but they should not impede meaning.

*Figure 5: Writing Achievement Levels for Grade 8*

**BASIC**

Eighth-grade students writing at the Basic level should be able to address the tasks appropriately and mostly accomplish their communicative purposes. Their texts should be coherent and effectively structured. Many of the ideas in their texts should be developed effectively. Supporting details and examples should be relevant to the main ideas they support. Voice should align with the topic, purpose, and audience. Texts should include appropriately varied uses of simple, compound, and complex sentences. Words and phrases should be relevant to the topics, purposes, and audiences. Knowledge of spelling, grammar, usage, capitalization, and punctuation should be made evident; however, there may be some errors in the texts that impede meaning.

**PROFICIENT**

Eighth-grade students writing at the Proficient level should be able to develop responses that clearly accomplish their communicative purposes. Their texts should be coherent and well structured, and they should include appropriate connections and transitions. Most of the ideas in the texts should be developed logically, coherently, and effectively. Supporting details and examples should be relevant to the main ideas they support, and contribute to overall communicative effectiveness. Voice should be relevant to the tasks and support communicative effectiveness. Texts should include a variety of simple, compound, and complex sentence types combined effectively. Words and phrases should be chosen thoughtfully and used in ways that contribute to communicative effectiveness. Solid knowledge of spelling, grammar, usage, capitalization, and punctuation should be evident throughout the texts. There may be some errors, but these errors should not impede meaning.

**ADVANCED**

Eighth-grade students writing at the Advanced level should be able to construct skillful responses that accomplish their communicative purposes effectively. Their texts should be coherent and well structured throughout, and they should include effective connections and transitions. Ideas in the texts should be developed logically, coherently, and effectively. Supporting details and examples should skillfully and effectively support and extend the main ideas in the texts. Voice should be distinct and enhance communicative effectiveness. Texts should include a well-chosen variety of sentence types, and the sentence structure variations should enhance communicative effectiveness. Words and phrases should be chosen strategically, with precision, and in ways that enhance communicative effectiveness. An extensive knowledge of spelling, grammar, usage, capitalization, and punctuation should be evident throughout the texts. Appropriate use of these features should enhance communicative effectiveness. There may be a few errors, but these errors should not impede meaning.

*Figure 6: Writing Achievement Levels for Grade 12*

**BASIC**

Twelfth-grade students writing at the Basic level should be able to respond effectively to the tasks and accomplish their communicative purposes. Their texts should be coherent and well structured. Most of the ideas in their texts should be developed effectively. Relevant details and examples should be used to support and extend the main ideas in the texts. Voice should support the communicative purposes of the texts. Texts should include appropriately varied simple, compound, and complex sentence types. Words and phrases should be suitable for the topics, purposes, and audiences. Substantial knowledge of spelling, grammar, usage, capitalization, and punctuation should be clearly evident. There may be some errors in the texts, but these errors should not generally impede meaning.

**PROFICIENT**

Twelfth-grade students writing at the Proficient level should address the tasks effectively and fully accomplish their communicative purposes. Their texts should be coherent and well structured with respect to these purposes, and they should include well-crafted and effective connections and transitions. Their ideas should be developed in a logical, clear, and effective manner. Relevant details and examples should support and extend the main ideas of the texts and contribute to their overall communicative effectiveness. Voice should be relevant to the tasks and contribute to overall communicative effectiveness. Texts should include a variety of simple, compound, and complex sentence types that contribute to overall communicative effectiveness. Words and phrases should be chosen purposefully and used skillfully to enhance the effectiveness of the texts. A solid knowledge of spelling, grammar, usage, capitalization, and punctuation should be evident throughout the texts. There may be some errors in the texts, but they should not impede meaning.

**ADVANCED**

Twelfth-grade students writing at the Advanced level should be able to address the tasks strategically, fully accomplish their communicative purposes, and demonstrate a skillful and creative approach to constructing and delivering their messages. Their texts should be coherent and well structured; they should include skillfully constructed and effective connections and transitions; and they should be rhetorically powerful. All of the ideas in their texts should be developed clearly, logically, effectively, and in focused and sophisticated ways. Supporting details and examples should be well crafted; they should skillfully support and extend the main ideas; and they should strengthen both communicative effectiveness and rhetorical power of the texts. A distinct voice that enhances the communicative effectiveness and rhetorical power of the texts should be evident. Texts should include a variety of sentence structures and types that are skillfully crafted and enhance communicative effectiveness and rhetorical power. Words and phrases should be chosen purposefully, with precision, and in ways that enhance communicative effectiveness and rhetorical power. A highly developed knowledge of spelling, grammar, usage, capitalization, and punctuation should be evident throughout the texts and function in ways that enhance communicative effectiveness and rhetorical power. There may be a few errors in the texts, but they should not impede meaning.

*Table 6: Writing Achievement Levels for Grade 4*

| Dimension | BASIC | PROFICIENT | ADVANCED |
|---|---|---|---|
| Addressing Communicative Purpose | Fourth-grade students writing at the Basic level should be able to address the tasks appropriately and at least partially accomplish their communicative purposes. | Fourth-grade students writing at the Proficient level should be able to address the tasks appropriately and accomplish their communicative purposes. | Fourth-grade students writing at the Advanced level should be able to address the tasks appropriately and accomplish their communicative purposes in effective ways. |
| Text Structure & Coherence | Their texts should be appropriately structured. | Their texts should be appropriately structured and coherent. | Texts should be well structured and coherent. |
| Idea Development | Many of the ideas in the texts should be developed | Most of the ideas in their texts should be developed effectively. | The ideas in the texts should be developed fully and effectively. |
| Details & Elaboration | Their texts should include supporting details and examples that are relevant to the topic, purpose, and audience. | Their texts should include supporting details and examples that support the main ideas in the texts. | Their texts should include supporting details and examples that are closely related to topics, purposes, and audiences and that enhance communicative effectiveness. |
| Sentence Structure & Complexity | Most sentences should be well structured, and texts may be composed mostly of simple sentences. | Texts should have well structured sentences and a variety of sentence types—simple, compound, and complex. | Sentences should be well structured, and texts should include a variety of sentence types—simple, compound, and complex—that enhance communicative effectiveness. |
| Word & Phrase Choice | Many of the words and phrases should be appropriate to the topics, purposes, and audiences. | Words and phrases should be thoughtfully selected and appropriate to the topics, purposes, and audiences. | Words and phrases should be chosen skillfully, and they should both enrich meaning and enhance communicative effectiveness. |
| Grammar Usage Mechanics | Spelling, grammar, usage, capitalization, and punctuation skills should be sufficiently accurate to convey general meaning, although there may be some errors that detract from meaning. | Spelling, grammar, usage, capitalization, and punctuation should be sufficiently accurate to communicate clearly with the reader. There may be some errors in the texts, but these errors should not impede meaning. | Spelling, grammar, usage, capitalization, and punctuation should be mostly accurate and well developed, and used appropriately. Grammatical, mechanical, and usage choices should contribute to communicative effectiveness. There may be a few errors, but they should not impede meaning. |

*Table 7: Writing Achievement Levels for Grade 8*

| Dimension | BASIC | PROFICIENT | ADVANCED |
|---|---|---|---|
| Addressing Communicative Purpose | Eighth-grade students writing at the Basic level should be able to address the tasks appropriately and mostly accomplish their communicative purposes. | Eighth-grade students writing at the Proficient level should be able to develop responses that clearly accomplish their communicative purposes. | Eighth-grade students writing at the Advanced level should be able to construct skillful responses that accomplish their communicative purposes effectively. |
| Text Structure & Coherence | Their texts should be coherent and effectively structured. | Their texts should be coherent and well structured, and they should include appropriate connections and transitions. | Their texts should be coherent and well structured throughout, and they should include effective connections and transitions. |
| Idea Development | Many of the ideas in their texts should be developed effectively. | Most of the ideas in the texts should be developed logically, coherently, and effectively. | The ideas in the texts should be developed logically, coherently, and effectively. |
| Details & Elaboration | Supporting details and examples should be relevant to the main ideas they support. | Supporting details and examples should be relevant to the main ideas they support, and contribute to overall communicative effectiveness. | Supporting details and examples should skillfully and effectively support and extend the main ideas in the texts. |
| Voice | Voice should align with the topic, purpose, and audience. | Voice should be relevant to the tasks and support communicative effectiveness. | Voice should be distinct and enhance communicative effectiveness. |
| Sentence Structure & Complexity | Texts should include appropriately varied uses of simple, compound, and complex sentences. | Texts should include a variety of simple, compound, and complex sentence types combined effectively. | Texts should include a well-chosen variety of sentence types, and the sentence structure variations should enhance communicative effectiveness. |
| Word & Phrase Choice | Words and phrases should be relevant to the topics, purposes, and audiences. | Words and phrases should be chosen thoughtfully and used in ways that contribute to communicative effectiveness. | Words and phrases should be chosen strategically, with precision, and in ways that enhance communicative effectiveness. |
| Grammar Usage Mechanics | Knowledge of spelling, grammar, usage, capitalization, and punctuation should be made evident; however, there may be some errors in the texts that impede meaning. | Solid knowledge of spelling, grammar, usage, capitalization, and punctuation should be evident throughout the texts. There may be some errors, but these errors should not impede meaning. | An extensive knowledge of spelling, grammar, usage, capitalization, and punctuation should be evident throughout the texts. Appropriate use of these features should enhance communicative effectiveness. There may be a few errors, but these errors should not impede meaning. |

*Table 8: Writing Achievement Levels for Grade 12*

| Dimension | BASIC | PROFICIENT | ADVANCED |
|---|---|---|---|
| Addressing Communicative Purpose | Twelfth-grade students writing at the Basic level should be able to respond effectively to the tasks and accomplish their communicative purposes. | Twelfth-grade students writing at the Proficient level should address the tasks effectively and fully accomplish their communicative purposes. | Twelfth-grade students writing at the Advanced level should be able to address the tasks strategically, fully accomplish their communicative purposes, and demonstrate a skillful and creative approach to constructing and delivering their messages. |
| Text Structure & Coherence | Their texts should be coherent and well structured. | Their texts should be coherent and well structured with respect to these purposes, and they should include well-crafted and effective connections and transitions. | Their texts should be coherent and well structured; they should include skillfully constructed and effective connections and transitions; and they should be rhetorically powerful. |
| Idea Development | Most of the ideas in their texts should be developed effectively. | Their ideas should be developed in a logical, clear, and effective manner. | All of the ideas in their texts should be developed clearly, logically, effectively, and in focused and sophisticated ways. |
| Details/ Elaboration | Relevant details and examples should be used to support and extend the main ideas in the texts. | Relevant details and examples should support and extend the main ideas of the texts and contribute to overall communicative effectiveness. | Supporting details and examples should be well crafted; they should skillfully support and extend the main ideas; and they should strengthen both communicative effectiveness and rhetorical power. |
| Voice | Voice should support the communicative purposes of the texts. | Voice should be relevant to the tasks and contribute to overall communicative effectiveness. | A distinct voice that enhances the communicative effectiveness and rhetorical power of the texts should be evident. |
| Sentence Structure & Complexity | Texts should include appropriately varied simple, compound, and complex sentence types. | Texts should include a variety of simple, compound, and complex sentence types that contribute to overall communicative effectiveness. | Texts should include a variety of sentence structures and types that are skillfully crafted and enhance communicative effectiveness and rhetorical power. |
| Word & Phrase Choice | Words and phrases should be suitable for the topics, purposes, and audiences. | Words and phrases should be chosen purposefully and used skillfully to enhance communicative effectiveness. | Words and phrases should be chosen purposefully, with precision, and in ways that enhance communicative effectiveness and rhetorical power. |
| Grammar Usage Mechanics | Substantial knowledge of spelling, grammar, usage, capitalization, and punctuation should be clearly evident. There may be some errors in the texts, but these errors should not generally impede meaning. | A solid knowledge of spelling, grammar, usage, capitalization, and punctuation should be evident throughout the texts. There may be some errors in the texts, but they should not impede meaning. | A highly developed knowledge of spelling, grammar, usage, capitalization, and punctuation should be evident throughout the texts and function in ways that enhance communicative effectiveness and rhetorical power. There may be a few errors in the texts, but they should not impede meaning. |

### *2.9.2 Body of Work Method*

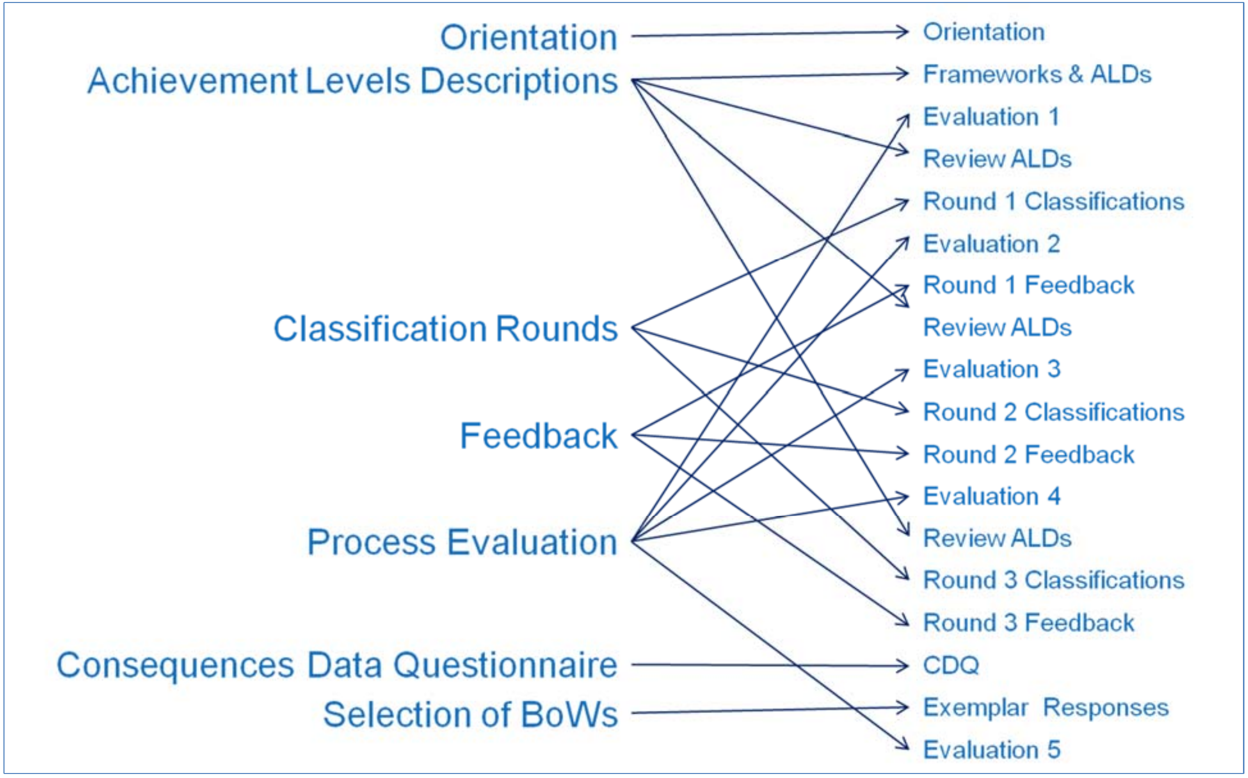
Measured Progress implemented the BoW method to set cut scores for the NAEP writing ALS process. The BoW method belongs to the holistic family of standard-setting methods in which the panelist's task consists of reviewing a series of examinee work samples, or bodies of work (BoWs), and assigning each sample to one of several performance categories (Hambleton & Pitoniak, 2006). Perhaps the most widely used of holistic methods (Cizek & Bunch, 2007), the BoW method (Kingston, Kahl, Sweeney, & Bay, 2001; Kingston & Tiemann, 2012) is the method deemed most appropriate for writing assessments, because it was developed specifically for use with performance assessments that are designed to measure student achievement using open-response items such as writing tasks.

The BoW standard-setting process includes an orientation and introduction to the assessment along with the purpose of the standard-setting meeting, a detailed review of the BoW method, activities for the purpose of gaining a common understanding of the ALDs, training in the BoW classification tasks, and three rounds of classifying student work samples, or BoWs, each followed by a process evaluation and presentation of feedback based on the classification round. For the classification tasks, each panelist assigns each BoW to an achievement level based on his or her understanding of the ALDs and the knowledge, skills, and abilities (KSAs) demonstrated in each BoW.

The BoW method was used in an implementation consistent with NAEP tradition; that is, with feedback information contributing to what was expected to be progressively better judgments by the selected panelists. Figure 7 presents the stages of the iterative process implemented for setting achievement levels for the grades 8 and 12 NAEP writing assessment. Each step is discussed more fully in subsequent sections.

The agendas for the pilot study and the ALS operational meeting reflect the iterative process.

*Figure 7: NAEP ALS Iterative Process*



The classification tasks for the traditional BoW method involve two distinct phases: rangefinding and pinpointing. In the rangefinding phase, BoWs representing the entire range of possible scores are presented for classification. Based on the cut scores resulting from the rangefinding phase, the pinpointing phase uses only work samples in the vicinity of the rangefinding cut scores to focus more precisely on the performance that best represents the standard. The perceived benefit of the pinpointing round is that BoWs are selected based on the location of the cut scores resulting from the previous round, such that the final cut scores recommended contain a higher degree of precision. Findings from the pilot study proved that this was not the case.

Determining a cut score based only on the classifications of pinpointing BoWs for one level was found to be problematic.

For the NAEP ALS process, the three rounds of ratings were originally planned to consist of two rangefinding rounds and one pinpointing round. This plan was implemented in the pilot study, but modified for the operational ALS meeting, where the third round was changed to an additional rangefinding round with a new set of BoWs.

After all panelists completed their ratings for each round, individual cut scores were calculated using logistic regression. The group's cut score is the median of individual panelists' cut scores. The median is the central tendency statistic of choice for this purpose because it is less susceptible to the effects of extreme values.

In statistics, logistic regression is a model used for prediction of the probability of occurrence of an event by fitting data to a logistic curve. In standard setting, an event consists of a panelist's classification of a work sample. By setting up dichotomies, denoting whether a work sample is classified below or above each achievement level, a logistic curve can be established. This logistic curve represents the empirical relationship among the scaled scores of all BoWs and a panelist's ratings. The inflection point of the logistic curve corresponds to an estimate of the panelist's cut score. For each panelist, a logistic curve is fit for each cut score and the estimate for each group's cut score is the median across panelists. Details of the computations are presented in the Technical Report.

After each round of ratings, panelists received several pieces of feedback based on the classifications they provided. After the first round, panelists were provided the group cut scores as well as distributional information of individual panelists' cut scores indicating interrater reliability. Panelists were also provided tally information on the individual work samples—that is, the number of panelists who classified each student work sample into each achievement level. This information was used in the discussion of specific BoWs prior to the second round of classifications. During this discussion, panelists were encouraged both to share their own point of view as well as to listen to the thoughts of their colleagues. The goal was to allow each panelist the opportunity to explain why he or she classified each BoW into one achievement level or another. Facilitators made sure the panelists understood that the purpose of the discussion was not to come to consensus; at every point throughout the standard-setting process, panelists were asked to provide their own best judgment. Once the discussions were complete, the panelists completed the Round 2 ratings using the same set of examinee work samples.

After the second round of ratings, panelists were again informed of the group cut scores and distributional information. Additionally, panelists were provided consequences data feedback. More commonly known as impact data, the consequences data provided to panelists included the percentage of students performing at or above each cut score. Similar feedback was provided to the panelists after Round 3. After the presentation of consequences data feedback based on Round 3 cut scores, panelists were administered a questionnaire asking whether they would recommend that the Governing Board adopt the achievement levels based on Round 3 cut scores. Panelists were given an opportunity to review the ALDs before each round of classification.

For the third component of the achievement levels, panelists selected BoWs that represented performance illustrative of each level. Panelists then filled out the last of the five evaluation questionnaires strategically scheduled after different milestones in the meeting.

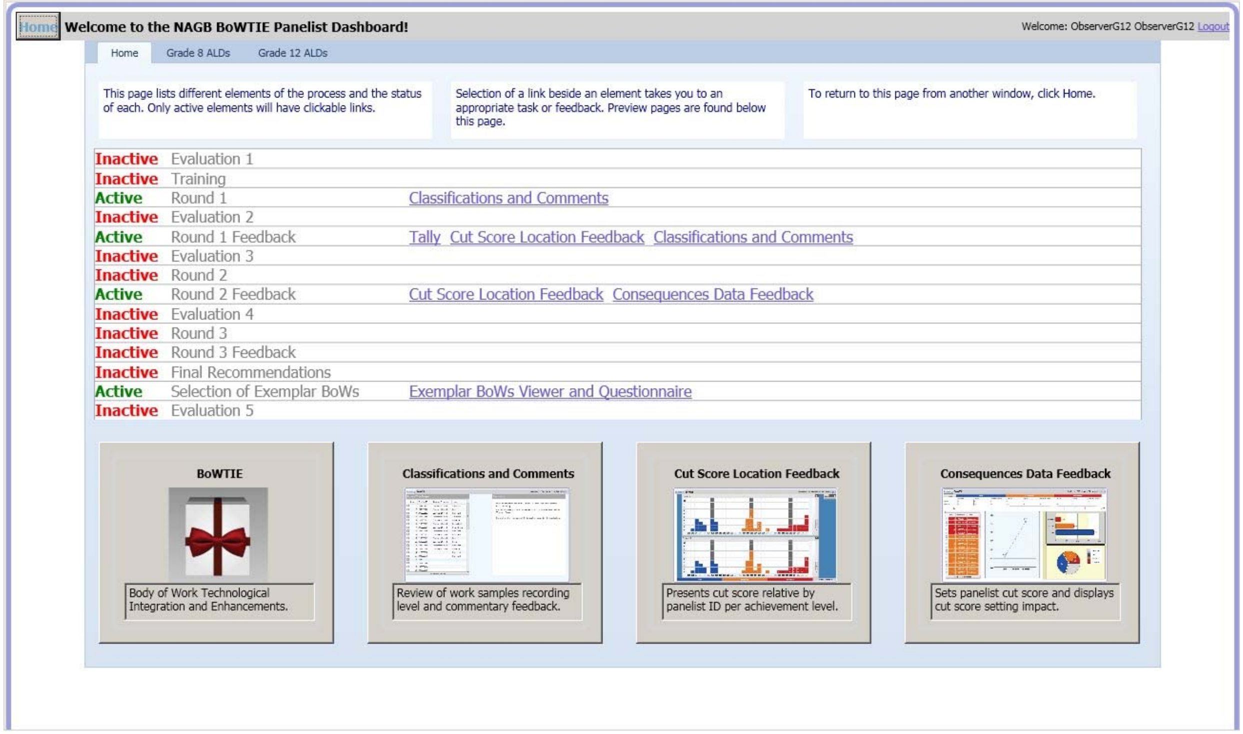
### 2.9.3 Body of Work Technological Integration and Enhancements

The Governing Board requested that some aspects of the ALS process be computerized to increase efficiency of implementation. Measured Progress computerized the entire process. The Body of Work Technological Integration and Enhancements (BoWTIE), a computer-based tool, was designed and developed specifically to aid in the BoW standard-setting process for the NAEP writing assessment in grades 8 and 12. This tool was designed to enhance the adequacy and efficiency of the standard-setting process. The design of the tool is described in this section.

Within BoWTIE, all aspects of the ALS process using the BoW method were integrated, including (a) selection of student work samples, (b) panelist training, (c) rounds of rating, (d) feedback, and (e) process evaluations. An additional feature of BoWTIE was the capability of providing interactive consequences data feedback.

Figure 8 presents the landing page, or Dashboard, for BoWTIE, which is the first page that panelists saw after they were securely logged-in to the system. The Dashboard provided a list of all the different parts of the ALS process for which BoWTIE was used. The links to different functionalities became active when the stage was activated by the CoSS. Figure 8 shows different stages in "Active" and "Inactive" modes, showing the links for functionalities corresponding to the "Active" stages. Note that BoWTIE was designed so that the appropriate stage could be activated by the CoSS as needed, while other stages remained inactive. Note also that the Dashboard includes a tab for the ALDs for each grade. Panelists had access to the ALDs through BoWTIE during the entire process. Hard copies of the ALDs were also provided for panelists' use.

*Figure 8: BoWTIE Dashboard*



Welcome to the NAGB BoWTIE Panelist Dashboard!

Welcome: ObserverG12 ObserverG12

This page lists different elements of the process and the status of each. Only active elements will have clickable links.

Selection of a link beside an element takes you to an appropriate task or feedback. Preview pages are found below this page.

To return to this page from another window, click Home.

| | | |
|---|---|---|
| Inactive | Evaluation 1 | |
| Inactive | Training | |
| Active | Round 1 | Classifications and Comments |
| Inactive | Evaluation 2 | |
| Active | Round 1 Feedback | Tally Cut Score Location Feedback Classifications and Comments |
| Inactive | Evaluation 3 | |
| Inactive | Round 2 | |
| Active | Round 2 Feedback | Cut Score Location Feedback Consequences Data Feedback |
| Inactive | Evaluation 4 | |
| Inactive | Round 3 | |
| Inactive | Round 3 Feedback | |
| Inactive | Final Recommendations | |
| Active | Selection of Exemplar BoWs | Exemplar BoWs Viewer and Questionnaire |
| Inactive | Evaluation 5 | |

BoWTIE

Body of Work Technological Integration and Enhancements.

Classifications and Comments

Review of work samples recording level and commentary feedback.

Cut Score Location Feedback

Presents cut score relative by panelist ID per achievement level.

Consequences Data Feedback

Sets panelist cut score and displays cut score setting impact.

The integration of all parts of the standard setting process enhanced the efficiency, security, and replicability. A fully computer-based system allowed greater ease in developing and preparing materials, ensuring consistency of materials among panelists, and simplifying the organization of materials. The wholly computer-based standard setting was both cost-effective and environmentally sensitive, as the need for hard-copy materials was minimized. Panelists accessed and annotated materials and entered their BoW classification directly in a database as a natural extension of computer-based assessment. The use of BoWTIE also enhanced security of the materials during standard setting by eliminating the potential for anyone to take materials from the panel meeting room.

BoWTIE, developed for the writing NAEP had the following features:

- observer access
    - access to all information and functionalities that panelists have

    - ability to classify BoWs without having classifications  used for computing cut scores
- quality assurance check for completeness of classification data
- interactive feedback
- personal annotation tools
- database of panelist information and classification data
- immediate availability of process evaluation reports

Another key advantage to a wholly computerized standard-setting process was the ability to allow panelists to focus more on their ratings and less on managing the vast quantity of material customarily distributed at a standard-setting meeting for writing. A paper-based implementation of the BoW method for NAEP writing would have required each panelist to have a stack of 50 BoWs for each round of classification for a total of between 50 to 100 sheets of papers.[8] BoWTIE provided panelists annotation and navigation tools that enabled them to view various BoWs without having to shuffle through large stacks of printed BoWs. Panelists could view actual student responses by simply clicking on a BoW number. This feature increased their ability to move from one BoW to the next and to flip back and forth to make comparisons between and among student responses.

Because panelists entered their classification data into BoWTIE, data analysis occurred automatically after panelists finished their classification task. All of the computations were programmed in BoWTIE such that cut scores and other feedback were computed and readily available shortly after the last panelist entered his or her

[8] A paper-based implementation of the BoW method for NAEP writing for grades 8 and 12 would have meant printing between 12,000 and 24,000 sheets of student responses.

final BoW classification. The pre-programmed computations have also benefited from a thorough check as part of the software quality assurance (QA).

Other built-in QA features also ensured that all ratings were within range and no blanks were left before panelists finished the rating session. The QA checks on the panelists' ratings, together with the fact that ratings did not have to be entered externally, meant that there were no data entry errors.  Time for checking was eliminated by having this built-in QA check.

The computerization of this NAEP standard-setting process was found to increase the efficiency of operations[9] by reducing the time required for panelists to complete most steps in the process and to complete data analysis and produce feedback to panelists between rounds of classification. It also enhanced the experience of the panelists by reducing the time and effort associated with the standard-setting tasks, as indicated by their positive evaluation of BoWTIE's ease of use. This supports the procedural validity of the standard-setting process.

BoWTIE was designed to meet all the technical and statistical adequacy criteria set by Berk (1986). In fact, BoWTIE addressed a limitation originally cited: ease of implementation. Ease of implementation was no longer a limitation because this computer-based tool eliminated the logistical challenge of preparing materials for the pinpointing round. For the pilot study, when results from Round 2 rangefinding were

[9] The COR made the following observation:

> Rather than waiting hours and hours to get results, results were ready within a minute—literally—of the final classification by the final panelist. And, it was possible to have responses to questionnaires back within a couple of hours for review by facilitators each evening. No more need to read each and every panelist's comments and review their responses. It is now easy to have a good feeling for how the process is going for panelists as a whole and to see if there are any specific panelists with issues or concerns. (S. Cooper Loomis, personal communication, May 18, 2012)

computed and approved by the COR to be used for the next round, BoWTIE selected in real time BoWs that were used for the pinpointing round.[10] [11]

### 2.9.4 Replicate Panels

Each grade-level panel was divided into two panels, referred to as Groups A and B, which were replicates, to the extent possible, with the panel's equivalent relative to the following demographic variables: (a) panelist type, (b) gender, (c) minority status, and (d) NAEP region.

Each group classified BoWs from a different form of the assessment with some forms in common, as described in the following section (2.9.5). The common forms made it possible to assign common BoWs to the two groups, which maximized the equivalence of the two sets of BoWs. The use of demographically equivalent groups and equivalent sets of assessment tasks allowed the reliability of the cut scores to be estimated in a straightforward way.

Tables 9 and 10 illustrate the composition of the pilot study replicate panels for grade 8 and grade 12, respectively. Similarly, Tables 11 and 12 illustrate the composition of the replicate panels for the operational ALS meeting for grade 8 and grade 12, respectively.

[10] The CoSS and the psychometrician had an opportunity to inspect the selections prior to their use in the pinpointing round. Unsuitable selections were manually overridden and replaced by more suitable ones.

[11] The computerization made it manageable to do pinpointing. But, analysis of the data and discussions with TACSS led to concerns about the method of computing cutscores. Ultimately, the decision was to use three rounds of rangefinding, using a new set of BoWs for the third round, instead of a pinpointing round. A full discussion of the pinpointing cut score computation is in section 4.3.4.

*Table 9: Pilot Study Replicate Panel Composition (Grade 8)*

| **Demographic Variable** | **Attribute** | **Group A** | | **Group B** | | **Goal** |
|---|---|---|---|---|---|---|
| | | **n** | **%** | **n** | **%** | **%** |
| Panelist Type | Teachers | 6 | 60 | 5 | 63 | 55 |
| | Nonteacher Educators | 2 | 20 | 2 | 25 | 15 |
| | General Public | 2 | 20 | 1 | 13 | 30 |
| Gender | Female | 8 | 80 | 7 | 88 | 50 |
| | Male | 2 | 20 | 1 | 13 | 50 |
| Race/Ethnicity | Caucasian | 8 | 80 | 7 | 88 | 80 |
| | Non-Caucasian | 2 | 20 | 1 | 13 | 20 |
| NAEP Region | Midwest | 3 | 30 | 3 | 38 | 35 |
| | Northeast | 2 | 20 | 2 | 25 | 20 |
| | South | 2 | 20 | 0 | 0 | 25 |
| | West | 3 | 30 | 3 | 38 | 20 |

*Table 10: Pilot Study Replicate Panel Composition (Grade 12)*

| **Demographic Variable** | **Attribute** | **Group A** | | **Group B** | | **Goal** |
|---|---|---|---|---|---|---|
| | | **n** | **%** | **n** | **%** | **%** |
| Panelist Type | Teachers | 6 | 67 | 6 | 67 | 55 |
| | Nonteacher Educators | 1 | 11 | 1 | 11 | 15 |
| | General Public | 2 | 22 | 2 | 22 | 30 |
| Gender | Female | 5 | 56 | 6 | 67 | 50 |
| | Male | 4 | 44 | 3 | 33 | 50 |
| Race/Ethnicity* | Caucasian | 6 | 67 | 6 | 75 | 80 |
| | Non-Caucasian | 3 | 33 | 2 | 25 | 20 |
| NAEP Region | Midwest | 3 | 33 | 4 | 44 | 35 |
| | Northeast | 3 | 33 | 2 | 22 | 20 |
| | South | 2 | 22 | 3 | 33 | 25 |
| | West | 1 | 11 | 0 | 0 | 20 |

*One panelist in Group B elected not to identify their ethnicity.

*Table 11: Operational ALS Meeting Replicate Panel Composition (Grade 8)*

| **Demographic Variable** | **Attribute** | **Group A** | | **Group B** | | **Goal** |
|---|---|---|---|---|---|---|
| | | **n** | **%** | **N** | **%** | **%** |
| Panelist Type | Teachers | 8 | 62 | 8 | 57 | 55 |
| | Nonteacher Educators | 2 | 15 | 3 | 21 | 15 |
| | General Public | 3 | 23 | 3 | 21 | 30 |
| Gender | Female | 10 | 77 | 12 | 86 | 50 |
| | Male | 3 | 23 | 2 | 14 | 50 |
| Race/Ethnicity | Caucasian | 11 | 85 | 12 | 86 | 80 |
| | Non-Caucasian | 2 | 15 | 2 | 14 | 20 |
| NAEP Region | Midwest | 3 | 23 | 3 | 21 | 35 |
| | Northeast | 2 | 15 | 3 | 21 | 20 |
| | South | 2 | 15 | 4 | 29 | 25 |
| | West | 6 | 46 | 4 | 29 | 20 |

*Table 12: Operational ALS Meeting Replicate Panel Composition (Grade 12)*

| **Demographic Variable** | **Attribute** | **Group A** | | **Group B** | | **Goal** |
|---|---|---|---|---|---|---|
| | | **n** | **%** | **n** | **%** | **%** |
| Panelist Type | Teachers | 7 | 50 | 8 | 57 | 55 |
| | Nonteacher Educators | 3 | 21 | 2 | 12 | 15 |
| | General Public | 4 | 29 | 4 | 29 | 30 |
| Gender | Female | 11 | 79 | 8 | 57 | 50 |
| | Male | 3 | 21 | 6 | 43 | 50 |
| Race/Ethnicity* | Caucasian | 11 | 92 | 14 | 100 | 80 |
| | Non-Caucasian | 1 | 8 | 0 | 0 | 20 |
| NAEP Region | Midwest | 3 | 21 | 4 | 29 | 35 |
| | Northeast | 2 | 14 | 2 | 14 | 20 |
| | South | 4 | 29 | 2 | 14 | 25 |
| | West | 5 | 36 | 6 | 43 | 20 |

*Two panelists in Group A elected not to identify their ethnicity.

### 2.9.5 Task Pool Division

There are 22 writing tasks for each grade; each task is specific to a purpose for writing—to convey, to explain, and to persuade. Each writing assessment form is composed of two tasks for different purposes. Table 13 presents the number of tasks for each purpose as well as the number of tasks that used the four different stimuli—image, text, audio, and video.

*Table 13: Number of Writing Tasks per Writing Purpose and Type of Task*

| Grade | Purpose for Writing | Total | Type of Task | | | | |
|---|---|---|---|---|---|---|---|
| | | | No Stimuli | Text | Visual | Audio | Video |
| 8 | Convey Experience | 6 | 2 | 0 | 1 | 1 | 2 |
| | Explain | 8 | 2 | 1 | 3 | 0 | 2 |
| | Persuade | 8 | 3 | 0 | 1 | 0 | 4 |
| 12 | Convey Experience | 5 | 4 | 1 | 0 | 0 | 0 |
| | Explain | 9 | 4 | 1 | 2 | 0 | 2 |
| | Persuade | 8 | 2 | 1 | 3 | 0 | 2 |

From the 22 tasks, 44 forms were created for the assessment, with each task appearing in exactly four forms—twice as a first task and twice as a second task. All writing tasks were used in setting achievement levels, but not all forms were used. Eleven forms were selected such that the 22 tasks were employed. These were the forms from which BoWs were selected for the panelist classification tasks. However, only seven forms were assigned to each group, to minimize the number of student responses reviewed by each panelist and thus minimize the cognitive demand on the panelists. Of the seven forms assigned to each group, three forms were common across the two groups and four were unique to each group. The assignment of the common forms allowed the consistency of results from the two groups to be checked. The final form assignment, displayed in Table 14, shows the balance between the two groups with respect to the average difficulty of the tasks assigned and the number of tasks for each writing purpose.

*Table 14: Task and Form Assignment*

| Grade | Group | Task Number | Task Information | Common Forms | | | Unique Forms | | | | Average Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Form 1 | Form 2 | Form 3 | Form 4 | Form 5 | Form 6 | Form 7 | |
| 8 | A | 1 | Task ID | 8_1* | 8_2* | 8_3 | 8_4 | 8_5 | 8_6 | 8_7 | 3.36 |
| | | | Purpose | Explain | Convey | Persuade | Explain | Explain | Convey | Explain | |
| | | | Average Score | 3.5610 | 3.6155 | 3.2771 | 3.2953 | 3.4343 | 3.7247 | 3.5348 | |
| | | 2 | Task ID | 8_8 | 8_9* | 8_10 | 8_11 | 8_12 | 8_13 | | |
| | | | Purpose | Persuade | Persuade | Explain | Persuade | Persuade | Persuade | Convey | |
| | | | Average Score | 3.2219 | 3.1934 | 3.5279 | 3.6134 | 3.3062 | 3.2235 | 3.6091 | |
| | B | 1 | Task ID | 8_1* | 8_2* | 8_3 | 8_15 | 8_16 | 8_17 8_14 | | 3.30 |
| | | | Purpose | Explain | Convey | Persuade | Persuade | Explain | Persuade | Explain | |
| | | | Average Score | 3.56100 | 3.6155 | 3.2771 | 3.2898 | 3.2310 | 3.3183 | 3.3171 | |
| | | 2 | Task ID | 8_8 | 8_9* | 8_10 | 8_19 | 8_20 | 8_21 8_18 | | |
| | | | Purpose | Persuade | Persuade | Explain | Explain | Convey | Convey | Convey | |
| | | | Average Score | 3.2219 | 3.1934 | 3.5279 | 3.4108 | 3.4386 | 3.4714 | 3.4752 | |
| 12 | A | 1 | Task ID | 12_1* | 12_2* | 12_3 | 12_4 | 12_5 | 12_6 8_22 | 12_7 | 3.72 |
| | | | Purpose | Explain | Persuade | Convey | Explain | Convey | Persuade | Explain | |
| | | | Average Score | 3.7215 | 3.7262 | 3.9798 | 3.7227 | 4.0122 | 3.6723 | 3.6916 | |
| | | 2 | Task ID | 12_8* | 12_9 | 12_10 | 12_11 | 12_12 | 12_13 | 12_14 | |
| | | | Purpose | Convey | Explain | Persuade | Convey | Explain | Explain | Persuade | |
| | | | Average Score | 3.9205 | 3.5692 | 3.5821 | 4.0119 | 3.6325 | 3.7811 | 3.6505 | |
| | B | 1 | Task ID | 12_1* | 12_2* | 12_3 | 12_15 | 12_16 | 12_17 | 12_18 | 3.66 |
| | | | Purpose | Explain | Persuade | Convey | Persuade | Explain | Explain | Persuade | |
| | | | Average Score | 3.7215 | 3.7262 | 3.9798 | 3.4960 | 3.6015 | 3.6477 | 3.5885 | |
| | | 2 | Task ID | 12_8* | 12_9 | 12_10 | 12_19 | 12_20 | 12_21 | 12_22 | |
| | | | Purpose | Convey | Explain | Persuade | Convey | Persuade | Persuade | Explain | |
| | | | Average Score | 3.9205 | 3.5692 | 3.5821 | 3.7656 | 3.7302 | 3.9505 | 3.7163 | |

*Item marked for release

Details on task assignment and form selection are discussed in the Technical Report.

### 2.9.6 Pseudo-NAEP Scales

Results of the NAEP become public only through the release of the Nation's Report Card. As a matter of security, the NAEP scale is masked during the ALS process. The score scale used for the ALS process was a linear transformation of the NAEP scale, and it was used to keep the results of the ALS process secure. A different set of transformation constants were used for each grade so that panelists were unable to compare the results across grades. Table 15 presents the relationship between the grades 8 and 12 NAEP scales and the pseudo-NAEP scales. Because the same slope is used for the NAEP scale and the pseudo-NAEP scale, these two scales differ only by a constant.

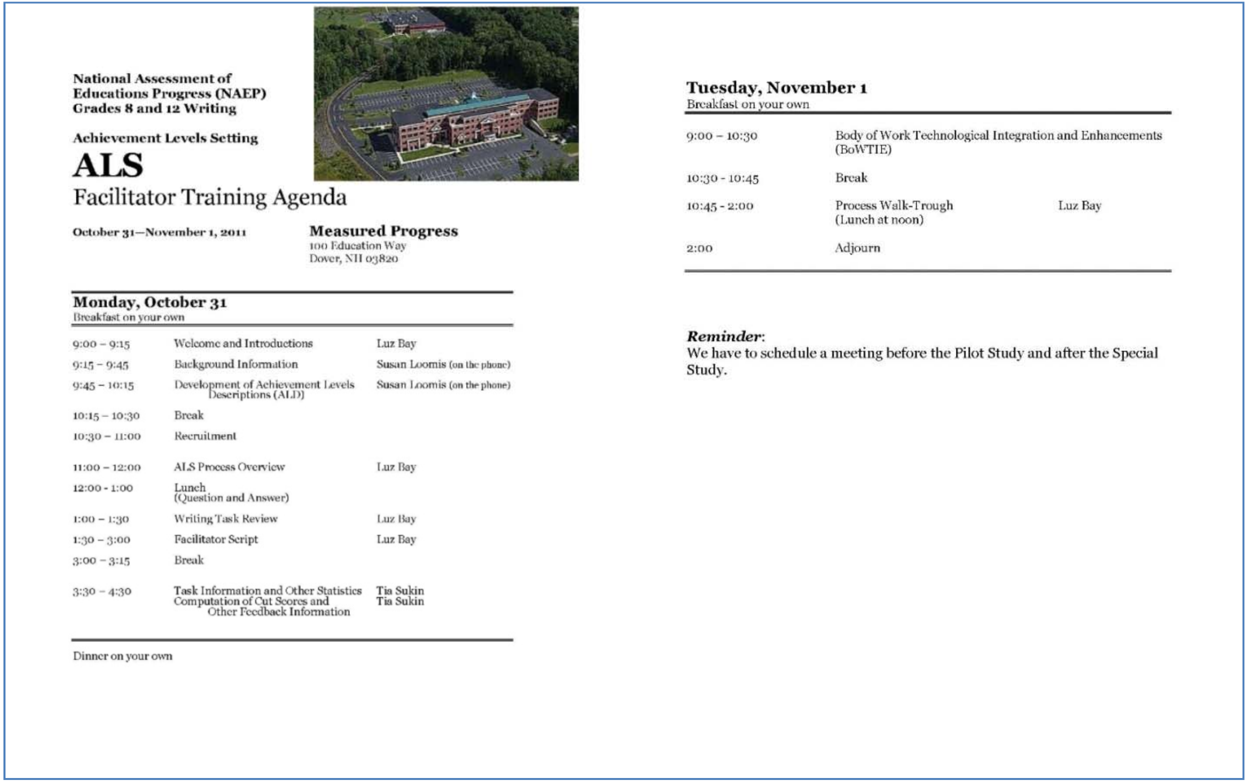
*Table 15: Transformation Constants From Theta*

| Grade | NAEP Scale | | Pseudo-NAEP Scale | |
|---|---|---|---|---|
| | Slope | Intercept | Slope | Intercept |
| 8 | 35.34109 | 150.0236 | 35.34109 | 249.0236 |
| 12 | 35.33142 | 150.0265 | 35.33142 | 566.0265 |

### 2.9.7 Facilitator Training

To maximize the extent to which the process is implemented identically across grade levels, the facilitators were trained by the chief of standard setting. A two-day facilitator training session was held in Dover, New Hampshire, on October 31–November 1, 2011, approximately two weeks before the pilot study meeting. The facilitator training agenda is shown in Figure 9.

*Figure 9: Facilitator Training Agenda*



**National Assessment of**
**Educations Progress (NAEP)**
**Grades 8 and 12 Writing**

**Achievement Levels Setting**

**ALS**

Facilitator Training Agenda

**October 31—November 1, 2011**

**Measured Progress**
100 Education Way
Dover, NH 03820

**Monday, October 31**
Breakfast on your own

| | | |
|---|---|---|
| 9:00 – 9:15 | Welcome and Introductions | Luz Bay |
| 9:15 – 9:45 | Background Information | Susan Loomis (on the phone) |
| 9:45 – 10:15 | Development of Achievement Levels Descriptions (ALD) | Susan Loomis (on the phone) |
| 10:15 – 10:30 | Break | |
| 10:30 – 11:00 | Recruitment | |
| 11:00 – 12:00 | ALS Process Overview | Luz Bay |
| 12:00 - 1:00 | Lunch<br>(Question and Answer) | |
| 1:00 – 1:30 | Writing Task Review | Luz Bay |
| 1:30 – 3:00 | Facilitator Script | Luz Bay |
| 3:00 – 3:15 | Break | |
| 3:30 – 4:30 | Task Information and Other Statistics<br>Computation of Cut Scores and Other Feedback Information | Tia Sukin<br>Tia Sukin |

Dinner on your own

**Tuesday, November 1**
Breakfast on your own

| | | |
|---|---|---|
| 9:00 – 10:30 | Body of Work Technological Integration and Enhancements (BoWTIE) | |
| 10:30 - 10:45 | Break | |
| 10:45 - 2:00 | Process Walk-Trough<br>(Lunch at noon) | Luz Bay |
| 2:00 | Adjourn | |

***Reminder***:
We have to schedule a meeting before the Pilot Study and after the Special Study.

In addition to the two-day training, a facilitator handbook was prepared for use in the pilot study for the purpose of standardizing the instructions given to the panels. The handbook was updated for the operational ALS meeting as appropriate, based on the process modifications and suggestions from the content and process facilitators. Facilitators were not permitted to change the instructions they provided to their panels unless the changes were approved by the CoSS to be implemented in both panels.

The facilitator handbook is a step-by-step set of instructions to be used during the entire standard-setting process. There is a section in the handbook for each grade-group session. Each section in the handbook also indicates the day and time for that specific section. In addition to instructions, each section provides the facilitator with information about the purpose of the activity, the role of each of the content and process facilitators, and the specific task for panelists in that session. The materials used in the session are also listed. The description of specific materials to be used for the first time in that session is also included. The facilitator handbook used at the operational ALS meeting is in Appendix D. The handbook, meeting agenda, and briefing booklet were constructed in a manner that allowed for easy cross-referencing.

For the pilot study and operational ALS meeting, facilitator training included a staff meeting the day before the meeting. The facilitation staff met with the CoSS, the COR, and the lead psychometrician to go over each step in the process and the instructions to be given to the panelists. Prior to the operational ALS meeting, the CoSS went over the modification to the pilot study implementation to clarify changes to be implemented for the operational ALS. In each of these meetings, the Facilitation Standards, which are included in the handbook, were reviewed. These are presented in Figure 10.

*Figure 10: Facilitation Standards*

To ensure that the ALS process is implemented as planned and intended in order to maintain procedural validity and security, the following guidelines are to be observed in the grade level panel rooms. It is the process facilitators' responsibilities that the guidelines are observed.

1. Panelists interact with each other and the facilitators only; there will be no interaction between the panelists and observers in the panel room at any time.

2. A facilitator must be present at all times.

3. Panelists are not allowed to use any means of electronic communications; cell phones may only be used outside the panel rooms during breaks. Panelists may not use any electronic device in the panel room other than the computers provided.

4. All materials distributed in the panel rooms are collected and/or accounted for by the Process Facilitator at the end of the day.

5. Facilitators, while working hand in hand, should maintain distinct and separate responsibilities; Content Facilitators are responsible for all matters related to NAEP writing including Achievement Levels Descriptions (ALDs), while the Process Facilitators are in charge of the implementation of every panelist task and keeping the panelists on schedule. Please see "Facilitator Roles" in each session description in the handbook.

6. When in doubt, please consult with the CoSS.

7. Be ready to attend a debriefing meeting shortly after the scheduled adjournment for each day. The meeting will be no longer than one hour, could be as short as 15 minutes. Please take this into consideration when making evening plans.

8. Attend a daily meeting at the beginning of each day, half an hour before the first session begins.

9. Attend a debriefing meeting within two weeks of the study.

Additionally, at the beginning of each day and prior to the panel meetings, the CoSS reviewed with the facilitation staff the processes that would be implemented for that day. PowerPoint presentations with instructions to the panelists were distributed to the process facilitators. The contents of the PowerPoint slides were taken directly from the Facilitator Handbook. These PowerPoint presentations are included as Appendix E.

### 2.9.8 Panelist Training

Panelist training was designed to prepare panelists to properly perform their tasks. Training was also designed to ensure that panelists understood the BoW procedures and writing NAEP and felt comfortable about the training and instructions. Sufficient time for training was designed to enhance procedural validity and the ability of panelists to make informed judgments that would result in achievement levels that are reasonable, valid, and informative to the public.

Sending advanced materials to panelists is considered the first step of their training (Cizek & Bunch, 2007; Raymond & Reid, 2001). Materials sent to panelists included the 2011 NAEP Writing Framework, the ALDs, the meeting agenda, and a briefing booklet. The briefing booklet describes all the steps in the process and includes the rationale and time allotted for each. During the meeting, panelists were repeatedly directed to the briefing booklet when they needed clarification on aspects of the process. A copy of the briefing booklet for the operational ALS meeting is in Appendix F.

The on-site panelist training provided during the operational ALS meeting was consistent with the training provided to the panelists for the 1998 NAEP ALS meetings, which exemplified a thorough training program for standard-setting panelists (Raymond & Reid, 2001) with modifications made as necessary to address needs specific to the standard-setting method. BoWTIE training was included in each aspect of the implementation process.

Panelist training occurred in both whole-group sessions and grade-groups sessions. All whole-group sessions were facilitated by the CoSS, and as mentioned earlier, all grade-group activities were facilitated by the grade-level process and content facilitators. The PowerPoint presentations used by the CoSS in general sessions are in Appendix E.

#### *2.9.8.1 General Orientation*

On-site training began with a general orientation to the NAEP program and the role of the Governing Board. An overview of the NAEP program was presented by the Governing Board COR. This overview was followed by a general introduction to the NAEP ALS process, which emphasized the steps related to the overall process that had already taken place and the steps that would occur after the conclusion of the operational ALS meeting. The intent was to provide the panelists as much context as possible so that they would be well informed when they started the BoW classification task. This information was provided to all panelists at the same time, ensuring that grade-level panels were provided the same training to the greatest extent possible.

#### *2.9.8.2 Taking a NAEP Exam*

Given that the goal of the ALS process is to determine what students should know and be able to do, it is logical for the panelists to become familiar with how students experienced the assessment. Early in the process, each panelist took a form of the NAEP at the appropriate grade level. This step was implemented in the grade-level group. This was the panelists' first exposure to the 2011 NAEP writing assessment. Panelists were given a brief orientation to the assessment form and manner of administration. Because the 2011 writing NAEP was administered by computer, panelists also took the assessment on the computer. Through Westat, the contractor for NAEP sampling and administration, NCES made computers available for the panelists to take a form of the writing assessment. These were the same computers used to administer the writing NAEP to grades 8 and 12 students in January through March in 2011. The test-taking computers also were used by the panelists to view the other tasks in the assessment.

After taking a form of NAEP writing, each panelist reviewed his or her own responses using the same scoring rubrics used by the operational scorers[12]. They were instructed to review their responses using scoring guides, although their tests would not be scored or used in any other way.

The form selected for this part of the training had two writing tasks that were marked for release. This form was common to both groups, A and B.[13] Exemplar BoWs were also selected from this form. It was important that panelists become familiar with this assessment form.

Fulcrum, a NAEP contractor for NCES, made a modification to the test-taking application to support this process. The modification ensured that the application behaved the same way for both students and panelists while they were taking the assessment, but the completed assessments were handled differently for panelists and students. Instead of being encrypted and sent for scoring, the responses written by panelists were saved in a Word file so that panelists could access their responses for self-scoring their performance. This software modification afforded panelists the ability to review their responses against the scoring rubrics, and was an important part of the panelists' training as they began to become familiar with the assessment.

[12] Pearson Assessment, NCES contractor for scoring the 2011 NAEP writing

[13] Details on how the form was selected are included in the description of form selection in the Technical Report section 2.3.1.

#### *2.9.8.3 Review of Writing Tasks and Scoring Rubrics*

To continue to become familiar with the tasks in the writing assessment, panelists reviewed each writing task in the forms from which they would classify student responses. Given the importance of closely replicating the student's experience of the NAEP administration for the panelists, panelists viewed each NAEP writing task on the same NAEP laptops used by students. For this part of the process, Fulcrum modified the test-taking application the following ways:

- A test package that had all 22 tasks was created for each group.
- The 22 tasks were ordered differently for each group. The specific ordering for each group followed the schema below:
  - The first 14 tasks were those in the group's task pool.
  - Tasks belonging to a selected form were presented consecutively and in the order that they appeared in that form.
  - The first three forms were common to the two groups.
  - The first form contained two tasks that had been selected to be released to the public when the Nation's Report Card is released.
- By clicking "Next," panelists were able to proceed to the next task without waiting for the 30-minute student response time to elapse.
- A "Back" button was added to the application so that panelists could conveniently access each task.[14]

During the task review, panelists were provided a list of the 14 tasks assigned to their group. The *Form Task Map* for each group is included in Appendix G. For a full review of each task, they were also provided the scoring guides used by the operational

[14] This functionality was requested after the field trial. For panelists to access already viewed tasks, the original application required that the panelists click "Next" until they reached the last item and then rerun the application and view and listen to the instructions again. The new functionality contributed tremendously to the efficiency of the process as well as reduced the amount of frustration experienced by panelists.

scorers. These are also included in Appendix G. Note that there is only one scoring guide for each purpose for writing. Further, for tasks *to convey experience,* the scoring guide applied to both grades 8 and 12.

Panelists were told that they may also review the remaining eight writing tasks for their grade level. These tasks were not listed in their *Form Task Map.*

#### 2.9.8.4 Orientation to the Method

Panelists once again came together as a group to train in the standard-setting method. Panelists were given an overview of the BoW method as well as an overview of the steps in the process, such as the rounds of classifications and the feedback provided after each round. The goal was to provide the panelists with information to help them understand their task without distracting them with too many details. The PowerPoint presentation for this general session is provided in Appendix E, which includes all the PowerPoints from general sessions.

#### 2.9.8.5 Presentation of NAEP Writing Framework and Achievement Levels Descriptions

Just as it was important that the panelists understand the assessment, it was important that they understand its framework. The framework is the ultimate source of information about the assessment. The content facilitators delivered a whole-group presentation on the framework. The goal of the presentation was to inform the panelists about the framework and achievement levels in a manner that would contribute to the confidence of the panelists and the integrity of the process. The PowerPoint used by the content facilitators in this presentation is also included in Appendix E.

Training on the ALDs continued in each grade-level group, where content facilitators encouraged panelists to ask questions specific to their grade level. To aid in the discussion, content facilitators used real student responses from a task marked for release to illustrate aspects of a student's response that demonstrated a knowledge, skill, or ability (KSA) specified in the ALDs. After the discussion, panelists were ready to participate in exercises intended to develop a common understanding of the ALDs.

#### *2.9.8.6 Achievement Level Training by Response Classification Exercise*

In order for the grade-level panelists to provide BoW classifications that would yield a set of reliable and valid cut scores, it was imperative that they gain a common understanding of the ALDs. In the absence of a collectively shared understanding of what students should know and be able to do, the cut scores resulting from the process would have no valid interpretation. The Response Classification Exercise was designed to help panelists continue to gain a common understanding of the ALDs. In this exercise, panelists applied their understanding of the ALDs to sample student responses for three writing tasks—one task for each purpose for writing.

For each of the three released tasks—one for each writing purpose—one student response was selected for each of the six score levels. A total of 18 student responses were used in this activity. The responses were randomly ordered relative to their scores, with the six sample responses for each task grouped together. A list revealing the scores of the sample student responses was provided only after all responses had been classified to achievement levels and the panelists had discussed their classifications based on their understanding of the ALDs. The specific instructions given to the panelists were as follows:

1. Examine each response and note the knowledge, skills, and abilities (KSAs) demonstrated in the response. Compare these KSAs with the ALDs. Using your understanding of the ALDs, classify the response into one of the following levels: Below Basic, Basic, Proficient, and Advanced.

2. After you have classified all of the 18 responses independently, discuss your classification with your colleagues. It is not necessary to agree on the classifications. Consensus is a goal but not a requirement. The important part of the exercise is that you discuss your classifications to gain deeper understandings of the ALDs and to become familiar with how your group approaches the classification task.

3. After you have classified all responses and have discussed all of your classifications with your colleagues, look at the rubric scores (from the score sheet) given to the responses that you classified at each level. Discuss the relations between the rubric scores of student responses and their achievement levels classifications, including why student responses with different scores may be classified into the same achievement level.

Reviewing the scores given to the sample responses helped panelists understand that there is not always a direct correspondence between the scores assigned to performances and their judgment of the achievement level represented by the response.

The exercise described above, called Paper Classification Exercise was implemented for ALS processes for previous NAEP writing which was paper-and-pencil. The paper selection was implemented for the 1992 ALS process as the way of making judgments about constructed response tasks (ACT, Inc., 1993). While it was never again used for setting cut scores, it was used as a training exercise in subsequent NAEP ALS processes. A modified Angoff method was used to set the NAEP achievement levels in 1998, and a paper selection process was used as a training exercise for that process. (Loomis & Hanick, 2000).

#### *2.9.8.7 Training in the Rating Method*

Training in the rating method prepared panelists for a major step in the ALS process—rounds of classifications and feedback. Great emphasis was placed on standardizing the instructions to the panelists across the two grade groups. Consistent with this effort, training in the rating method started with a general session to explain the BoW classification task and how to implement the software.

The general session was also used to explain to panelists how the 50 BoWs for the classification task were selected. The two main criteria for selection were as follows:

- uniformity across the forms—from each of the seven forms assigned to a group, seven or eight BoWs were selected
- uniformity across the scaled score range—the number of BoWs in any interval of the same length is about the same

BoWs with a total raw score of 0, 1, or 2 were excluded from the selection.[15]

Because the first round of classification was also the first time that the panelists used a major BoWTIE functionality, use of BoWTIE in the classification process was demonstrated at the general session. When the panelists reconvened in the grade room, they had an opportunity to practice using BoWTIE in a training round.

The facilitator led the panelists in a practice round to help the panelists become familiar with the classification task. A common form with two tasks was included in the rating pool for each group (previously explained in section 2.9.4).  A sample of six BoWs was selected from the common form for use in the practice session. . The BoWs

[15] The Technical Report section 2.3.2 includes further details on the selection of the 50 BoWs.

in the sample were rank-ordered from highest to lowest score. Panelists reviewed the BoWs with the facilitator and discussed them as a group. The facilitator led the panelists through the practice classification including using BoWTIE to access the responses and record their KSA comments before they select an achievement level classification for each of the six BoWs.

### 2.9.9 Body of Work Classification and Feedback

For each round of classification and feedback, panelists were introduced to tasks and given instructions as a whole group, and they were given individualized instructions and information in their smaller grade groups.

Classification data were collected for each round and were analyzed so that results and feedback information could be given to the panelists to inform their ratings for the subsequent rounds. Feedback was given to the panelists to guide their judgments and provide indications of how they may wish to adjust their judgments in subsequent rounds. Further, feedback "provides evidence for the quality of the conduct of the process as well as a direct indication that the standard setters considered relevant information when participating in the process" (Reckase, 2001, p. 161).

#### 2.9.9.1 Round 1 Classifications: Rangefinding Set 1

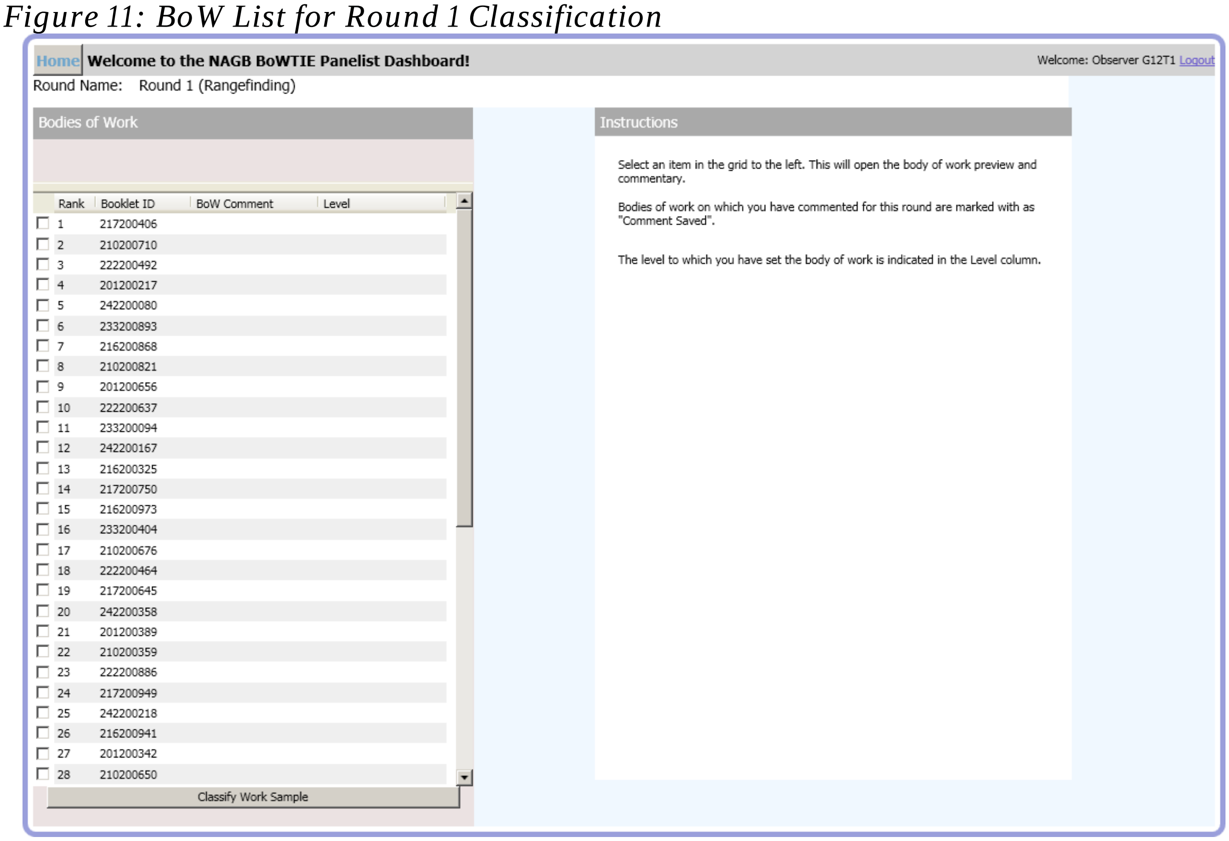
In each round of classifications, panelists were instructed to use their best judgment in classifying BoWs for setting individual cut scores. Their judgments were to be based on the correspondence between the performance demonstrated in each BoW and the ALDs.

During this round, panelists examined 50 BoWs that were distributed across the full score range. Each panelist's classifications of all 50 BoWs were used for computing that panelist's Basic, Proficient, and Advanced cut scores. The BoWs were presented in

BoWTIE in order of highest to lowest performance based on estimated scores.[16] Panelists were informed that even though the BoWs were presented from highest to lowest score, they could classify them in any order.

Figure 11 shows the list of BoWs that were to be classified by panelists in Round 1. Each BoW was accessed through the click of a mouse. Figure 12 shows what a panelist first saw when accessing a BoW: the first writing task for that first BoW. Clicking the response button displayed the student response to that task, as shown in Figure 13. Clicking the radio button for Task 2 showed the second task, and clicking the corresponding response button allowed the panelist to access the response to the second task.

*Figure 11: BoW List for Round 1 Classification*



**Welcome to the NAGB BoWTIE Panelist Dashboard!**

Welcome: Observer G12T1

Round Name: Round 1 (Rangefinding)

Bodies of Work

| | Rank | Booklet ID | BoW Comment | Level |
|---|---|---|---|---|
| □ | 1 | 217200406 | | |
| □ | 2 | 210200710 | | |
| □ | 3 | 222200492 | | |
| □ | 4 | 201200217 | | |
| □ | 5 | 242200080 | | |
| □ | 6 | 233200893 | | |
| □ | 7 | 216200868 | | |
| □ | 8 | 210200821 | | |
| □ | 9 | 201200656 | | |
| □ | 10 | 222200637 | | |
| □ | 11 | 233200094 | | |
| □ | 12 | 242200167 | | |
| □ | 13 | 216200325 | | |
| □ | 14 | 217200750 | | |
| □ | 15 | 216200973 | | |
| □ | 16 | 233200404 | | |
| □ | 17 | 210200676 | | |
| □ | 18 | 222200464 | | |
| □ | 19 | 217200645 | | |
| □ | 20 | 242200358 | | |
| □ | 21 | 201200389 | | |
| □ | 22 | 210200359 | | |
| □ | 23 | 222200886 | | |
| □ | 24 | 217200949 | | |
| □ | 25 | 242200218 | | |
| □ | 26 | 216200941 | | |
| □ | 27 | 201200342 | | |
| □ | 28 | 210200650 | | |

Instructions

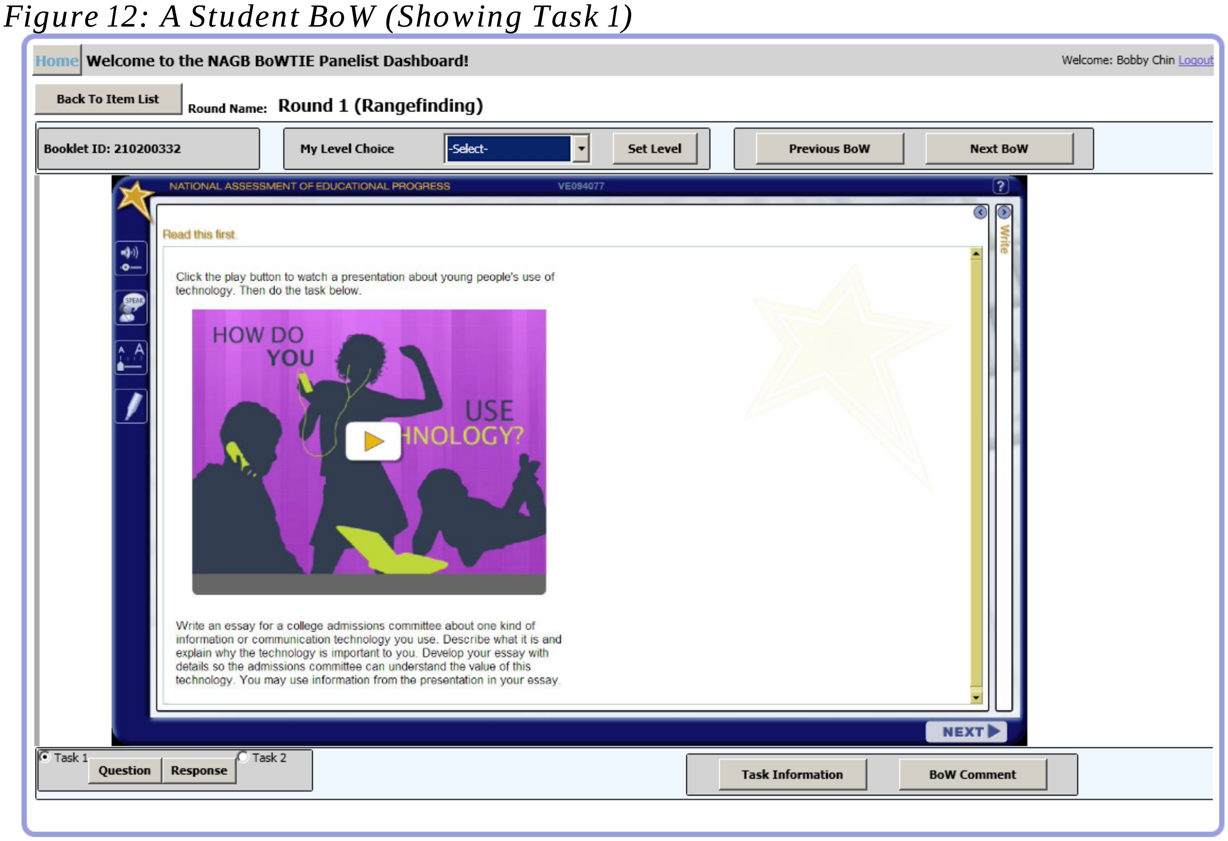
Select an item in the grid to the left. This will open the body of work preview and commentary.

Bodies of work on which you have commented for this round are marked with as "Comment Saved".

The level to which you have set the body of work is indicated in the Level column.

[16] Details on the computation of *expected a posteriori* scores are provided in the Technical Report section 2.2.

*Figure 12: A Student BoW (Showing Task 1)*



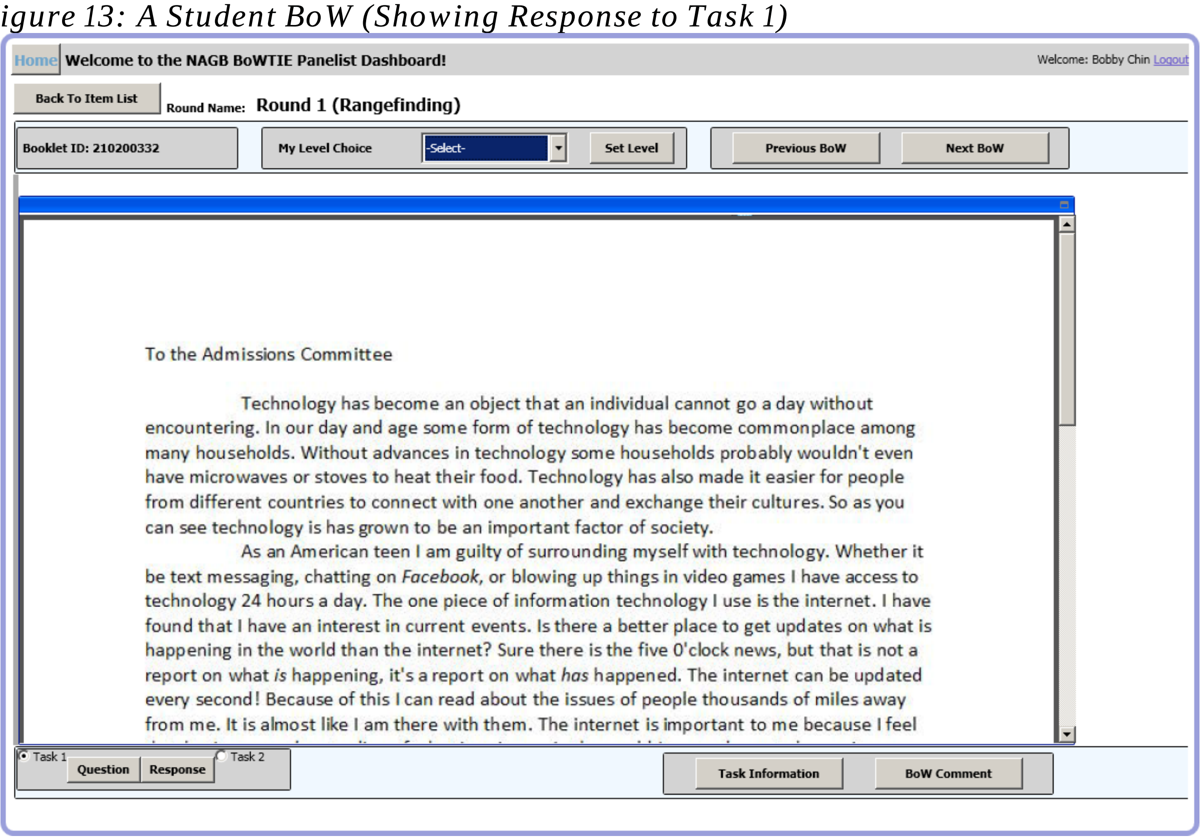
Home
Welcome to the NAGB BoWTIE Panelist Dashboard!
Welcome: Bobby Chin Logout
Back To Item List
Round Name: Round 1 (Rangefinding)
Booklet ID: 210200332
My Level Choice
-Select-
Set Level
Previous BoW
Next BoW
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS
VE094077
Write
HOW DO YOU USE TECHNOLOGY?
NEXT
Task 1
Question
Response
Task 2
Task Information
BoW Comment

Read this first.

Click the play button to watch a presentation about young people's use of technology. Then do the task below.

Write an essay for a college admissions committee about one kind of information or communication technology you use. Describe what it is and explain why the technology is important to you. Develop your essay with details so the admissions committee can understand the value of this technology. You may use information from the presentation in your essay.

*Figure 13: A Student BoW (Showing Response to Task 1)*



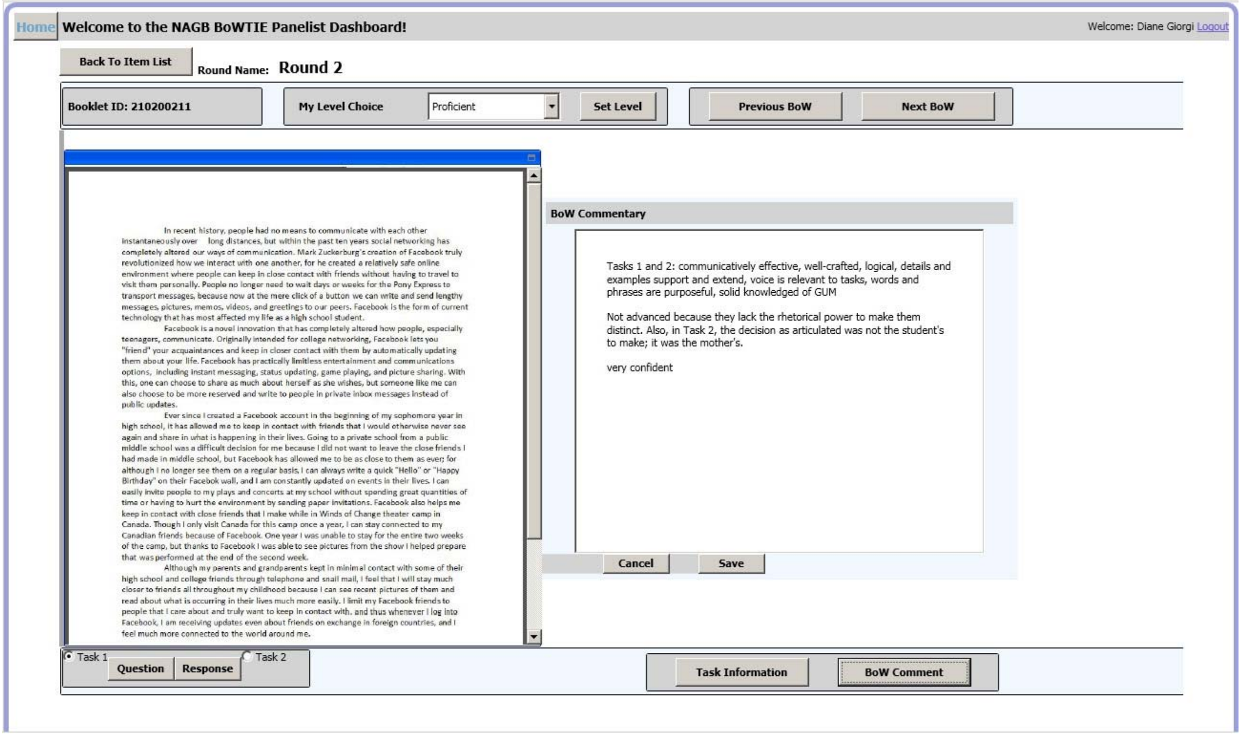
To the Admissions Committee

Technology has become an object that an individual cannot go a day without encountering. In our day and age some form of technology has become commonplace among many households. Without advances in technology some households probably wouldn't even have microwaves or stoves to heat their food. Technology has also made it easier for people from different countries to connect with one another and exchange their cultures. So as you can see technology is has grown to be an important factor of society.

As an American teen I am guilty of surrounding myself with technology. Whether it be text messaging, chatting on *Facebook*, or blowing up things in video games I have access to technology 24 hours a day. The one piece of information technology I use is the internet. I have found that I have an interest in current events. Is there a better place to get updates on what is happening in the world than the internet? Sure there is the five 0'clock news, but that is not a report on what *is* happening, it's a report on what *has* happened. The internet can be updated every second! Because of this I can read about the issues of people thousands of miles away from me. It is almost like I am there with them. The internet is important to me because I feel

Annotation functionality was available to panelists for making notes about each BoW, as shown in Figure 14. This functionality was accessible through the BoW Comment button on both the task screen (Figure 12) and the response screen (Figure 13). In Figure 14, note that there is only one comment window. Panelist comments, whether for the response to the first or second task, were saved together. This is because panelist classifications are based on the demonstrated performance on the two writing tasks combined. Panelists were told to treat their annotations as "notes to self" that they could later access when they were discussing their classification of specific BoWs. Panelists were encouraged not only to make notes for each response but to indicate in the notes how they would classify student performance based on a response. However, it was made clear that their final classification should be based on KSAs demonstrated in both responses and how well the responses matched the ALDs.

*Figure 14: BoW Comment*



Round Name: Round 2

Booklet ID: 210200211

My Level Choice: Proficient

In recent history, people had no means to communicate with each other instantaneously over long distances, but within the past ten years social networking has completely altered our ways of communication. Mark Zuckerburg's creation of Facebook truly revolutionized how we interact with one another, for he created a relatively safe online environment where people can keep in close contact with friends without having to travel to visit them personally. People no longer need to wait days or weeks for the Pony Express to transport messages, because now at the mere click of a button we can write and send lengthy messages, pictures, memos, videos, and greetings to our peers. Facebook is the form of current technology that has most affected my life as a high school student.

Facebook is a novel innovation that has completely altered how people, especially teenagers, communicate. Originally intended for college networking, Facebook lets you "friend" your acquaintances and keep in closer contact with them by automatically updating them about your life. Facebook has practically limitless entertainment and communications options, including instant messaging, status updating, game playing, and picture sharing. With this, one can choose to share as much about herself as she wishes, but someone like me can also choose to be more reserved and write to people in private inbox messages instead of public updates.

Ever since I created a Facebook account in the beginning of my sophomore year in high school, it has allowed me to keep in contact with friends that I would otherwise never see again and share in what is happening in their lives. Going to a private school from a public middle school was a difficult decision for me because I did not want to leave the close friends I had made in middle school, but Facebook has allowed me to be as close to them as ever; for although I no longer see them on a regular basis, I can always write a quick "Hello" or "Happy Birthday" on their Facebok wall, and I am constantly updated on events in their lives. I can easily invite people to my plays and concerts at my school without spending great quantities of time or having to hurt the environment by sending paper invitations. Facebook also helps me keep in contact with close friends that I make while in Winds of Change theater camp in Canada. Though I only visit Canada for this camp once a year, I can stay connected to my Canadian friends because of Facebook. One year I was unable to stay for the entire two weeks of the camp, but thanks to Facebook I was able to see pictures from the show I helped prepare that was performed at the end of the second week.

Although my parents and grandparents kept in minimal contact with some of their high school and college friends through telephone and snail mail, I feel that I will stay much closer to friends all throughout my childhood because I can see recent pictures of them and read about what is occurring in their lives much more easily. I limit my Facebook friends to people that I care about and truly want to keep in contact with, and thus whenever I log into Facebook, I am receiving updates even about friends on exchange in foreign countries, and I feel much more connected to the world around me.

Tasks 1 and 2: communicatively effective, well-crafted, logical, details and examples support and extend, voice is relevant to tasks, words and phrases are purposeful, solid knowledged of GUM

Not advanced because they lack the rhetorical power to make them distinct. Also, in Task 2, the decision as articulated was not the student's to make; it was the mother's.

very confident

Panelists indicated their final BoW classifications by using a dropdown menu, available in all the BoW screens shown in Figures 11 through 14, to select one of the four levels of Below Basic, Basic, Proficient, and Advanced. When panelists clicked the Save button, their BoW classification data were uploaded to the ALS database. Panelists were able to review and modify their classifications before submitting them for analysis.

Panelists were instructed to notify the process facilitator when they concluded their classifications. Functionality within BoWTIE allowed the facilitator to check whether each panelist had classified all 50 BoWs.

#### *2.9.9.2 Feedback from Round 1*

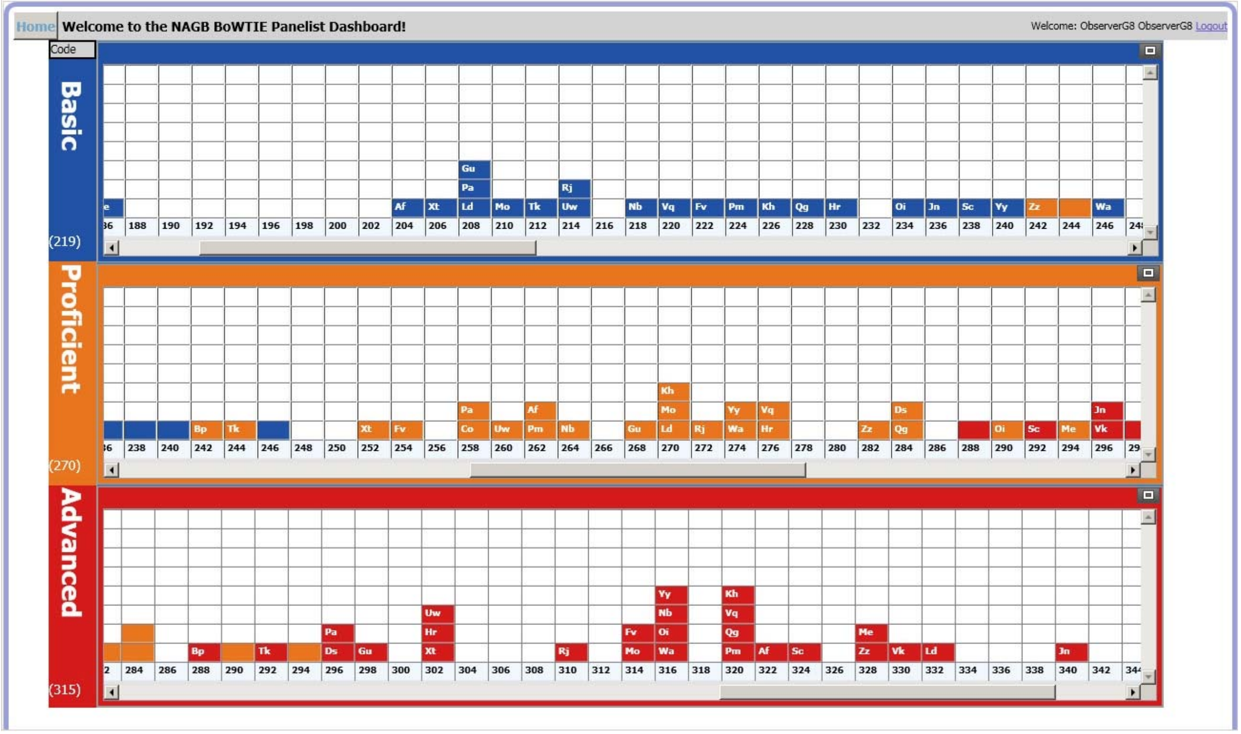
Grade-level cut scores, cut score distribution and location charts, and a tally of panelists' classifications from Round 1 were presented to panelists to inform their second round of classifications. Per the Governing Board's ALS tradition, feedback was presented to the panelists in the general session. Thus, each panel was informed of the cut scores of the other grade-level panel, although they were reminded that the cut scores should not be compared between grades because the score scales were unique to each grade level.

##### Cut Score Location Charts

The cut score location charts graphically present the distribution of cut scores set by panelists, where each panelist's cut score is indicated by a two-letter code known only to him or her. This feedback was presented in BoWTIE and provided interrater consistency information to panelists by showing where they set their cut scores in relation to cut scores of other panelists and the median cut score for the overall grade group. An example is provided in Figure 15.

One cut score location chart was produced for each achievement level, with three charts stacked together for presentation. Note that in Figure 15, each of the three charts corresponds to an achievement level cut score, with the median cut score shown on the left-hand side. On each chart, the horizontal axis was the pseudo-NAEP scale. The entire scale may be viewed by scrolling horizontally. Each square on the chart represents a cut score set by a panelist resulting from the logistic regression computation. A chart for a particular level identifies panelists' cut scores for that level, while the cut scores for other levels are not identified by code. The cut score location chart shows some other interesting results. For example, a panelist, identified as Wa, set his or her Basic cut score higher than other panelists' Proficient cut scores, as shown in Figure 15. In this example, panelist Wa's Basic cut score was 246, which was higher than the Proficient cut scores set by Bp and Tk, who set their cut scores at 242 and 244, respectively. The overlap in the ranges of the two cut score levels signaled that the panelists involved should revisit their understanding of the achievement levels.

*Figure 15: Cut Score Location Chart*



Home Welcome to the NAGB BoWTIE Panelist Dashboard!

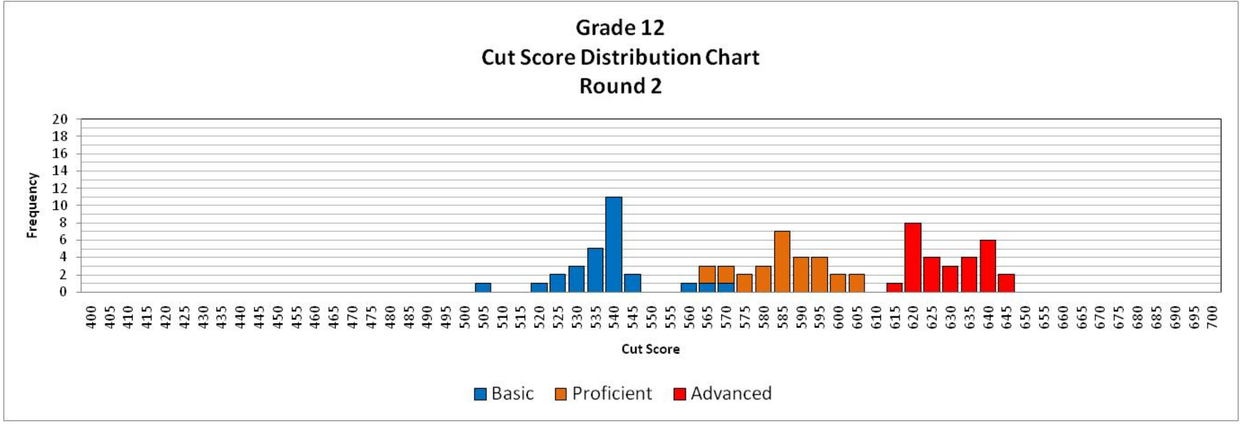
Welcome: ObserverG8 ObserverG8 Logout

## Cut Score Distribution Chart

Other feedback provided to the panelists after the first round included the cut score distribution chart, shown in Figure 16, which is a histogram of panelist cut scores. In a sense, this chart, which shows color-coded scores for all levels, provides the same information as the cut score location chart; however, it does not identify where individual panelists set their cut scores. Because the entire pseudo-NAEP scale is visible on the cut score distribution chart, it was easy for panelists to see where their cut scores were relative to the whole scale.

The cut score distribution chart was not presented in BoWTIE. The TACSS strongly recommended that this chart be included in the feedback presented to panelists after every round of classification. However, because this recommendation was made when it was too late to build this functionality into the software, the chart was shown on an overhead screen during the general session as well as during the grade-group sessions.

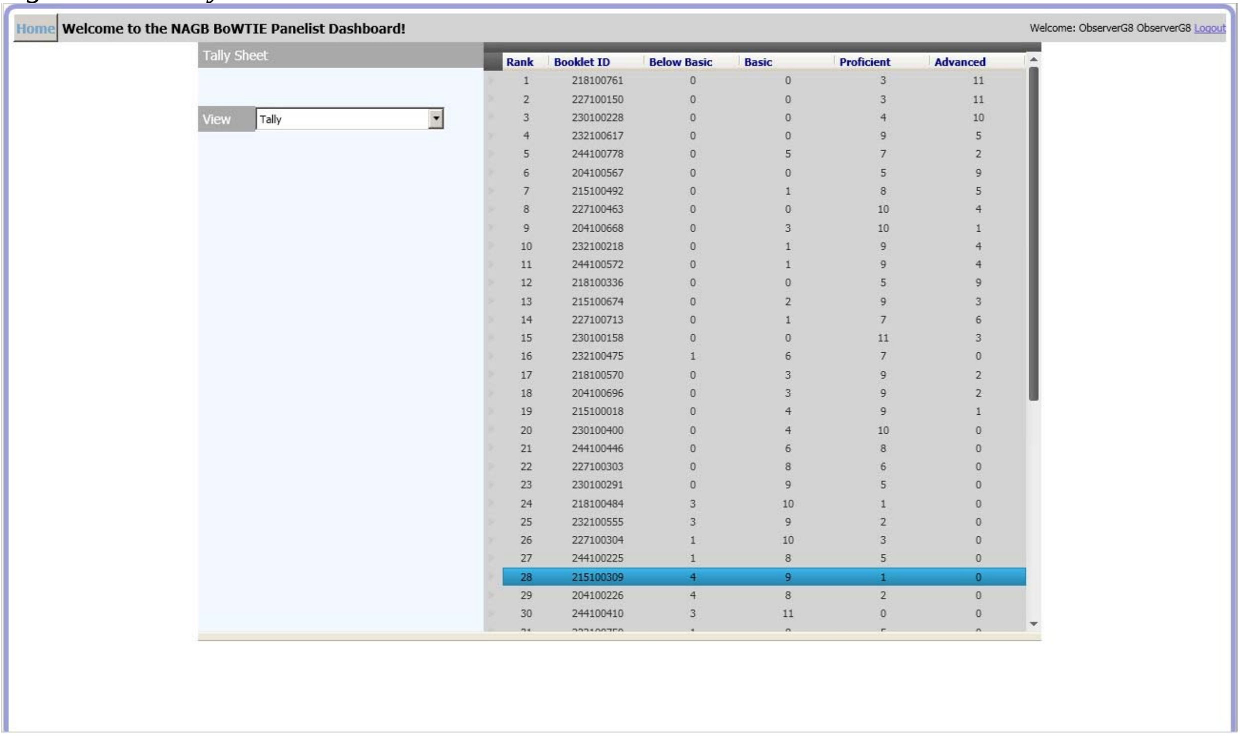
*Figure 16: Cut Score Distribution Chart*



### Tally and Discussion of Entropic Bodies of Work

The tally is a tabular presentation of the number of panelists that classified each BoW to a particular achievement level. Each panelist saw the tally for his or her group (A or B) as well as across groups for common BoWs. Figures 17 and 18 are sample tallies provided to the panelists through BoWTIE, where Figure 17 is group-specific and Figure 18 lists BoWs that were common to the two groups. Note that for the across-group tally of common BoWs, each panelist saw the within-group rank of each BoW.

*Figure 17: Tally*



Home Welcome to the NAGB BoWTIE Panelist Dashboard!

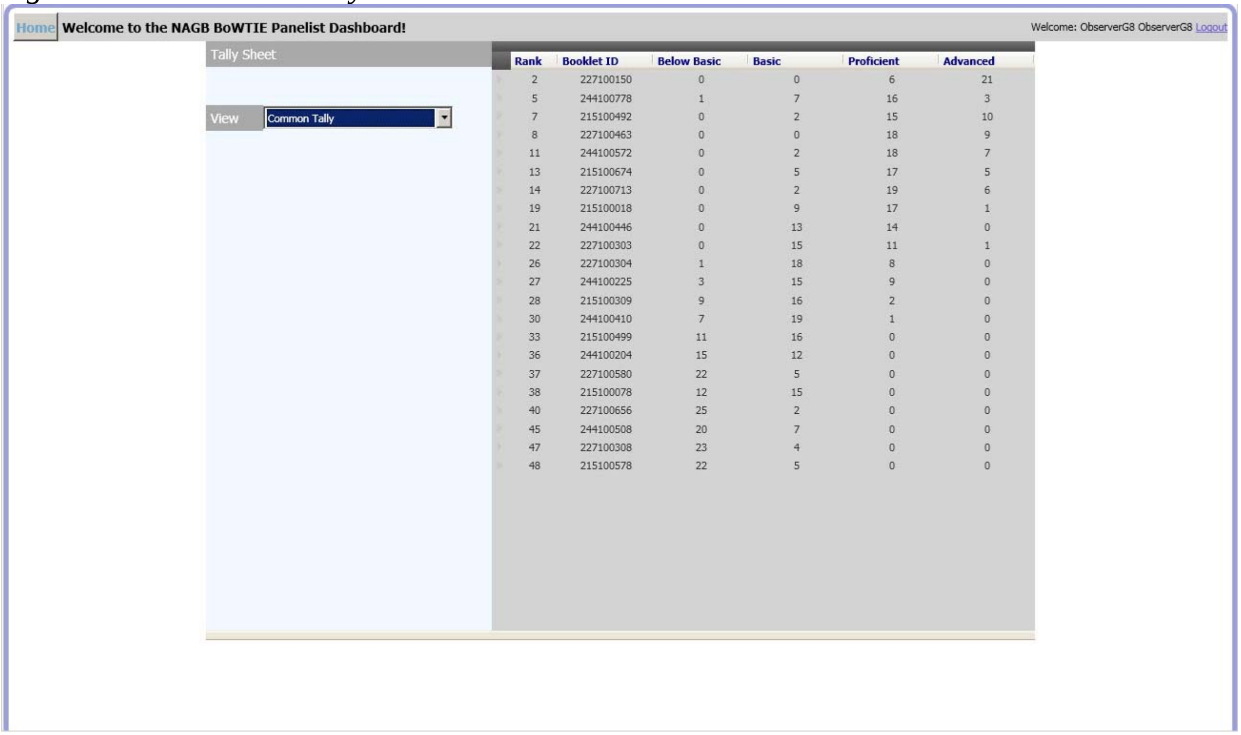
Welcome: ObserverG8 ObserverG8 Logout

Tally Sheet

View: Tally

| Rank | Booklet ID | Below Basic | Basic | Proficient | Advanced |
|---|---|---|---|---|---|
| 1 | 218100761 | 0 | 0 | 3 | 11 |
| 2 | 227100150 | 0 | 0 | 3 | 11 |
| 3 | 230100228 | 0 | 0 | 4 | 10 |
| 4 | 232100617 | 0 | 0 | 9 | 5 |
| 5 | 244100778 | 0 | 5 | 7 | 2 |
| 6 | 204100567 | 0 | 0 | 5 | 9 |
| 7 | 215100492 | 0 | 1 | 8 | 5 |
| 8 | 227100463 | 0 | 0 | 10 | 4 |
| 9 | 204100668 | 0 | 3 | 10 | 1 |
| 10 | 232100218 | 0 | 1 | 9 | 4 |
| 11 | 244100572 | 0 | 1 | 9 | 4 |
| 12 | 218100336 | 0 | 0 | 5 | 9 |
| 13 | 215100674 | 0 | 2 | 9 | 3 |
| 14 | 227100713 | 0 | 1 | 7 | 6 |
| 15 | 230100158 | 0 | 0 | 11 | 3 |
| 16 | 232100475 | 1 | 6 | 7 | 0 |
| 17 | 218100570 | 0 | 3 | 9 | 2 |
| 18 | 204100696 | 0 | 3 | 9 | 2 |
| 19 | 215100018 | 0 | 4 | 9 | 1 |
| 20 | 230100400 | 0 | 4 | 10 | 0 |
| 21 | 244100446 | 0 | 6 | 8 | 0 |
| 22 | 227100303 | 0 | 8 | 6 | 0 |
| 23 | 230100291 | 0 | 9 | 5 | 0 |
| 24 | 218100484 | 3 | 10 | 1 | 0 |
| 25 | 232100555 | 3 | 9 | 2 | 0 |
| 26 | 227100304 | 1 | 10 | 3 | 0 |
| 27 | 244100225 | 1 | 8 | 5 | 0 |
| 28 | 215100309 | 4 | 9 | 1 | 0 |
| 29 | 204100226 | 4 | 8 | 2 | 0 |
| 30 | 244100410 | 3 | 11 | 0 | 0 |

*Figure 18: Common Tally*



Home Welcome to the NAGB BoWTIE Panelist Dashboard!

Welcome: ObserverG8 ObserverG8 Logout

Tally Sheet

View: Common Tally

| Rank | Booklet ID | Below Basic | Basic | Proficient | Advanced |
|---|---|---|---|---|---|
| 2 | 227100150 | 0 | 0 | 6 | 21 |
| 5 | 244100778 | 1 | 7 | 16 | 3 |
| 7 | 215100492 | 0 | 2 | 15 | 10 |
| 8 | 227100463 | 0 | 0 | 18 | 9 |
| 11 | 244100572 | 0 | 2 | 18 | 7 |
| 13 | 215100674 | 0 | 5 | 17 | 5 |
| 14 | 227100713 | 0 | 2 | 19 | 6 |
| 19 | 215100018 | 0 | 9 | 17 | 1 |
| 21 | 244100446 | 0 | 13 | 14 | 0 |
| 22 | 227100303 | 0 | 15 | 11 | 1 |
| 26 | 227100304 | 1 | 18 | 8 | 0 |
| 27 | 244100225 | 3 | 15 | 9 | 0 |
| 28 | 215100309 | 9 | 16 | 2 | 0 |
| 30 | 244100410 | 7 | 19 | 1 | 0 |
| 33 | 215100499 | 11 | 16 | 0 | 0 |
| 36 | 244100204 | 15 | 12 | 0 | 0 |
| 37 | 227100580 | 22 | 5 | 0 | 0 |
| 38 | 215100078 | 12 | 15 | 0 | 0 |
| 40 | 227100656 | 25 | 2 | 0 | 0 |
| 45 | 244100508 | 20 | 7 | 0 | 0 |
| 47 | 227100308 | 23 | 4 | 0 | 0 |
| 48 | 215100578 | 22 | 5 | 0 | 0 |

From the common tally, a list of BoWs was selected to be discussed by the whole group. The general criteria for selection were (a) the diversity of classification provided by panelists and (b) the number of BoWs that could reasonably be discussed within the time allocated for this activity. The following criteria, applied hierarchically, were developed for selecting the specific BoWs:

- most levels—the greatest number of achievement levels into which the panelists classified a particular BoW
- most spread—the greatest distance between the lowest category and the highest category where a particular BoW was classified
- split—the classification of a BoW into two or more categories by approximately the same number of panelists
- reversal—the classification of a BoW by a majority of panelists that is inconsistent with the modal classifications of the BoWs around it

A more detailed explanation of the selection is provided in the Technical Report. About eight BoWs were selected for discussion in each grade. The list of BoWs selected for discussion was provided to the facilitators only.

Discussion of the selected BoWs was to enhance the panelists' common understanding of the ALDs. Panelists discussed the rationale for their classifications and tried to understand reasons for the differences. During the discussion, panelists had read-only access to their classifications and comments from the first round in BoWTIE.

The process facilitator led the first part of the discussion for the grade group, and the content facilitator provided expertise as needed. The discussion helped panelists to differentiate performance between adjacent achievement levels. After discussing student work samples from the common forms, panelists were encouraged to continue the discussion within each table group using BoWs from forms unique to the rating group. Panelists were encouraged to discuss BoWs for which they had a high rate of disagreement or specific student work samples that they wanted to review with other panelists at the table.

#### *2.9.9.3 Round 2 Classifications: Rangefinding Set 1*

During the second round of classifications, panelists were presented with the same set of BoWs, along with the classifications and comments they provided in Round 1. Their task was to provide an achievement level classification for each BoW, similar to their task in Round 1, but based on new information in the feedback from the first round of classifications. This may be viewed as an adjustment to the ratings provided in Round 1. Panelists could change their classifications for all, some, or none of the BoWs.

Panelists who wanted to change their cut scores were provided the following general guidelines:

- If you want to set a cut score higher, you should classify more BoWs into a lower achievement level. For example, to set the Proficient cut score higher, you should reclassify some BoWs from Proficient to Basic.
- If you want to set a cut score lower, you should classify more BoWs into a higher achievement level. For example, to set the Proficient cut score lower, you should reclassify some BoWs from Basic to Proficient.

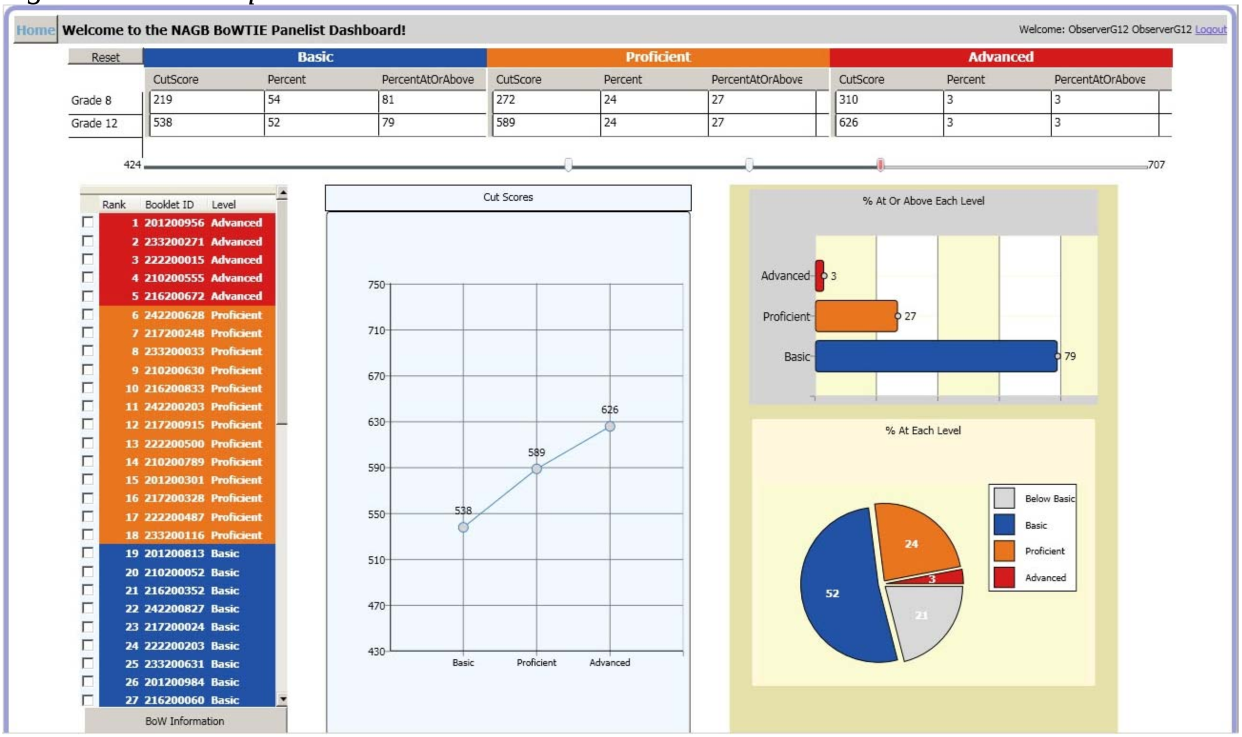
Panelists were reminded that their classifications should ultimately be based on the match between the ALDs and the KSAs demonstrated in each BoW. Just as in Round 1, they were to provide their classifications independently without discussion with other panelists.

#### *2.9.9.4 Feedback from Round 2*

The consequences data feedback informed the panelists of the proportion of student performances that would score at or above the cut score of each achievement level, based on the grade-level cut scores. If the proportion of examinees did not match individual panelists' expectations, based on the ALDs and their own experience with students, they were instructed to reexamine their ratings relative to the ALDs and determine if adjustments were needed. BoWTIE included an interactive mechanism for the consequences data feedback. Figure 19 shows a screen shot of this mechanism.

The consequences data feedback software is an interactive tool that instantaneously provides the new consequences data resulting from changing cut scores. The software calculates the percentage of students at each achievement level as well as highlights the student work samples associated with that achievement level.

*Figure 19: Consequences Data Feedback*



Welcome to the NAGB BoWTIE Panelist Dashboard!

Welcome: ObserverG12 ObserverG12 Logout

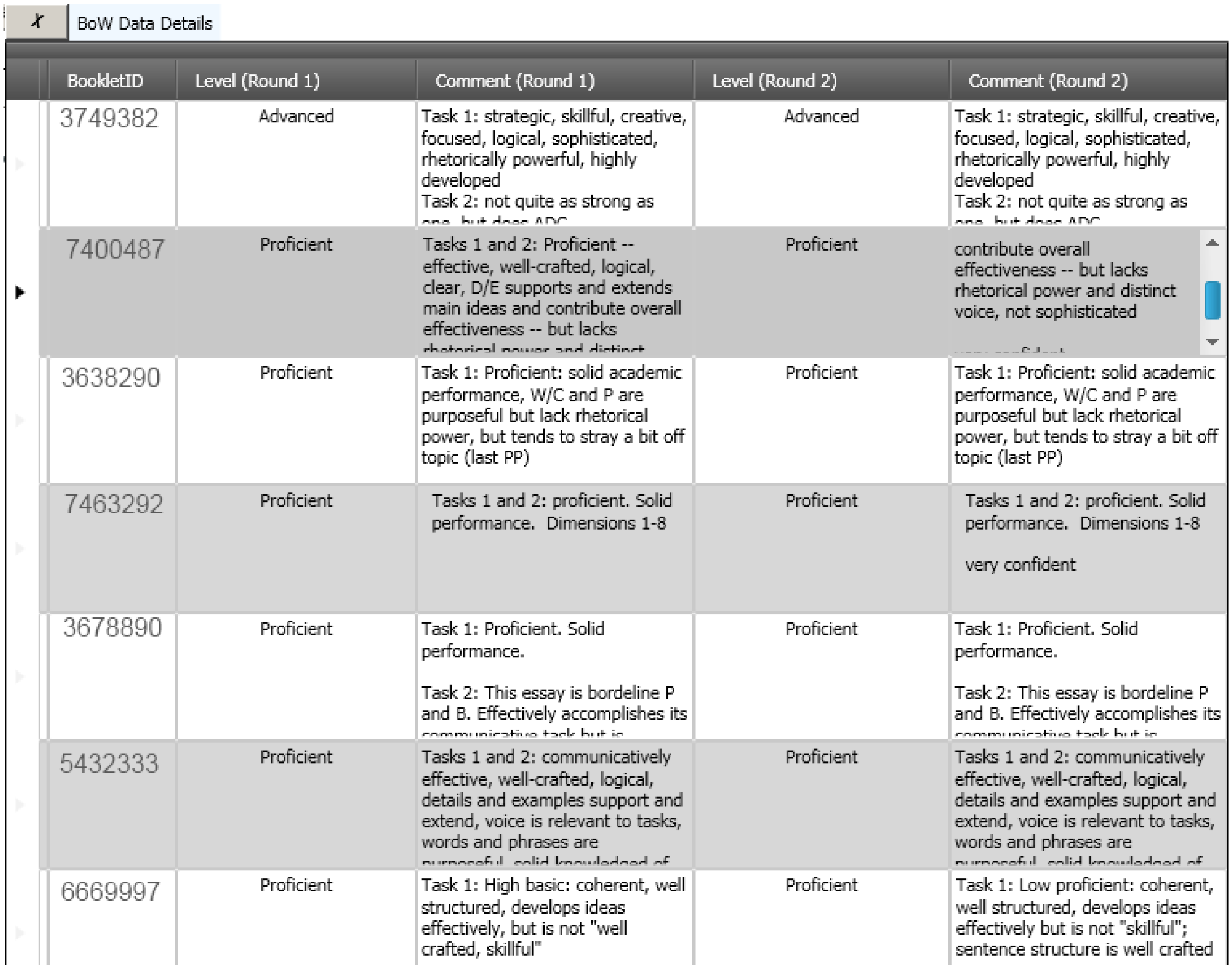
| Reset | Basic | | | Proficient | | | Advanced | | |
|---|---|---|---|---|---|---|---|---|---|
| | CutScore | Percent | PercentAtOrAbove | CutScore | Percent | PercentAtOrAbove | CutScore | Percent | PercentAtOrAbove |
| Grade 8 | 219 | 54 | 81 | 272 | 24 | 27 | 310 | 3 | 3 |
| Grade 12 | 538 | 52 | 79 | 589 | 24 | 27 | 626 | 3 | 3 |

| | Rank | Booklet ID | Level |
|---|---|---|---|
| □ | 1 | 201200956 | Advanced |
| □ | 2 | 233200271 | Advanced |
| □ | 3 | 222200015 | Advanced |
| □ | 4 | 210200555 | Advanced |
| □ | 5 | 216200672 | Advanced |
| □ | 6 | 242200628 | Proficient |
| □ | 7 | 217200248 | Proficient |
| □ | 8 | 233200033 | Proficient |
| □ | 9 | 210200630 | Proficient |
| □ | 10 | 216200833 | Proficient |
| □ | 11 | 242200203 | Proficient |
| □ | 12 | 217200915 | Proficient |
| □ | 13 | 222200500 | Proficient |
| □ | 14 | 210200789 | Proficient |
| □ | 15 | 201200301 | Proficient |
| □ | 16 | 217200328 | Proficient |
| □ | 17 | 222200487 | Proficient |
| □ | 18 | 233200116 | Proficient |
| □ | 19 | 201200813 | Basic |
| □ | 20 | 210200052 | Basic |
| □ | 21 | 216200352 | Basic |
| □ | 22 | 242200827 | Basic |
| □ | 23 | 217200024 | Basic |
| □ | 24 | 222200203 | Basic |
| □ | 25 | 233200631 | Basic |
| □ | 26 | 201200984 | Basic |
| □ | 27 | 216200060 | Basic |

The consequences data feedback in BoWTIE were accessible to individual panelists. The table at the top of the screen shows the cut scores and consequences data for each grade level. The rest of the display is grade specific. Immediately below this table is an interactive slider. Moving or sliding the cut score via the interactive slider resulted in the following changes:

- Percentage of students scoring at or above the new cut score.
- Percentages of students scoring in the adjacent levels. That is, the percentages of students in the Basic and Proficient levels both change when the Proficient cut score is moved. Changes in the percentage scoring at or above the cut score are reflected in the table of numerical results and on the bar graph. Changes to the percentage of students in each achievement level are reflected in the pie chart.
- Data in the table of numerical results and on the line graph are updated.
- Highlighting of the student work samples by achievement level classification, based on the new cut score locations. On the vertical display, the scale scores and work samples move relative to the highlighting when the cut score is moved. Easy access to the work samples at or around the cut scores helps the panelists maintain the necessary connection between the cut scores and the ALDs. See Figure 20 for an example.

*Figure 20: Notes on Some Selected BoWs Accessed Through the Consequences Data Feedback*



| BookletID | Level (Round 1) | Comment (Round 1) | Level (Round 2) | Comment (Round 2) |
|---|---|---|---|---|
| 3749382 | Advanced | Task 1: strategic, skillful, creative, focused, logical, sophisticated, rhetorically powerful, highly developed<br>Task 2: not quite as strong as one, but does ADC | Advanced | Task 1: strategic, skillful, creative, focused, logical, sophisticated, rhetorically powerful, highly developed<br>Task 2: not quite as strong as one, but does ADC |
| 7400487 | Proficient | Tasks 1 and 2: Proficient -- effective, well-crafted, logical, clear, D/E supports and extends main ideas and contribute overall effectiveness -- but lacks rhetorical power and distinct | Proficient | contribute overall effectiveness -- but lacks rhetorical power and distinct voice, not sophisticated<br>very confident |
| 3638290 | Proficient | Task 1: Proficient: solid academic performance, W/C and P are purposeful but lack rhetorical power, but tends to stray a bit off topic (last PP) | Proficient | Task 1: Proficient: solid academic performance, W/C and P are purposeful but lack rhetorical power, but tends to stray a bit off topic (last PP) |
| 7463292 | Proficient | Tasks 1 and 2: proficient. Solid performance. Dimensions 1-8 | Proficient | Tasks 1 and 2: proficient. Solid performance. Dimensions 1-8<br>very confident |
| 3678890 | Proficient | Task 1: Proficient. Solid performance.<br>Task 2: This essay is bordeline P and B. Effectively accomplishes its communicative task but is | Proficient | Task 1: Proficient. Solid performance.<br>Task 2: This essay is bordeline P and B. Effectively accomplishes its communicative task but is |
| 5432333 | Proficient | Tasks 1 and 2: communicatively effective, well-crafted, logical, details and examples support and extend, voice is relevant to tasks, words and phrases are purposeful, solid knowledged of | Proficient | Tasks 1 and 2: communicatively effective, well-crafted, logical, details and examples support and extend, voice is relevant to tasks, words and phrases are purposeful, solid knowledged of |
| 6669997 | Proficient | Task 1: High basic: coherent, well structured, develops ideas effectively, but is not "well crafted, skillful" | Proficient | Task 1: Low proficient: coherent, well structured, develops ideas effectively but is not "skillful"; sentence structure is well crafted |

### 2.9.9.5 Round 3 Classifications[17]: Rangefinding Set 2

To set the final cut scores, panelists classified a new set of 50 BoWs into the four achievement level categories. This new sample was selected using the same

[17] Per the Design Document, Round 3 was planned to be a pinpointing round. Part of the reason for developing BoWTIE was to support the selection of BoWs for pinpointing and producing the appropriate materials, which was the most logistically challenging part of a traditional implementation of the BoW method. The version of BoWTIE used for the pilot study had all the necessary functionalities related to pinpointing. Findings from the pilot study highlighted other challenges associated with pinpointing. A decision was made based on the recommendations of the TACSS to forgo the pinpointing round in favor of a replicate rangefinding for Round 3 classifications. Details of pinpointing will be discussed in the chapter about the pilot study.

methodology employed to select the first set. The only additional criterion for selection was that BoWs selected for the first set could not be selected for the second set.[18]

#### 2.9.9.6 Feedback from Round 3

Feedback from Round 3 was the same as Round 2 but with information reflective of Round 3 classifications. Again, the feedback included grade-level cut scores, consequences data, and cut score location and distribution charts.

### 2.9.10 Consequences Data Questionnaire

After the presentation and discussion of consequences data from Round 3, panelists were asked to fill out a questionnaire indicating whether they wanted to make additional changes to any of the cut scores after learning the consequences of those cut scores. Panelists were given the opportunity to recommend a change for any or all of the three cut scores, but they were aware that this was only provided as a recommendation to the Governing Board and would not result in a change to the computation of cut scores. Responses to the questionnaire provided more information to the Governing Board to inform their policy decision on the achievement levels.[19]

### 2.9.11 Selection of Exemplar Performance

Exemplar performances are part of the ALS reporting and one of the products of the ALS process. For writing, exemplar performances were recommended to represent student performance on an assessment form. This recommendation is based on its consistency with the method used in setting cut scores. BoWs were selected to exemplify what students know and can do at each of the achievement levels. Exemplar BoWs were selected from the NAEP writing form consisting of two marked-for-release

[18] Details of BoW selection are discussed in the Technical Report section 2.3.2.

[19] A copy of the Consequences Data Questionnaire is in Appendix I, along with all questionnaires used in the ALS process.

tasks. Note that for each of grades 8 and 12, this form was a common form for which eight BoWs were included in each of the rangefinding sets.

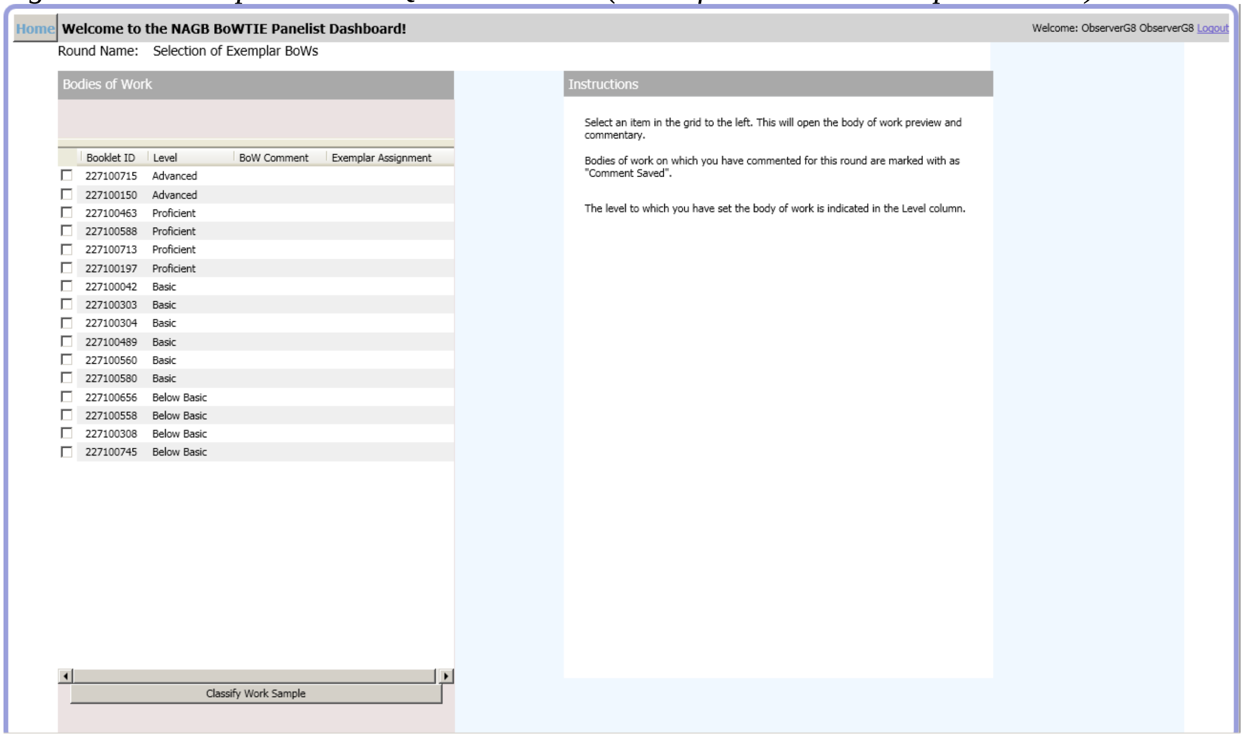
There were two stages in the process for selecting BoWs to exemplify performance at each achievement level. In the first stage, the BoWs eligible for selection at each level were identified, given the cut scores set in Round 3. In the second stage, panelists used their understanding of the ALDs to recommend BoWs to illustrate what students at each achievement level. Through BoWTIE, panelists were presented with the list of potential exemplar BoWs for each achievement level. Figure 21 shows an example of this list. The achievement level assigned to each BoW on the list is the level for which the BoW was a potential exemplar. As shown in Figure 22, when panelists accessed a BoW, the interface was similar to the rangefinding interface, with a window for writing comments and a dropdown menu for classification. Panelists were asked whether they would recommend each BoW as an exemplar for the achievement level indicated on the list. Their response choices were "Very Good," "Okay," or "Do Not Use." Panelists were encouraged to comment on each BoW, especially if they recommended that a specific BoW not be used as an exemplar. Panelists were also encouraged to discuss their selection with other panelists, but were told that their ratings should be made independently.

Data from the questionnaire were summarized and presented to the TACSS at a subsequent meeting. One BoW was selected for each grade for each level based on further criteria:

- At least 50% of the panelists rated it as "Very Good."
- Very few (i.e., three or fewer) panelists rated it as "Do Not Use."

Each of the above criteria was adjusted as needed to recommend the best exemplars to the Governing Board.

*Figure 21: Exemplar BoWs Questionnaire (List of Potential Exemplar BoWs)*



Home
Welcome to the NAGB BoWTIE Panelist Dashboard!
Welcome: ObserverG8 ObserverG8 Logout

Round Name: Selection of Exemplar BoWs

Bodies of Work

| | Booklet ID | Level | BoW Comment | Exemplar Assignment |
|---|---|---|---|---|
| ☐ | 227100715 | Advanced | | |
| ☐ | 227100150 | Advanced | | |
| ☐ | 227100463 | Proficient | | |
| ☐ | 227100588 | Proficient | | |
| ☐ | 227100713 | Proficient | | |
| ☐ | 227100197 | Proficient | | |
| ☐ | 227100042 | Basic | | |
| ☐ | 227100303 | Basic | | |
| ☐ | 227100304 | Basic | | |
| ☐ | 227100489 | Basic | | |
| ☐ | 227100560 | Basic | | |
| ☐ | 227100580 | Basic | | |
| ☐ | 227100656 | Below Basic | | |
| ☐ | 227100558 | Below Basic | | |
| ☐ | 227100308 | Below Basic | | |
| ☐ | 227100745 | Below Basic | | |

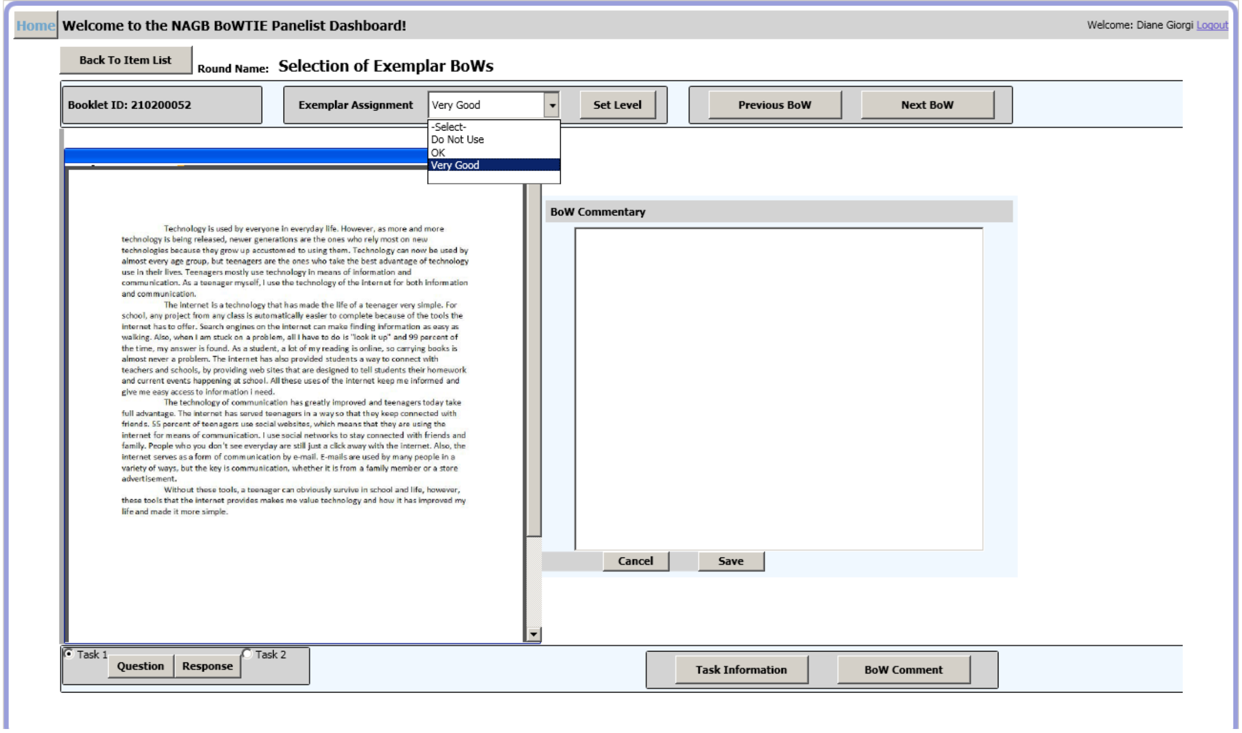
Instructions

Select an item in the grid to the left. This will open the body of work preview and commentary.

Bodies of work on which you have commented for this round are marked with as "Comment Saved".

The level to which you have set the body of work is indicated in the Level column.

*Figure 22: Exemplar BoW Questionnaire (Ratings and Comments Screen)*



Technology is used by everyone in everyday life. However, as more and more technology is being released, newer generations are the ones who rely most on new technologies because they grow up accustomed to using them. Technology can now be used by almost every age group, but teenagers are the ones who take the best advantage of technology use in their lives. Teenagers mostly use technology in means of information and communication. As a teenager myself, I use the technology of the internet for both information and communication.

The internet is a technology that has made the life of a teenager very simple. For school, any project from any class is automatically easier to complete because of the tools the internet has to offer. Search engines on the internet can make finding information as easy as walking. Also, when I am stuck on a problem, all I have to do is "look it up" and 99 percent of the time, my answer is found. As a student, a lot of my reading is online, so carrying books is almost never a problem. The internet has also provided students a way to connect with teachers and schools, by providing web sites that are designed to tell students their homework and current events happening at school. All these uses of the internet keep me informed and give me easy access to information i need.

The technology of communication has greatly improved and teenagers today take full advantage. The internet has served teenagers in a way so that they keep connected with friends. 55 percent of teenagers use social websites, which means that they are using the internet for means of communication. I use social networks to stay connected with friends and family. People who you don't see everyday are still just a click away with the internet. Also, the internet serves as a form of communication by e-mail. E-mails are used by many people in a variety of ways, but the key is communication, whether it is from a family member or a store advertisement.

Without these tools, a teenager can obviously survive in school and life, however, these tools that the internet provides makes me value technology and how it has improved my life and made it more simple.

### 2.9.12 Process Evaluations

At the end of the first day and after each round, panelists were provided with an evaluation form designed to assess their understanding of instructions, tasks, and materials. Five questionnaires were administered over the course of the operational ALS meeting. The schedule of the five evaluations is described in Table 16. Panelists accessed the process evaluation questionnaires through BoWTIE and panelists' responses were saved directly to the BoWTIE database. Most panelist responses to the evaluations were collected on Likert scales, but several responses were collected as narratives, to address specific aspects of the process. Copies of all process evaluation questionnaires are included in Appendix I.

*Table 16: Schedule of Process Evaluations*

| Evaluation | Schedule |
|---|---|
| 1 | End of Day 1 |
| 2 | End of Day 2 After Round 1 |
| 3 | Before Lunch on Day 3 After Round 1 Feedback |
| 4 | End of Day 3 After Round 2 Feedback |
| 5 | End of Day 4 After Round 3 Feedback and Before Debriefing |

Summary results of each evaluation questionnaire were made available to the facilitation staff as well as the COR and TACSS member observing the process shortly after all the data were collected for each questionnaire. These evaluations were reviewed at the end of each day and any sources of confusion, dissatisfaction, or other concerns were identified for clarification with individual panelists or the panel as a whole.

## 2.10 Special Study

The Governing Board requested that a special study be implemented to explore the relationship between performance on the 2011 assessment, based on the new writing framework, and performance on the 2007 assessment, based on the writing framework first implemented for the 1998 NAEP. The special study methodology implemented and described here was based on the recommendations from the first meeting of the TACSS and was not the study that was originally requested by the Governing Board. Discussions of the purposes of the special study led to the modification that the special study be implemented at the end of the pilot ALS process. Further, the design of the study was to address performance relative to the 2011 ALDs for both 2007 and 2011 assessments.

Measured Progress implemented the special study at the end of the pilot ALS study, as designed, with the participation of all the pilot study panelists. Findings from the pilot study led to evaluation and revision of the ALDs. Because the revision of the ALDs rendered the results of the special study moot, a second special study was implemented to explore the relationship between performance on the 2007 and 2011 writing assessments was implemented immediately following the operational ALS meeting. The need for these changes was evidenced after some of the panelists for the operational ALS meeting had been recruited. As a result, participation in the special study was not a requirement for participation in the ALS, and only a subset of the panelists from the operational ALS meeting was able to participate in the special study.

Changes in the writing NAEP, based on the framework first implemented in 2001, relative to the framework last implemented in 2007, required careful consideration. In addition to the revision of the ALDs and cut scores based on those ALDs, other elements of change included the transition from paper-and-pencil to computer-based administration, the prompts, the scoring rubrics, and the student population. As these changes prevented direct comparisons between performance on the 2011 and 2007 writing NAEP, a design that involves evaluating responses to the 2007 assessment relative to the 2011 ALDs was developed. This evaluation was planned and implemented in a separate round of rating after the ALS meeting.

For each of the pilot study and operational ALS meetings, panelists again came together as a group to be given an overview of the purpose of the special study and the steps in the process. At this point, the panelists had experienced extensive training on the BoW procedures. Therefore, the implementation was an abbreviated version with only the steps necessary for the special study: namely, taking the 2007 writing NAEP, becoming familiar with the 2007 prompts and rubrics, and classifying student work using the BoW methodology designed for the 2011 writing NAEP. It was important for panelists to take the 2007 writing NAEP because all NAEP ALS processes have included this step and it made the process for panelists the same as that designed for the ALS. Additionally, taking the test allowed them to experience the difference between taking the paper-and-pencil test (2007) and the computer-based version (2011). No feedback was provided to panelists following the classification task. The agenda for the special study is appended to the agenda of the pilot and operational ALS. The following subsections discuss the different features of the special study.

### 2.10.1 The 2007 Writing NAEP

The special study involved enough forms to ensure that all genre were represented in proportion to the framework specifications. All panelists reviewed the same test forms and the same work samples. Panelists were given a brief orientation to the test and the test-taking situation. Since the 2007 writing NAEP was not administered by computer, panelists took the paper form of the test and reviewed scanned images of handwritten student responses.

### 2.10.2 Review of Prompts and Scoring Rubrics

After panelists were introduced to the 2007 writing assessment, they reviewed all prompts and scoring rubrics for their grade level. To foster a common application of the 2011 ALDs to the 2007 prompts, panelists were presented with a training set of student work samples at each score level for three selected prompts, one for each genre.

### 2.10.3 Classification of Student Work

Panelists examined sets of BoWs from the 2007 administration distributed across the score range and classified them into the Basic, Proficient, and Advanced levels based on the 2011 ALDs. The BoWs were scanned images presented on the computer. BoW classifications were made in BoWTIE; thus, no further training in the classification process was provided.

### 2.10.4 Analyses

A series of cross-tabular analyses were conducted in an attempt to understand the relationship between the performance on the 2007 assessment using the 2007 achievement levels and performance on the 2007 assessment using the 2011 ALDs. The goal was to compare the achievement level classification from the reported 2007 results to the achievement level classification using the 2011 ALDs. Because each student is not assigned an official achievement level classification, Measured Progress ran the comparison based on classifications from the student's plausible values as well as the classification that would have arisen had the student been assigned a single EAP score. In addition, the classifications based on the 2011 ALDs were examined using the individual panelist classifications as well as the classifications that result when cut scores are calculated using logistic regression.

# Chapter 3—Field Trial

The first of the achievement levels-setting (ALS) meeting was a small-scale study for the primary purpose of testing the logistics of an entirely computer-based standard setting. The procedures implemented in the field trial were an abbreviated version of the procedures designed for setting the 2011 NAEP writing achievement levels. The field trial used a simplified, scaled-down version of the ALS sampling process to select a single panel of 20 for the two-day study. Table 17 presents the extent of abbreviation of the field trial relative to the operational ALS meeting.

*Table 17: Extent of Abbreviation of the Field Trial Relative to the ALS Process Planned for the Pilot Study*

| **ALS Process** | **Field Trial** |
|---|---|
| Four Days | Two Days |
| Orientation and Training | Abbreviated Orientation and Training |
| Evaluation 1 | |
| Round 1 (Rangefinding) | Round 1 (Rangefinding) |
| Evaluation 2 | Evaluation 1 |
| Round 1 Feedback | Round 1 Feedback |
| Evaluation 3 | Evaluation 2 |
| Round 2 (Rangefinding) | |
| Round 2 Feedback | |
| Evaluation 4 | |
| Round 3 (Pinpointing) | Round 2 (Pinpointing) |
| Round 3 Feedback | |

The field trial was conducted at an off-site hotel venue, following a series of in-house trials at Measured Progress and user acceptance testing of the Body of Work

Technological Integration and Enhancements (BoWTIE) software. The field trial allowed Measured Progress to emulate the 2011 procedures, identify logistical weaknesses, and adjust the procedures as necessary for further evaluation in the pilot study, where the exact procedures planned for the ALS operational sessions would be carried out.

Because part of the logistics being evaluated was the amount of time it took to implement each specific part of the process, it was deemed sufficient to implement the field trial for grade 12 only. Reading student responses is the most time-consuming part of any implementation of the Body of Work (BoW) method. There is an expectation that grade 12 responses are generally longer than grade 8 responses. Adjusting the time to accommodate reading grade 12 responses ensured that enough time would be allocated for both grades during the operational ALS meeting.

### 3.1 Field Trial Panelists

As described earlier, panelists for the field trial were recruited from within a 50-mile radius of the standard-setting site. Consistent with the demographic goals for the field trial panel, the selected panel was composed of 11 teachers, three nonteacher educators, and six general public panelists, or 55%, 15%, and 30%, respectively; additionally, 18 (90%) of the panelists represented public school districts and two (10%) represented private schools. This distribution is represented in Table 18.

*Table 18: Field Trial Panelist Distribution*

| Demographic Variable | Attribute | Grade 12 | | Goal |
|---|---|---|---|---|
| | | n | % | % |
| Panelist Type | Teachers | 11 | 55 | 55 |
| | Nonteacher Educators | 3 | 15 | 15 |
| | General Public | 6 | 30 | 30 |
| Publicity | Public | 18 | 90 | 90 |
| | Private | 2 | 10 | 10 |

Teacher panelists were classroom teachers who teach writing in secondary schools. Nonteacher educator panelists were curriculum specialists in writing and other educators with a background in writing; these educators were not active classroom teachers in grades K–12. Furthermore, faculty members in writing at public and private two-year and four-year postsecondary schools were considered nonteacher educator panelists. General public panelists were members of the general public who were in a position in their professional practice to evaluate writing samples such as reports and general memoranda. Specifically, they were not current or former educators. Included among the general public panelists were a town manager, an author, a university lecturer of English, and a communications business owner.

## 3.2 Logistics

The field trial focused on the logistical elements of the meeting that are directly impacted by the use of computers. In particular, our evaluation centered on five main elements: (a) hardware, (b) room configuration, (c) test administration and task review, (d) presentation of static information including the ALDs, and (e) presentation of student work samples. Although each of these logistical elements was tested prior to the field trial by an internal group at Measured Progress, the field trial served as an operational investigation of each element during an actual implementation of the ALS process.

Implementing a computer-based procedure involves transport of a large amount of computer equipment. This was the main reason for holding the field trial on a site near Measured Progress. The corporate headquarters for Measured Progress in Dover, New Hampshire was the original location selected for the field trial. However, during the first TACSS meeting on December 2–3, 2011, a recommendation was made to change the location for the field trial to an off-site location that would require implementation of procedures with conditions more similar to those likely to be encountered in the operational implementations of the process. The logistics being examined included transporting, setting up, and packing up the equipment, which included two computers to be used by each panelist. The field trial was first scheduled for July 9–10, 2011, in Portsmouth, New Hampshire. The field trial date was changed to late September so that actual operational student data and responses could be used. The item parameters used in estimating BoW scores as well as distribution information used for consequences data feedback were then considered only preliminary.

### *3.2.1 Hardware*

Because NAEP writing is a computer-based assessment (CBA) and the ALS process was also computer-based using BoWTIE, the standard-setting implementation required that each panelist use two laptop computers. It has always been the interest of the Governing Board that the logistics involved in having the panelists use two computers did not disrupt the ALS process.

For taking a form of NAEP writing, panelists used the actual computers used by students. NCES provided the laptops for use of panelists during the meeting. Westat, the contractor for NAEP administration, sent the CBA laptops directly to the meeting hotel. These laptops were already loaded with the form selected for panelists. The test-taking application was modified for the panelists so that they were able to access their responses and review them against the scoring rubrics as part of the ALS process. The laptops were also loaded with the rest of the writing tasks ordered per the specification provided by Measured Progress. All of the CBA laptops ran locally without any wired or wireless connection to any type of server or the Internet.

The ALS laptops were the same make and model as the CBA laptops, and also provided courtesy of NCES. Westat sent the ALS laptops ready to be configured to run BoWTIE. Measured Progress staff was responsible for the transportation, storage, and networking of the laptop computers. This included ensuring that appropriate equipment (e.g., CBA and ALS laptops, extension cords, Ethernet cords, routers) was available and working properly on-site. The Office of Information Technology (OIT) Infrastructure staff from Measured Progress configured the ALS laptops to be hardwired to a local server. Particular attention was given to the security of the laptops. Specifically, they were configured to allow access only to the standard-setting server. The intention was to limit distractions from e-mails, the Internet, and so forth, and eliminate security breaches. Each of the above elements was carefully examined by the appropriate Measured Progress staff (OIT and program management) to ensure optimal configuration for the achievement level setting.

In summary, the two computers and their specific uses were as follows:

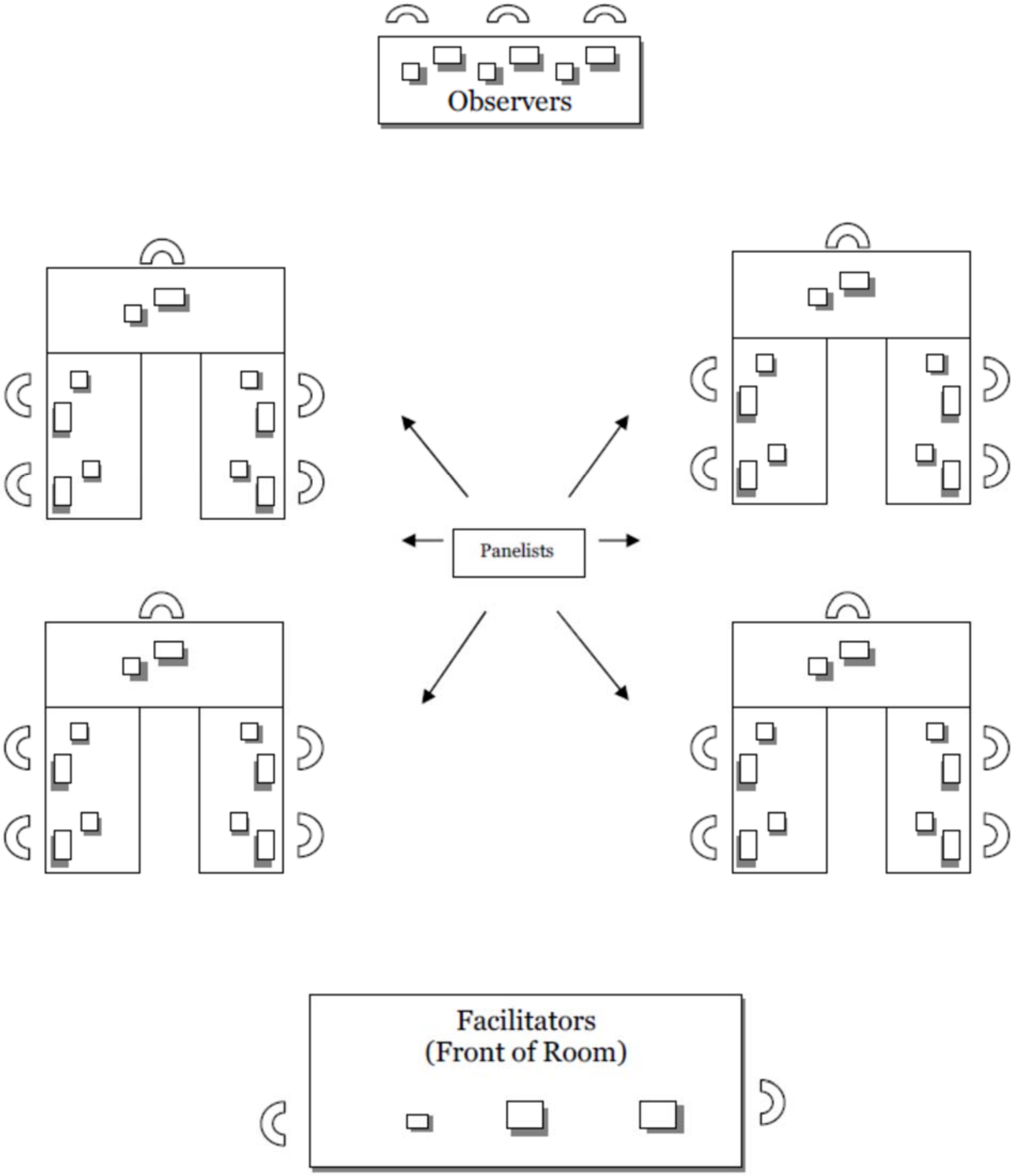
- CBA computer (Laptop 1)
    - taking a form of the assessment
    - viewing all NAEP writing tasks
- ALS computer (Laptop 2)

- classifications and comments
  - rangefinding
  - pinpointing
- feedback
  - tally
  - cut score location chart
  - consequences data
- process evaluation questionnaires

### 3.2.2 Room Configuration

A critical aspect of any ALS effort is the room configuration for the meeting. The room must be set up to facilitate discussion while simultaneously allowing the panelists to work independently. For this purpose, the room was configured with four sets of three six-foot tables arranged in an open square to seat five panelists per table with all panelists able to face the front of the room. Site visits were conducted to examine a mock setup of the rooms to verify that these specifications could be used. Figure 23 displays the room configuration diagram sent in advance of the site visit. In addition to the configuration of the tables, particular attention was given to the placement of all cords and cables. The safety of the room was of paramount importance. An extra day was scheduled prior to the beginning of the meeting to set up, configure, and test the computers and other equipment in advance. The number of staff and amount of time required to set up, test, and pack up the equipment was evaluated in order to optimize efficiency.

*Figure 23: Room Configuration for NAEP Writing Field Trial*



### 3.2.3 Test Administration and Task Review

It is critical to the success of standard setting to give panelists the opportunity to fully comprehend the test instrument and its administration to students. To achieve this goal, each panelist took a form of the NAEP writing assessment using a NAEP laptop computer. In preparation for this part of the process, the Chief of Standard Setting (CoSS), psychometrician, and a process facilitator observed a test administration of NAEP writing, each one at a different site. The goal was to ensure that panelists' experience of taking the NAEP replicated the students' experience to the extent possible.

To take a form of the assessment, panelists had to go through two levels of login: first to the Windows operating system and then to the student test-taking application. Once logged in, panelists then had to select their "school" and session. Lastly, panelists entered a 10-digit BoW ID to access the first part of the assessment, which was included in the multimedia directions for taking the test. Only after viewing the video for the directions were they able to get to the first writing task. This series of logins proved to be challenging for many panelists. Once logged in, they all appreciated being able to respond to the writing tasks using the word-processing tool included in the application.

The test-taking application did not allow the panelists to move on to the second writing task until the 30-minute response time limit had lapsed. Panelists who finished responding to the first task before the 30 minutes elapsed were allowed to take a break until they were able to start responding to the second task. Some panelists, on the other hand, were still writing when the period for responding to a task was over, so their essays ended midsentence. Later observations indicated that this was also the case for some high-achieving students.

Panelists were able to review their responses by opening a rich text format file for each of their responses. Panelists were provided printed scoring guides so they could roughly gauge how well they did. There were no logistical issues apparent during this part of the process.

To review the rest of the writing tasks assigned to each group, panelists went through the same login process and watched the directions video to get to the first task. They did not have to respond to a task or wait for 30 minutes to lapse before they could move on to the next task. If panelists wanted to return to a task or inadvertently clicked the “Next” button multiple times and skipped a task, they did not have a way of going back to that task other than going through the entire login process and beginning with the first task. This caused a lot of wasted time during the process as well as frustration on the part of the panelists.

### *3.2.4 Static Information*

The implementation of a computer-based process offers the distinct advantage of a paperless meeting. There was recognition, however, that some information, such as static information, may be better provided on paper. Information distributed to panelists is considered static if it does not change throughout the standard-setting process. More specifically, it is information that is independent of the cut scores: ALDs, *p*-value feedback, Reckase charts, and scoring guides. For the field trial, all of the above were planned to be presented to panelists on hard copy.

#### Achievement Level Descriptions

For the field trial, the ALDs were presented to panelists four different ways: (a) projected onto an overhead screen, (b) accessible on individual computer screens, (c) displayed on large posters around the room, and (d) printed on paper. Panelists completed an evaluation form with questions regarding the ease and preference for the display of static information.

### *p*-Value Feedback

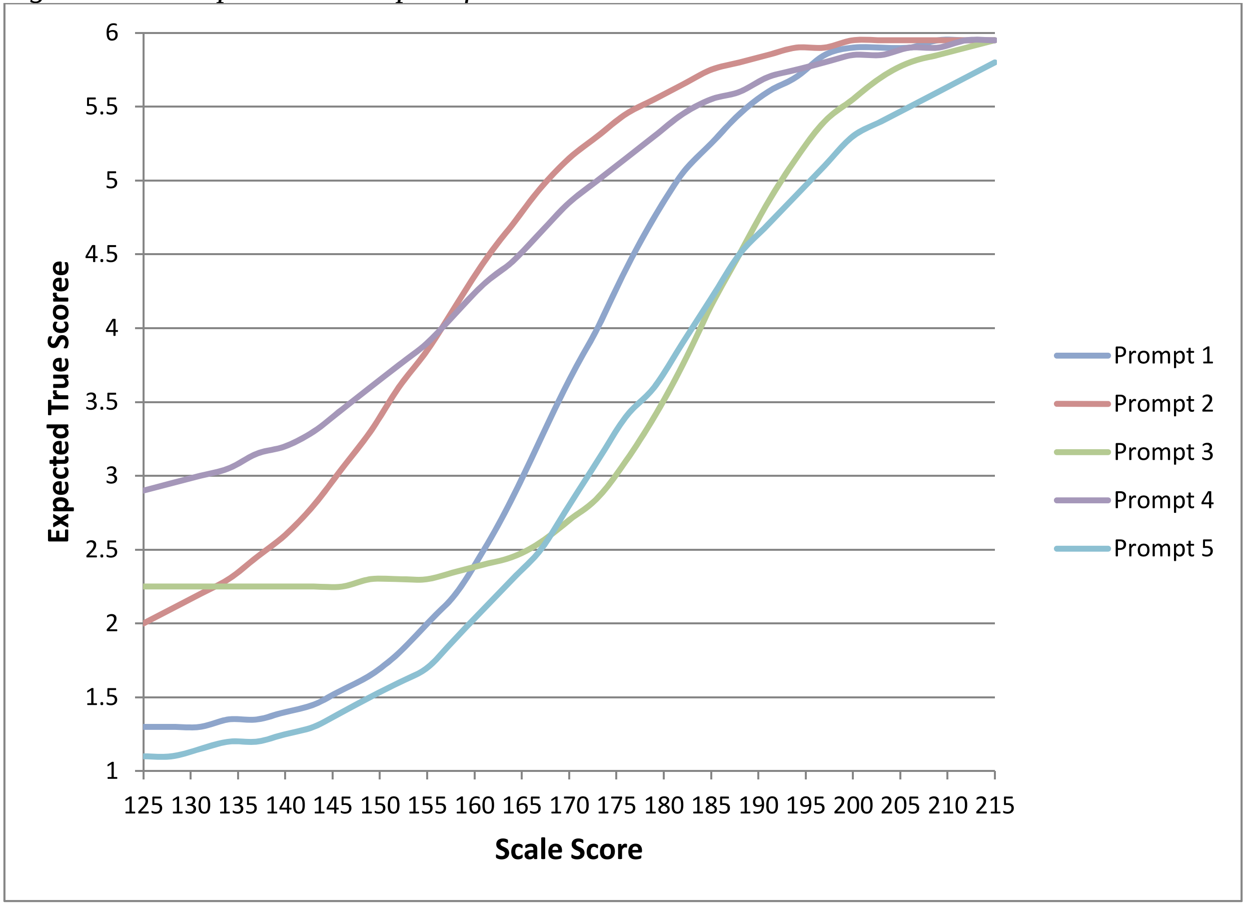
Important as a reality check, the *p*-value feedback provides information on the level of difficulty of each writing task. For NAEP writing ALS, the plan was to provide the cumulative percentage of students who received each score level. A copy of this feedback is included in Appendix D of the Technical Report. Because a computer presentation would have been cumbersome and required too much scrolling, this feedback was presented on paper. Logistics aside, it was found that such information was too confusing for panelists to comprehend.

### Reckase Charts

Per the Design Document (Measured Progress, 2011), a graphical version of the Reckase chart (Cizek and Bunch, 2009), was to be provided to the panelists before the second round of classification. The Reckase chart is a feedback mechanism for showing how panelists' ratings relate to the performance of students on the writing prompts. The Reckase chart shows the relative difficulty of the writing tasks and the rate at which the performance on the writing tasks changes as the performance of examinees increases on the pseudo-NAEP scale. Figure 24 shows an abbreviated graphical example of a Reckase chart with five prompts. The horizontal axis is a pseudo-NAEP scale. The vertical axis is the expected score of examinees who have a particular scale score. For example, for the examinees estimated to have a scaled score of 152, the expected average score would be 3.75 on Prompt 4. Prompt 5 is substantially harder for these students because their expected score on this prompt is 1.6. The Reckase chart provides the conditional difficulty of each writing task.

The purpose was to provide panelists with information on how tasks vary by difficulty at different parts of the scale. However, because of the similarity of the task characteristic curves, it was decided prior to the field trial based on consultation with the COR and a TACSS member not to provide such feedback. The actual Reckase charts prepared for the field trial are included in Appendix D of the Technical Report.

*Figure 24: Graphical Example of a Reckase Chart*



### Scoring Guides

The scoring guides were presented to the panelists in the exact format that was used by operational scorers. The scoring guides were not task specific. Instead there was only one scoring guide for each purpose for writing. The scoring guides are provided in Appendix G. Note that the scoring guide for "to convey experience" was the same for grades 8 and 12.

### 3.2.5 Presentation of Student Work Samples

The plan was to implement a BoW standard setting that included both rangefinding and pinpointing stages. For the rangefinding stage, a set of pre-identified work samples was selected to represent the full range of achievement and a balance of passage type across forms. For the pinpointing stage, a second set of work samples was selected to target a finer range of the achievement continuum, based on the resulting cut scores from the rangefinding stage, while still maintaining passage type balance across forms. The selection procedures and criteria are detailed in the Technical Report. BoWTIE aided in the student work sample presentation for the following aspects of the ALS procedures:

- storing and presenting the rangefinding work samples
- ordering the work samples by their *expected a posteriori* (EAP) score estimate, from lowest to highest
- selecting the second set of work samples targeted to the cut scores identified during rangefinding

Panelists were asked to evaluate the cognitive load and logistical ease involved in reviewing both rangefinding and pinpointing bodies of work (BoWs) using BoWTIE, as it is possible that the classification task would become more difficult when panelists were asked to make distinctions between work samples that represent similar achievement, especially in the pinpointing stage. Additionally, particular attention was given to panelists' evaluation of the number of BoWs assigned for review. Because of the expectations for longer responses written by 12th graders, the single-panel field trial was implemented for grade 12 NAEP writing. During the debriefing, panelists were very vocal about whether the BoWs should have been presented in score order.

Panelists expressed concern that their classifications might have been biased by the ordering of the BoWs.

### 3.2.6 Field Trial Methodology

A single-panel standard-setting process was conducted using data from the operational administration of the 2011 assessment. The procedures employed were an abbreviated version of those proposed for the pilot study and operational ALS meeting. The agenda for the field trial is provided in Appendix A, along with all the ALS meeting agendas.

Excluding time for setup, the implementation of the field trial was scheduled for two full days. In comparison, the operational ALS meeting would span approximately three and a half days. Because the field trial was meant to test the logistics and computer-based components of the meeting and only grade 12 standard setting was conducted, some agenda items were abbreviated: introductions, framework and ALD reviews, training, and classification rounds. Thus, panelists completed one round of rangefinding during the first day followed by an evaluation (Evaluation 1) to assess both the ease of transitioning between computers and the logistics of accessing BoWs and entering classifications into BoWTIE. The second round of rangefinding was not conducted. Instead, all cut score feedback was presented on the second day followed by an evaluation (Evaluation 2) to assess which modes of presentation (on-screen, projected, displayed on posters, or printed) were most helpful. Panelists continued by providing their classifications for the pinpointing stage (Round 2). This stage was followed by the final evaluation (Evaluation 3), which was intended to assess the cognitive load and ease of the classification task during the pinpointing stage. Finally, panelists were dismissed after a short debriefing of the two-day field trial.

For the field trial, cut scores were computed from the rangefinding round for the purpose of providing overall cut scores, cut score location feedback, and consequences data feedback. The overall cut scores were also used to select pinpointing BoWs based on the procedures described in the Technical Report. More importantly, the rangefinding classification data was very useful as additional verification that BoWTIE was computing the logistic regression cut scores. After the field trial, the data were sent to the data analysis group at Measured Progress to parallel process the cut scores.

No cut scores were computed from the pinpointing round. As planned, panelists were not provided feedback from the pinpointing round. A debriefing session began very shortly after the last panelist finished with pinpointing classification.

## 3.3 Field Trial Findings

True to its purpose, the field trial yielded important findings. These findings were presented to the TACSS and became the bases for some recommendations regarding details of ALS implementation. These logistical findings are organized around four logistical aspects: (a) CBA and ALS computers, (b) ALDs, (c) BoW selection and classification, and (d) between round feedback.

Panelist appraisals of the activities during the field trial were collected through process evaluation questionnaires and a debriefing session. The summary of responses to evaluation questions are in Appendix I and the notes from the debriefing session are in Appendix J. Specifically, panelists were asked about the amount of time and work space, ease of computer operation and manipulation, the room setup, the elements of the meeting that worked best, and the things they would change.

### *3.3.1 Laptop Computers*

Even though panelists were not vexed by having to use two laptop computers, it was deemed unnecessary for the panelists to have both computers for the duration of the ALS process. Panelists needed access to the CBA computers for taking a form of the writing NAEP and reviewing the rest of the writing tasks that were in their pool. They also needed access to the writing tasks during the rangefinding and pinpointing classification rounds. Because panelists needed only reminders of the tasks for these parts of the process, screenshots of the tasks were provided through BoWTIE. Additionally, panelists were provided printed screenshots of the tasks that they could write notes on if they chose to. They had access to the printed screenshots for the duration of the process.

In the ALS process, taking a form of the NAEP and reviewing the writing tasks occurred before the ALS laptops were needed, which began with the first evaluation questionnaire. For the first part of the process, panelists used the CBA laptops, which were then replaced with the ALS laptops when appropriate. This arrangement reduced the clutter in panelists' work spaces, but panelists still needed to be able to view the writing tasks the same way that students viewed them. This was accomplished by providing a CBA laptop for each group on each side of the room after the individual CBA laptops were switched to the ALS laptops.

During the task review, panelists found it cumbersome to go back to a previously viewed task. The only way to go back to a writing task was to exit the application, restart the application, listen to the directions, and go through the tasks one by one until the desired task was reached. This issue was resolved with a modification to the application that added a "Back" button to complement the "Next" button. Working with the NCES contractor, a plan was developed whereby the modification was made for use in the ALS process.

The field trial revealed difficulty for panelists to log on to the CBA laptop. The multiple logons (to the Windows operating system, then to the CBA application, then to the actual assessment form) were very confusing for the panelists. Although these logon protocols were the same ones used in the student administration of the assessment, it was determined that additional time and better logon directions such as a written copy of the instructions should be provided for panelists.

The logon procedures designed for the BoWTIE did not present any problems for panelists.  Logon issues that arose were due to inaccurate entry or inaccurate logon information. However, during the field trial, software issues were discovered in the BoWTIE application that affected both the selection of pinpointing BoWs and the presentation of the BoWs to panelists for classification. These issues were fixed prior to the pilot study.

### *3.3.2 Achievement Levels Descriptions*

Because the ALDs are the basis for the BoW classification from which the cut scores are computed, panelists' access to the ALDs is always of paramount importance in any standard-setting effort. Given that the ALS process was computer-based, one consideration was whether to make the process paperless in every way. Panelists in all previous ALS processes used paper copies of the ALDs, which were useful for noting issues to take into consideration during the rating or classification task. For the field trial, in addition to the hard copy, the electronic copy of the ALDs was also provided through BoWTIE and on an overhead screen projected by the process facilitator. Additionally, as in a previous NAEP ALS meeting, the ALDs were printed on 24″x 36″

posters, with one level described on each poster. Two copies of each poster were strategically placed around the room.

Panelists were asked if they accessed the ALDs in each medium. The most frequently used medium was paper, with 14 panelists reporting that they accessed the ALDs through the printed copy, followed by 10 panelists who accessed the ALDs through BoWTIE, and nine who looked at the projected copy of the ALDs. Only six panelists looked at the ALDs printed on the posters. When asked about their preferred presentation of the ALDs, 10 of the 20 panelists preferred paper, and six preferred BoWTIE. Only two panelists preferred the projected version and one panelist preferred the poster.

It was concluded that the operational ALS meeting would not be totally paperless because the ALDs should be printed for panelists. The electronic ALDs were already available in BoWTIE, but the practice of providing paper copies was retained for the operational ALS meeting.

### *3.3.3 Selection and Classification of Bodies of Work*

Even though cut score computation was not the primary interest of the field trial, an important finding was related to not having Advanced cut scores for some panelists. This was an artifact of some panelists not having classified any rangefinding BoWs to the Advanced category. Note that for the field trial, BoWs with scores of 1s or 6s for both tasks were excluded from selection. The original reason was that inclusion of these BoWs would not add information to the process, given that two 1s will always be Below Basic and two 6s will be Advanced. This decision resulted in a set of BoWs in which the highest raw scores were a 5 and a 6. Because some panelists did not believe that the BoWs, even those with the highest scores included in rangefinding, exhibited the KSAs that matched the description for the Advanced level, they did not have any data to which to fit a logistic regression curve to compute Advanced level cut scores. This led to a proposal to include BoWs with raw scores of two 6s. This was supported and recommended by the TACSS.

Two other exclusions were discussed with the TACSS based on the field trial experience. First, one of the student responses included graphic descriptions that some panelists considered inappropriate and disturbing. Instead of applying a rule for exclusion, the TACSS recommended that concerns about student responses be handled on a case-by-case basis. Thus, the selected BoWs would be reviewed in preparation for the pilot study (and subsequently, the operational ALS study) and any concerns about student responses would be submitted to the TACSS and the COR. The second exclusion issue discussed with the TACSS was that of BoW raw scores that were more than one point apart. The TACSS recommended that these BoWs not be excluded from selection.

One primary logistical issue for any BoW implementation is the number of BoWs to be classified by panelists—the number has to be large enough for computing stable cut scores but not so large that panelists will get too fatigued to provide reliable classifications. Between 30 and 50 is the number that Measured Progress has used over the years for state large-scale and alternate assessment programs. With 50 NAEP writing BoWs, it took panelists up to three hours to complete the rangefinding classifications. This was the case even though most panelists had a slow start because the directions provided were focused on the mechanics of making classifications using BoWTIE.

A lesson learned from the field trial is that the directions should emphasize the purpose of tasks and target panelists' conceptual understanding of the activities, as well as indicate the considerations that panelists should take into account in making their judgments. Based on the field trial experience, another improvement to the directions was to emphasize to panelists that the ratings should be based on the BoW consisting of both responses to two tasks. Thus, a step-by-step direction was developed to ensure that panelists would read and make notes on each task before making a decision on how to classify each BoW. Figure 25 is an excerpt from the facilitator handbook, included in Appendix D.

*Figure 25: Directions to the Panelists for BoW Classification*

Classify the BoW to an achievement level independently by going through the following steps:

1. Read the student response to the first task and note KSAs demonstrated by the examinee in the Comments box in BoWTIE. Leave the My Level Choice drop down menu at 'Select a Level' until you complete commenting on the second task response.
2. Read the student response to the second task and note KSAs demonstrated by the examinee in the Comments box. There is only one comment box, which will contain your notes on both tasks.
3. Use your comments on the student work to classify the BoW into an achievement level. Use the My Level Choice drop down menu to select the achievement level that you feel best represents the performance demonstrated by the student on the combination of the two tasks. Your classification should be based on how the KSAs seen in both tasks correspond to your understanding of the ALDs. Your choice of an achievement level will be visible in the drop down menu for both tasks and on the BoW list on your Home page.

Repeat [the above steps] for each of the rest of the BoWs.

During the debriefing session, some panelists were very vocal about their dissatisfaction regarding the fact that the BoWs were presented in score order for the rangefinding. They were concerned that the ordering affected their classifications. That is, seeing low-level performance first affected their expectations, which in turn affected their classifications. It was called to their attention that they did not have to classify the BoWs in the order they were presented. The ordering was for expedience—it was expected that it would have taken panelists a significantly longer period of time to classify 50 BoWs into achievement levels if the BoWs had not been ordered by score in some fashion. Some panelists shared the different ordering strategies they used to classify the 50 BoWs. For example, one general public panelist shared that the first two BoWs he read were at the two ends of the distribution, and the third was in the middle of the distribution. This gave him a realistic expectation of student performance of the assessment. Panelists were very engaged in the discussion about ordering. At the end of the debriefing session, they generally agreed that they would recommend that rangefinding BoWs should be presented in score order. A copy of the debriefing notes is included in Appendix J.

### 3.3.4 Between-Round Feedback

In addition to the overall cut scores, information provided to the panelists after the rangefinding round of the field trial included (a) *p*-value feedback, (b) cut score location charts, and (c) consequences data.

Given panelists' difficulty in understanding the *p*-value chart provided, it was recommended that it not be used for the operational implementation. The TACSS also recommended that in addition to the cut score location chart, another chart showing cut score distribution relative to the whole scale be shown to the panelists. Although this information is available to the panelists through the cut score location chart, the TACSS deemed it important for panelists to see the whole distribution on one screen without having to scroll.

For the field trial, the consequences data feedback was provided after the first round. This was by virtue of an abbreviated process. The consequences data feedback was planned to be provided after the second and third of classifications for the ALS process, and this remained the plan.

Between Rounds 1 and 2, the classification tally from Round 1 was provided to the panelists. From this tally, the BoWs with most disagreement between panelist classifications were discussed for the purpose of enhancing panelists' common understanding of the ALDs. Field trial experience and discussion with the TACSS led a definition of "most disagreement" as "entropy." Further, it was recommended that no more than 10 BoWs should be discussed in the operational ALS meeting, and that the discussion would be led by the content facilitator.

# Chapter 4—Pilot Study

A pilot study using the exact procedures designed to set the achievement levels for the writing assessments was implemented in November 2011. The intent was for procedural results from this study to provide information regarding operational aspects of the procedure, feedback presentation, and the amount of time and understanding necessary for a smooth implementation of the operational achievement levels–setting (ALS) methodology. The 20 panelists per grade were selected using the same nomination and recruitment scheme used for selecting ALS panelists. The goal was to ensure that every detail of the pilot study was as similar as possible to the planned procedures for the operational ALS meeting.

For the pilot study, the version of the achievement levels descriptions (ALDs) used by panelists was the one that was provisionally approved by the Governing Board at the August 2011 quarterly meeting. Results from the pilot study led to a revision of the ALDs, as discussed in Section 2.9.1. Thus, the ALDs used during the pilot study were different from the final ALDs. The provisional version of the ALDs is in Appendix K.

Shortly after the conclusion of the pilot study, the special study, described in Section 2.10, was implemented with the same set of panelists. The process implemented in the special study emulated an ALS process, with a few exceptions: (a) the orientation was concentrated on the purpose of the study, (b) the training provided was geared to assessment elements related to the 2007 NAEP writing, and (c) the process concluded after one round of classification and a process evaluation. No feedback was provided to the panelists. The same Body of Work Technological Integration and Enhancements (BoWTIE) application was used, and the only difference was that the bodies of work (BoWs) classified by the panelists were responses to writing prompts from the 2007 NAEP writing administration.

## 4.1 Pilot Study Panelists

The recruitment efforts for the pilot study resulted in the confirmation of two grade-level panels of 18 members each, including Teacher of the Year award recipients and nominees, adjunct university instructors, a teacher consultant for the National Writing Project, a Fulbright Scholar, a Scripps Howard Foundation National Journalism Award winner, and others with notable qualifications and recognitions. Clearly well-qualified, the panels were fairly broadly representative, the most notable exceptions being general public representation and representation from the South in grade 8, and representation from the West in grade 12. The male-to-female ratio for both panels was characteristically weighted in favor of female representation.

Whereas the Special Study was presented during recruitment as one of the responsibilities of the Pilot Study panelists, the same grade-level panels participated in the Special Study. Panel composition for the Pilot Study and Special Study may be seen below in Table 19.

*Table 19: Pilot Study/Special Study Panel Composition*

| Demographic Variable | Attribute | Grade 8 | | Grade 12 | | All | | Goal |
|---|---|---|---|---|---|---|---|---|
| | | n | % | n | % | n | % | % |
| Panelist Type | Teachers | 11 | 61 | 10 | 56 | 21 | 58 | 55 |
| | Nonteacher Educators | 4 | 22 | 3 | 17 | 7 | 19 | 15 |
| | General Public | 3 | 17 | 5 | 28 | 8 | 22 | 30 |
| Gender | Female | 15 | 83 | 11 | 61 | 26 | 72 | 50 |
| | Male | 3 | 17 | 7 | 39 | 10 | 28 | 50 |
| Race/Ethnicity* | Caucasian | 15 | 83 | 12 | 71 | 27 | 77 | 80 |
| | Non-Caucasian | 3 | 17 | 5 | 29 | 8 | 23 | 20 |
| NAEP Region | Midwest | 6 | 33 | 7 | 39 | 13 | 36 | 35 |
| | Northeast | 4 | 22 | 5 | 28 | 9 | 25 | 20 |
| | South | 2 | 11 | 5 | 28 | 7 | 19 | 25 |
| | West | 6 | 33 | 1 | 6 | 7 | 19 | 20 |

*One grade 12 panelist elected not to identify ethnicity.

## 4.2 Pilot Study Process

The pilot study was held November 15—18, 2011, at the St. Louis Ritz-Carlton in Clayton, Missouri. For the 2011 NAEP ALS meeting, this hotel was selected to be the site for three reasons directly related to process implementation. It has an amphitheater with tiered desk seating for 140 people, an elevated stage, and permanent audio-visual equipment that includes rear-screen projection complemented by plasma TV screens on either side. The setup lends itself to the general sessions planned for the ALS process. More importantly, the computer networking infrastructures of the hotel building allowed us to set up our own private secure network. Our secure network setup was such that our computer server was located in the server room of the hotel. Our computers in the two grade-level rooms, the office, and the amphitheater were connected to our server, on which BoWTIE operated. Lastly, the meeting rooms were large enough to accommodate the grade-level panel room setup for the operational ALS meeting.

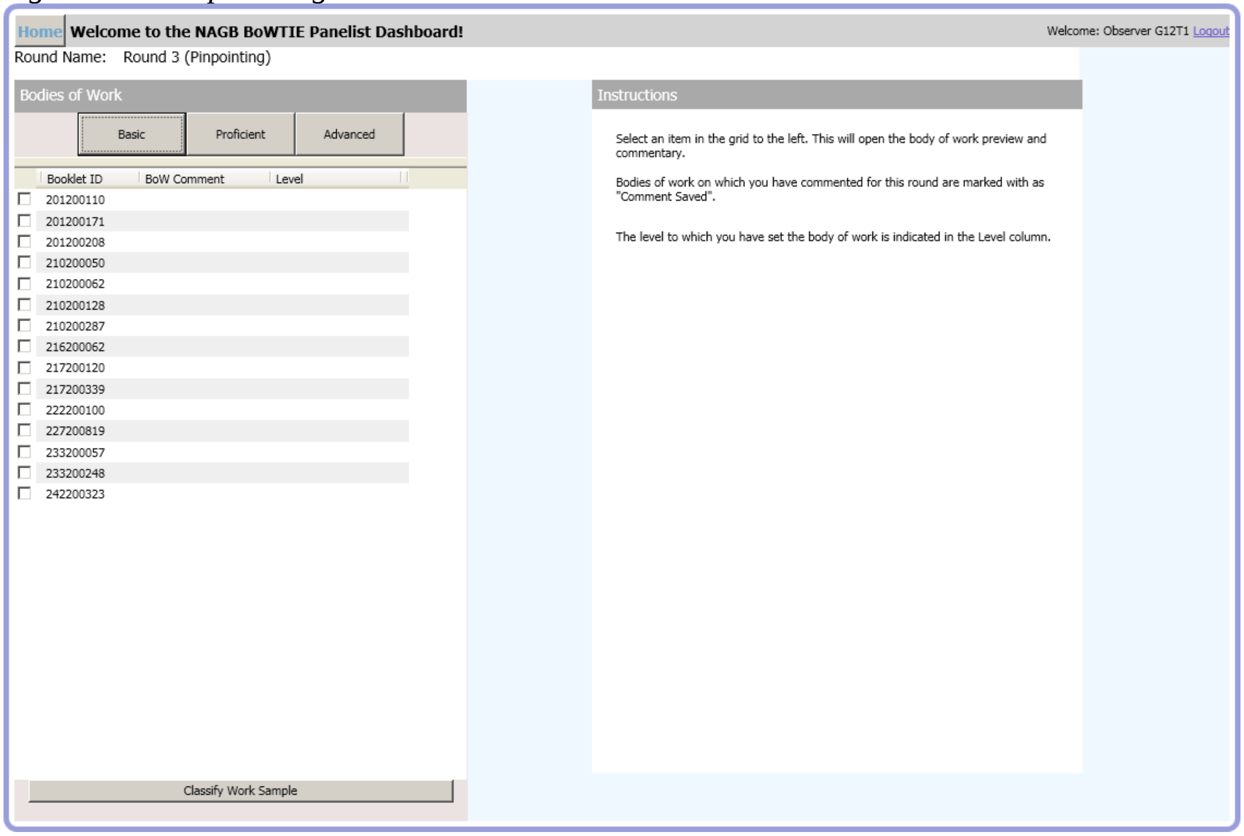
The process implemented during the pilot study was intended to be the exact process used for the ALS meeting, conditional on the results of the pilot study. Thus, the process implementation for the pilot study was as described in Chapter 2, Section 2.9, with some exceptions. The differences between the operational implementation and the pilot study are discussed in the following subsections.

### 4.2.1 Round 3 Classification: Pinpointing

Based on cut scores computed from the Round 2 rangefinding round, sample BoWs were selected with scores in the vicinity of the cut scores. These samples were the BoWs that panelists classified into achievement levels during the pinpointing round, which was the third and last classification round. The classification task was similar to the task for Rounds 1 and 2, except panelists had a new set of BoWs to classify into achievement levels. Technical details of the pinpointing BoW selection are in the Technical Report.

For each cut score, panelists were presented 15 BoWs with EAP scores around that cut score. The lowest score in the sample was lower than the lowest individual cut score set by a panelist in the grade group. Similarly, the highest score in the sample was higher than the highest individual cut score set by a panelist in the grade group. The scores of the 15 student work samples were uniformly distributed within the specified range. Unlike the rangefinding BoWs, pinpointing BoWs were not presented in rank order according to score. Within the pinpointing sample for a cut score, the BoWs were presented in order of their student “booklet” ID. Figure 26 shows a BoWTIE screenshot from the pinpointing round. Note that there are three tabs, one for each cut score. Given that the first three digits of the booklet ID compose the form number, the BoWs were presented to the panelists by form. Relative to the EAP scores, the BoWs were presented in no particular order. The booklet ID order might have helped the panelists to be more efficient in their classifications because they performed their classification task by form.

*Figure 26: Pinpointing*



Welcome to the NAGB BoWTIE Panelist Dashboard!

Welcome: Observer G12T1 Logout

Round Name: Round 3 (Pinpointing)

| | Booklet ID | BoW Comment | Level |
|---|---|---|---|
| □ | 201200110 | | |
| □ | 201200171 | | |
| □ | 201200208 | | |
| □ | 210200050 | | |
| □ | 210200062 | | |
| □ | 210200128 | | |
| □ | 210200287 | | |
| □ | 216200062 | | |
| □ | 217200120 | | |
| □ | 217200339 | | |
| □ | 222200100 | | |
| □ | 227200819 | | |
| □ | 233200057 | | |
| □ | 233200248 | | |
| □ | 242200323 | | |

Select an item in the grid to the left. This will open the body of work preview and commentary.

Bodies of work on which you have commented for this round are marked with as "Comment Saved".

The level to which you have set the body of work is indicated in the Level column.

For pinpointing, the panelists’ task was to classify each BoW below or above the cut score based on their understanding of the ALDs and the level of performance demonstrated in the responses to the two tasks. The classification task was performed separately for each cut score. The classification and annotation interface for pinpointing was different from that of rangefinding on only one count—there were only two choices when setting each level:

- Below Basic vs. Basic or Above
- Below Proficient vs. Proficient or Above
- Below Advanced vs. Advanced

A subsection included in the results section discusses issues regarding computation of cut scores from the pinpointing round. Upon presentation of information to the Technical Advisory Committee on Standard Setting (TACSS), a recommendation was made not to employ a pinpointing round. The Round 3 classification task was changed to another rangefinding round with a second set of BoWs.

### *4.2.2 Presentation Order of the Rangefinding BoWs*

For the pilot study, the rangefinding BoWs were presented to the panelists in ascending order of EAP scores. This presentation order was consistent with the more recent BoW method implementations by Measured Progress as well as the booklet classification studies implemented by ACT as part of investigating different methods for setting achievement levels for the 1998 NAEP writing (Hanson, Bay, & Loomis, 1998; Hanson & Bay, 1999; Chen, Loomis, & Fisher, 2000) and the 1994 NAEP geography and US history (Bay & Loomis, 1995; Kane & Bay, 1996).

After the pilot study, there was a sense that panelists were, for the most part, classifying rangefinding BoWs as they were presented—from low to high EAP scores. If the order of presentation was indeed biasing the panelists' classification, presenting the BoWs in highest to lowest for half of the panel was considered. The logistical challenges associated with the different ordering was a consideration for the TACSS's recommendation to present the rangefinding BoWs from highest to lowest for all panelists. Interestingly, personal communication with one of the original authors of the method informed the Chief of Standard Setting (CoSS) that highest to lowest was the BoW ordering used in the original implementation of the method.

### 4.2.3 Training to Enhance Understanding of the Achievement Levels Descriptions

For the pilot study, training on the ALDs included the following:

- presentation of the 2011 NAEP Writing Framework and ALDs (whole-group session)
- review and discussion of ALDs (grade-group session)
- review of responses to writing tasks for common understanding of the ALDs (grade-group session)

Based on the sessions, per se, the structure of the ALDs training for the pilot study was not different from that for the operational ALS meeting. After panelists listened to the presentation in the general session, they had an opportunity for more discussion in the following grade-group session. The opportunity for panelists to discuss the ALDs particular to the grade level for which they were setting achievement levels was very important. Such discussion cannot be implemented sufficiently in the whole group given some differences between the grade levels. The intent for the grade-group discussion was to go over the details of the different elements of the ALDs and what they meant to the panelists.

In the grade-group sessions, the discussion led by the respective content facilitators went quite differently. The discussion in grade 12 focused on the different dimension of writing performance that led to the matrix form of the ALDs. The content facilitators created the matrix version of the ALDs, which the panelists considered very helpful to use during their classification tasks, after the ALDs were modified following the pilot study. The matrix version of the ALDs was used by the grades 8 and 12 panelists during the operational ALS meeting, but it is not intended for public consumption.

The grade 8 discussion was based on student responses (materials intended for the next session), which the content facilitator read aloud to the panelists. After reading each response, the content facilitator asked the panelists what the classification for that response should be. Although this activity was similar to the exercise intended for the next session, reading the responses aloud to the panelists presented an element that was not intended the responses were edited as they were read aloud. In other words, the response that the panelists heard and classified was not necessarily the same response that they would have read themselves.

For the third part of the training on the ALDs, 15 student responses were selected for the panelists to discuss and, as a group, classify into achievement levels based on the knowledge, skills, and abilities (KSAs) demonstrated in the responses and panelist understanding of the ALDs. These were student responses to the tasks that were marked for release. The set of 15 responses included one sample response for each of the score levels 2 through 6. The sample responses were presented to the panelists in order from low to high raw score within each task. They were discussed in the grade groups in about the same order, with no specification as to which task was discussed first. Facilitators were told the following with regard to this session:

- The purpose of the activity is to discuss the ALDs in order to gain a common understanding among panelists.
- When the group discusses achievement level classifications, consensus is a goal but not a requirement.

- The amount of discussion for each response depends on how discrepant the panelists' classifications are. Discussing responses that all panelists classify the same way is not very helpful in gaining common understanding of the ALDs. Richer discussions are expected when panelists classify a response to different achievement level categories.

For the pilot study implementation of the ALS process, the TACSS did not think that panelists were provided enough opportunities to gain awareness and be cognizant of the raw scores assigned to the individual responses that became the basis for the scores used in the ordering of the BoWs. This part of the training on ALDs was then modified to be the Response Classification Exercise described in section 2.9.8.6. This exercise is not dissimilar to the Paper Selection Exercise that was a staple in NAEP ALS implementations done by ACT for the Governing Board in the 1990s. The Paper Selection Exercise was also considered to be the method for setting the 1998 NAEP writing achievement level cut scores (Loomis, & Hanick, 2000).

## 4.3 Pilot Study Results

Results from the pilot study were intended to inform the operational ALS meeting. The numerical results from the pilot study were not intended to be used for reporting results but may have been considered a replication had there not been changes to the operational process relative to the pilot study process. Given the changes to the implementation discussed in this report, numerical results from the pilot study were not suitable for comparisons to the results from the operational ALS meeting. The process evaluation results and notes from the debriefing, in conjunction with the numerical results, resulted in revised ALDs, which became another source of difference in numerical results from the pilot study and the operational ALS meeting.

Another significant difference in the implementation of the pilot study and the operational ALS meeting was the removal of the pinpointing round. Computation issues surrounding the cut scores from the pinpointing round led to the change. Results from the study on cut score computation from pinpointing are discussed in detail later in this section (see Section 4.3.4).

Results from the original implementation of the special study (implemented immediately following the pilot study) were rendered moot relative to the primary purpose of the study because of the subsequent modifications to the ALDs. The results are presented in this section to the extent that the numbers relate to the numerical result of the pilot study.

### 4.3.1 Numerical Results

The overall cut scores and related measures of cut score variability across rounds and across panelists resulting from the pilot study are presented in this section. All cut scores are reported on the pseudo-NAEP scale, as discussed in Section 2.9.6. The measures of variability reported here are those that have been used in previous ALS processes implemented for NAEP (e.g., ACT, 2007; Loomis & Hanick, 2000). Results from the pilot study were presented to the TACSS on December 13, 2011, during an online meeting.

Table 20 presents the cut scores across the three rounds of classifications. Also presented in Table 20 is the percentage of students at or above each cut score. It was noted during the TACSS meeting that both overall cut scores and percentages changed very little from round to round. Table 21 presents the number and percentage of panelists who changed their cut scores from round to round. The changes in overall cut scores or lack thereof are due to little movement in panelists' cut scores, not to large changes in panelists' cut scores that cancel the effect of each other.

*Table 20: Pilot Study Cut Scores and Percentages At or Above*

| Grade | Achievement Level | Round 1 | | Round 2 | | Round 3 | |
|---|---|---|---|---|---|---|---|
| | | Cut Score | % At or Above | Cut Score | % At or Above | Cut Score | % At or Above |
| 8 | Basic | 189 | 95 | 189 | 95 | 184 | 96 |
| | Proficient | 223 | 77 | 229 | 72 | 225 | 76 |
| | Advanced | 282 | 17 | 286 | 15 | 284 | 16 |
| 12 | Basic | 511 | 93 | 512 | 93 | 510 | 93 |
| | Proficient | 556 | 63 | 553 | 66 | 553 | 66 |
| | Advanced | 630 | 2 | 632 | 2 | 628 | 3 |

*Table 21: Pilot Study Changes in Panelists' Cut Scores*

| Grade | Achievement Level | Round | Increased | | No Change | | Decreased | |
|---|---|---|---|---|---|---|---|---|
| | | | n | % | N | % | N | % |
| 8 | Advanced | R1:R2 | 11 | 61.11 | 3 | 16.67 | 4 | 22.22 |
| | | R2:R3 | 5 | 27.78 | 1 | 5.56 | 12 | 66.67 |
| | Proficient | R1:R2 | 8 | 44.44 | 5 | 27.78 | 5 | 27.78 |
| | | R2:R3 | 3 | 16.67 | 1 | 5.56 | 14 | 77.78 |
| | Basic | R1:R2 | 5 | 27.78 | 6 | 33.33 | 7 | 38.89 |
| | | R2:R3 | 1 | 5.56 | 0 | 0.00 | 17 | 94.44 |
| 12 | Advanced | R1:R2 | 5 | 27.78 | 7 | 38.89 | 6 | 33.33 |
| | | R2:R3 | 5 | 27.78 | 0 | 0.00 | 13 | 72.22 |
| | Proficient | R1:R2 | 6 | 33.33 | 5 | 27.78 | 7 | 38.89 |
| | | R2:R3 | 8 | 44.44 | 0 | 0.00 | 10 | 55.56 |
| | Basic | R1:R2 | 4 | 22.22 | 10 | 55.56 | 4 | 22.22 |
| | | R2:R3 | 4 | 22.22 | 1 | 5.56 | 13 | 72.22 |

Given that the overall cut scores were medians of panelists' cut scores, the mean absolute deviation (MAD) is an appropriate measure of variability across panelists. MAD is the average of the absolute differences between the median cut score and the panelists' respective cut scores. Figures 27 and 28 present the MAD values for grades 8 and 12 for each round of classification.

The increase in the MAD from Round 1 to Round 2 for grade 8 was noted during the TACSS meeting. This was not expected, since variability of cut scores across panelists tends to decrease across rounds. However, when the MAD charts for grades 8 and 12 were compared, it was conjectured that maybe the increase was due to the unusually low values for Round 1 MADs for grade 8. Note that the feedback provided to panelists between Rounds 1 and 2 included the overall cut scores and cut score distribution information by way of cut score location charts and cut score distribution charts. The consequences data feedback was provided only after Round 2. All feedback presentation was updated for subsequent rounds after its initial presentation.

*Figure 27: Pilot Study MAD—Grade 8*

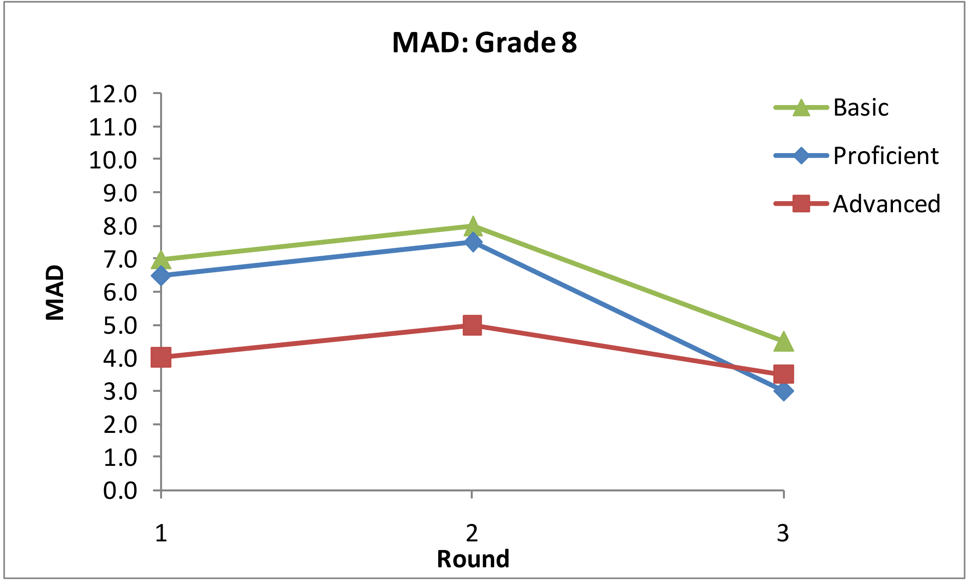
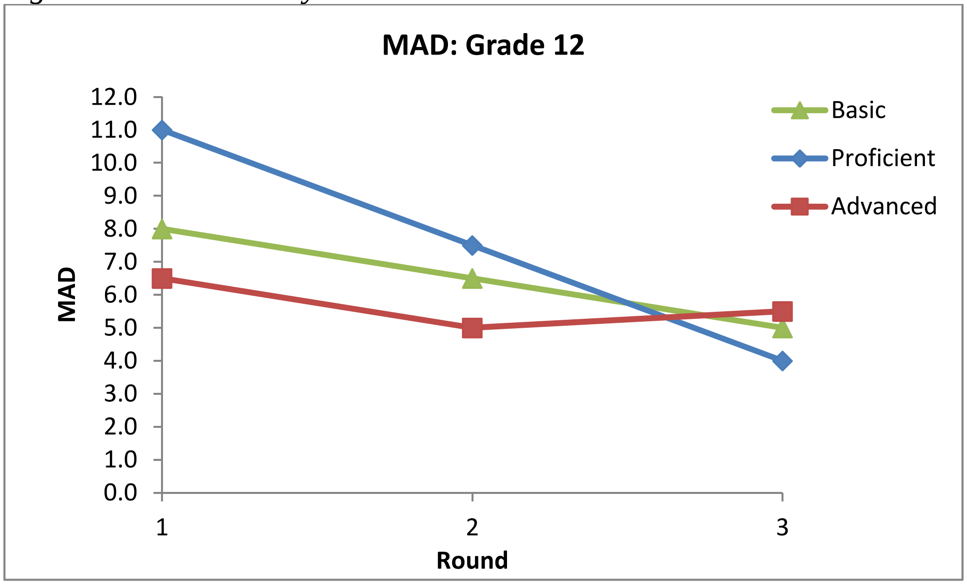


*Figure 28: Pilot Study MAD—Grade 12*



The standard error of the overall cut scores was also examined. Unlike the mean, the median does not have a standard error computation that is de rigueur in psychometrics. Two nonparametric methods of computing the standard errors used in previous ALS processes were used here—one is empirical based (Maritz & Jarrett, 1978) and the other is based on a bootstrapping technique (Efron & Gong, 1983). Details of the computation are presented in the Technical Report. Table 22 presents the standard errors of the cut scores for all three rounds. The standard errors of the cut scores by group and table group are also presented in the Technical Report.

*Table 22: Pilot Study Standard Error of the Cut Scores*

| Grade | Achievement Level | Round 1 | | Round 2 | | Round 3 | |
|---|---|---|---|---|---|---|---|
| | | EmpSE[A] | BootSE[B] | EmpSE | BootSE | EmpSE | BootSE |
| 8 | Basic | 2.81 | 2.76 | 3.64 | 3.51 | 1.68 | 1.68 |
| | Proficient | 3.86 | 3.77 | 2.33 | 2.35 | 1.38 | 1.39 |
| | Advanced | 1.40 | 1.41 | 1.88 | 2.04 | 1.57 | 1.79 |
| 12 | Basic | 3.99 | 4.09 | 4.25 | 4.44 | 3.03 | 3.22 |
| | Proficient | 6.19 | 6.37 | 3.52 | 3.51 | 1.56 | 1.59 |
| | Advanced | 2.10 | 2.01 | 2.50 | 2.41 | 2.22 | 2.20 |

[A] EmpSE is the empirical standard error
[B] BootSE is the bootstrap standard error

Another measure of variability of cut scores is the standard error based on two observations (Brennan, 2002). For ALS processes, the two observations are the median cut scores for the two equivalent groups, A and B. Note that each group classified BoWs from equivalent writing task pools. The standard error was computed as the difference between the group cut scores divided by two. Tables 23 and 24 present the standard error and the 95% confidence interval around the mean of the two median cut scores.

*Table 23: Pilot Study Standard Error and 95% Confidence Interval Around the Mean of the Group Cut Scores: Grade 8*

| Achievement Level | Round | Cut Score | | | Standard Error | 95% Confidence Level | |
|---|---|---|---|---|---|---|---|
| | | Panel A | Panel B | Mean | | Upper Limit | Lower Limit |
| Basic | 1 | 120 | 122 | 121.0 | 1.0 | 122.96 | 119.04 |
| | 2 | 121 | 120 | 120.5 | 0.5 | 121.48 | 119.52 |
| | 3 | 115 | 123 | 119.0 | 4.0 | 126.84 | 111.16 |
| Proficient | 1 | 172 | 171 | 171.5 | 0.5 | 172.48 | 170.52 |
| | 2 | 171 | 176 | 173.5 | 2.5 | 178.40 | 168.60 |
| | 3 | 173 | 172 | 172.5 | 0.5 | 173.48 | 171.52 |
| Advanced | 1 | 221 | 216 | 218.5 | 2.5 | 223.40 | 213.60 |
| | 2 | 221 | 217 | 219.0 | 2.0 | 222.92 | 215.08 |
| | 3 | 211 | 212 | 211.5 | 0.5 | 212.48 | 210.52 |

*Table 24: Pilot Study Standard Error and 95% Confidence Interval Around the Mean of the Group Cut Scores: Grade 12*

| Achievement Level | Round | Cut Score | | | Standard Error | 95% Confidence Level | |
|---|---|---|---|---|---|---|---|
| | | Panel A | Panel B | Mean | | Upper Limit | Lower Limit |
| Basic | 1 | 119 | 121 | 120.0 | 1.0 | 121.96 | 118.04 |
| | 2 | 122 | 122 | 122.0 | 0.0 | 122.00 | 122.00 |
| | 3 | 122 | 120 | 121.0 | 1.0 | 122.96 | 119.04 |
| Proficient | 1 | 176 | 163 | 169.5 | 6.5 | 182.24 | 156.76 |
| | 2 | 174 | 167 | 170.5 | 3.5 | 177.36 | 163.64 |
| | 3 | 175 | 170 | 172.5 | 2.5 | 177.40 | 167.60 |
| Advanced | 1 | 220 | 212 | 216.0 | 4.0 | 223.84 | 208.16 |
| | 2 | 218 | 209 | 213.5 | 4.5 | 222.32 | 204.68 |
| | 3 | 209 | 210 | 209.5 | 0.5 | 210.48 | 208.52 |

This manner of computing standard error of the cut scores was first used in 1992 for setting achievement levels for NAEP reading and mathematics. A criticism of this standard error computation is that the assumption of independence between the two observations was not met. The two groups were not independent because they were provided the same feedback between rounds. The equivalent task pools were not independent because of the common tasks included in their pool.

After feedback from the third round of classifications was presented, panelists were asked to respond to the Consequences Data Questionnaire (CDQ). There were three primary questions on the CDQ:

> Given your understanding of student performance at the [Basic/Proficient/Advanced] achievement level, does this percentage reflect your expectation about the proportion of students whose NAEP score would be at or above the Basic cut score?
>
> Having seen the data on the percentages of students whose score on the NAEP was at or above the cut score your panel set for each achievement level, would you change one or more of the achievement levels you have set if you could?
>
> What is your final [Basic/Proficient/Advanced] cut score recommendation to the Governing Board? Please enter a scale value keeping in mind that raising the cut score would lead to a smaller percentage of students scoring at or above the [Basic/Proficient/Advanced] level and lowering the cut score would lead to a larger percentage of students scoring at or above the [Basic/Proficient/Advanced] level.

Between 72% and 89% of panelists indicated that the percentages at or above each achievement level reflected their expectations. Less than a third indicated that they would change one or more cut scores if they could. Responses to the third question were deemed unreliable because more panelists responded to this question than there were panelists who indicated that they would change one or more cut scores if they could. Panelists had been instructed to skip the third question if they responded negatively to the second question. Details of the CDQ response summary are included in Appendix I.

Based on cut scores from the third round, panelists were presented potential exemplar BoWs for their recommendation. These BoWs were from the form with two tasks marked for release. Recall that there were eight BoWs from this form that were selected for rangefinding. These eight BoWs were evenly distributed across the score range. The number of BoWs presented as potential exemplars for the two grades is shown in Table 25. Note that for grade 8, the cut score was so low that even the lowest-scored rangefinding BoW from the form with two tasks marked for release fell in the Basic category.

*Table 25: Number of Potential Exemplar BoWs Presented to Pilot Study Panelists*

| Grade | Achievement Level | Number of BoWs |
|---|---|---|
| 8 | Basic | 3 |
| | Proficient | 2 |
| | Advanced | 3 |
| 12 | Basic | 2 |
| | Proficient | 4 |
| | Advanced | 1 |

Because this was the pilot study, no further action was taken based on panelists' ratings of the possible exemplar BoWs. Nevertheless, a summary of panelists' ratings as well as their comments were presented to the TACSS. It was noted during the TACSS meeting that comments made by panelists indicated misunderstanding of the ALDs.

### 4.3.2 Process Evaluation

Process evaluation was the primary source of procedural validity evidence of the ALS results. Five process evaluation questionnaires, administered at strategic stages of the process, were filled out by panelists. The evaluations were designed to assess their understanding of instructions, tasks, and materials. Most responses were collected on four different five-level Likert-type scales for agreement, adequacy, length, and clarity. On each of the four scales, a level 5 is associated with a high level of magnitude on the scale, whereas a level 1 is associated with a low magnitude. The 20 actual labels for the response levels are appropriately worded for each scale (see Table 26). Several items included the opportunity to provide narratives that addressed specific aspects of the process. These evaluations were reviewed at the end of each day, and any sources of confusion or misunderstanding were identified for clarification with individual panelists or the group as a whole. Any Likert question that received an average response lower than 3.5 was flagged and examined to determine whether there was a related issue that needed to be addressed. An exception was questions regarding the length of time allocated to sessions, in which case an average response that could not be rounded to 3.0 was considered a concern. The summaries for all the process evaluation questionnaires are included in Appendix I. Table 26 presents summaries for selected questions. Overall, panelists indicated that they understood their tasks and the materials, felt comfortable and confident in making their decisions, felt free to make independent judgments, and found the computerization of the process helpful. Specific feedback includes the following:

- Almost all panelists indicated that BoWTIE was helpful.
- All panelists indicated that the instructions were clear.
- All panelists indicated an understanding of what they were supposed to do in each round.
- All panelists indicated willingness to sign a statement recommending the use of the cut scores.

Further, when the same questions were asked across multiple evaluations, the percentage of positive responses tended to increase, as expected, as the workshops progressed.

*Table 26: Summary of Selected Questions from the Pilot Study Process Evaluations*

| Question | n | Grade | Round | Totally Agree (5) | Agree (4) | Somewhat Agree (3) | Disagree (2) | Totally Disagree (1) | Mean | SD |
|---|---|---|---|---|---|---|---|---|---|---|
| When we discussed BoWs that were selected because panelists tended to disagree on how to classify them, I felt that my comments were valued. | 28 | 8 |  | 2 (7%) | 12 (43%) | 14 (50%) | 0 (0%) | 0 (0%) | 4.5 | 0.5 |
|  | 28 | 12 |  | 0 (0%) | 13 (46%) | 12 (43%) | 3 (11%) | 0 (0%) | 4.4 | 0.7 |
| I understand what students in each of the achievement levels can do based on the panel cut scores. | 28 | 8 | 1 | 3 (11%) | 4 (14%) | 15 (54%) | 5 (18%) | 1 (4%) | 3.9 | 0.7 |
|  | 28 | 12 | 1 | 1 (4%) | 8 (29%) | 18 (64%) | 1 (4%) | 0 (0%) | 4.3 | 0.5 |
|  | 28 | 8 | 2 | 1 (4%) | 5 (18%) | 18 (64%) | 3 (11%) | 1 (4%) | 4 | 0.7 |
|  | 28 | 12 | 2 | 2 (7%) | 11 (39%) | 14 (50%) | 1 (4%) | 0 (0%) | 4.4 | 0.6 |
|  | 27 | 8 | 3 | 1 (4%) | 13 (48%) | 12 (44%) | 1 (4%) | 0 (0%) | 4.5 | 0.6 |
|  | 28 | 12 | 3 | 2 (7%) | 12 (43%) | 13 (46%) | 1 (4%) | 0 (0%) | 4.4 | 0.6 |
| I believe my round X classification of bodies of work into achievement levels is consistent with the Achievement Level Descriptions. | 28 | 8 | 2 | 4 (14%) | 9 (32%) | 15 (54%) | 0 (0%) | 0 (0%) | 4.4 | 0.5 |
|  | 28 | 12 | 2 | 2 (7%) | 14 (50%) | 12 (43%) | 0 (0%) | 0 (0%) | 4.5 | 0.5 |
|  | 27 | 8 | 3 | 1 (4%) | 12 (44%) | 14 (52%) | 0 (0%) | 0 (0%) | 4.5 | 0.5 |
|  | 28 | 12 | 3 | 1 (4%) | 11 (39%) | 15 (54%) | 1 (4%) | 0 (0%) | 4.4 | 0.6 |
| I would be willing to sign a statement (after reading it of course) recommending the use of the cut scores resulting from this ALS process. | 27 | 8 |  | 3 (11%) | 11 (41%) | 11 (41%) | 1 (4%) | 1 (4%) | 4.3 | 0.8 |
|  | 28 | 12 |  | 1 (4%) | 19 (68%) | 8 (29%) | 0 (0%) | 0 (0%) | 4.7 | 0.5 |

continued

| Question | n | Grade | Round | Totally Adequate (5) | Adequate (4) | Somewhat Adequate (3) | Inadequate (2) | Totally Inadequate (1) | Mean | SD |
|---|---|---|---|---|---|---|---|---|---|---|
| My understanding of the tasks I was to accomplish during each round was: | 27 | 8 |  | 4 (15%) | 16 (59%) | 6 (22%) | 1 (4%) | 0 (0%) | 4.7 | 0.6 |
|  | 28 | 12 |  | 1 (4%) | 14 (50%) | 11 (39%) | 2 (7%) | 0 (0%) | 4.4 | 0.6 |
| **Question** | **n** | **Grade** | **Round** | **Very Helpful (5)** | **Helpful (4)** | **Somewhat Helpful (3)** | **Not Helpful (2)** | **Not at All Helpful (1)** | **Mean** | **SD** |
| In general, during the achievement levels-setting process, I found using BoWTIE to be: | 27 | 8 |  | 1 (4%) | 19 (70%) | 6 (22%) | 1 (4%) | 0 (0%) | 4.7 | 0.5 |
|  | 28 | 12 |  | 1 (4%) | 20 (71%) | 7 (25%) | 0 (0%) | 0 (0%) | 4.7 | 0.4 |
| **Question** | **n** | **Grade** | **Round** | **Very Confident (5)** | **Confident (4)** | **Somewhat Confident (3)** | **Not Confident (2)** | **Not at All Confident (1)** | **Mean** | **SD** |
| The most accurate description of my level of confidence in my body of work classifications is: | 28 | 8 | 1 | 4 (14%) | 4 (14%) | 15 (54%) | 5 (18%) | 0 (0%) | 4 | 0.6 |
|  | 27 | 12 | 1 | 0 (0%) | 1 (4%) | 20 (74%) | 6 (22%) | 0 (0%) | 3.8 | 0.5 |
|  | 28 | 8 | 2 | 4 (14%) | 7 (25%) | 14 (50%) | 1 (4%) | 2 (7%) | 4.1 | 0.8 |
|  | 28 | 12 | 2 | 3 (11%) | 11 (39%) | 14 (50%) | 0 (0%) | 0 (0%) | 4.4 | 0.5 |
|  | 27 | 8 | 3 | 0 (0%) | 6 (22%) | 18 (67%) | 3 (11%) | 0 (0%) | 4.1 | 0.6 |
|  | 28 | 12 | 3 | 1 (4%) | 8 (29%) | 18 (64%) | 1 (4%) | 0 (0%) | 4.3 | 0.5 |
| The most accurate description of my level of confidence in the cut score recommendations I provided was: | 27 | 8 |  | 1 (4%) | 9 (33%) | 15 (56%) | 2 (7%) | 0 (0%) | 4.3 | 0.6 |
|  | 28 | 12 |  | 1 (4%) | 16 (57%) | 10 (36%) | 1 (4%) | 0 (0%) | 4.6 | 0.6 |

### *4.3.3 Special Study*

The purpose of the special study was to gather information to indicate relationships between performance on the 2007 NAEP writing and the 2011 NAEP writing. In brief, the task of the panelists for the special study was to classify 2007 BoWs using 2011 ALDs. Details of the cross-tabulation analyses of this study are included in the Technical Report. Note that results from this study were rendered moot relative to its purpose, given the modifications to the ALDs after the pilot study. However, the results of the special study helped to determine that revisions to the ALDs were necessary. BoW classification data from this study were used to compute cut scores that were then applied to the 2007 score frequency distribution. The percentages at or above the resulting cut scores are presented in Figures 29 and 30, and they are compared to the percentages reported on the Nation's Report Card for 2007.

*Figure 29: Initial Special Study Results—Grade 8*

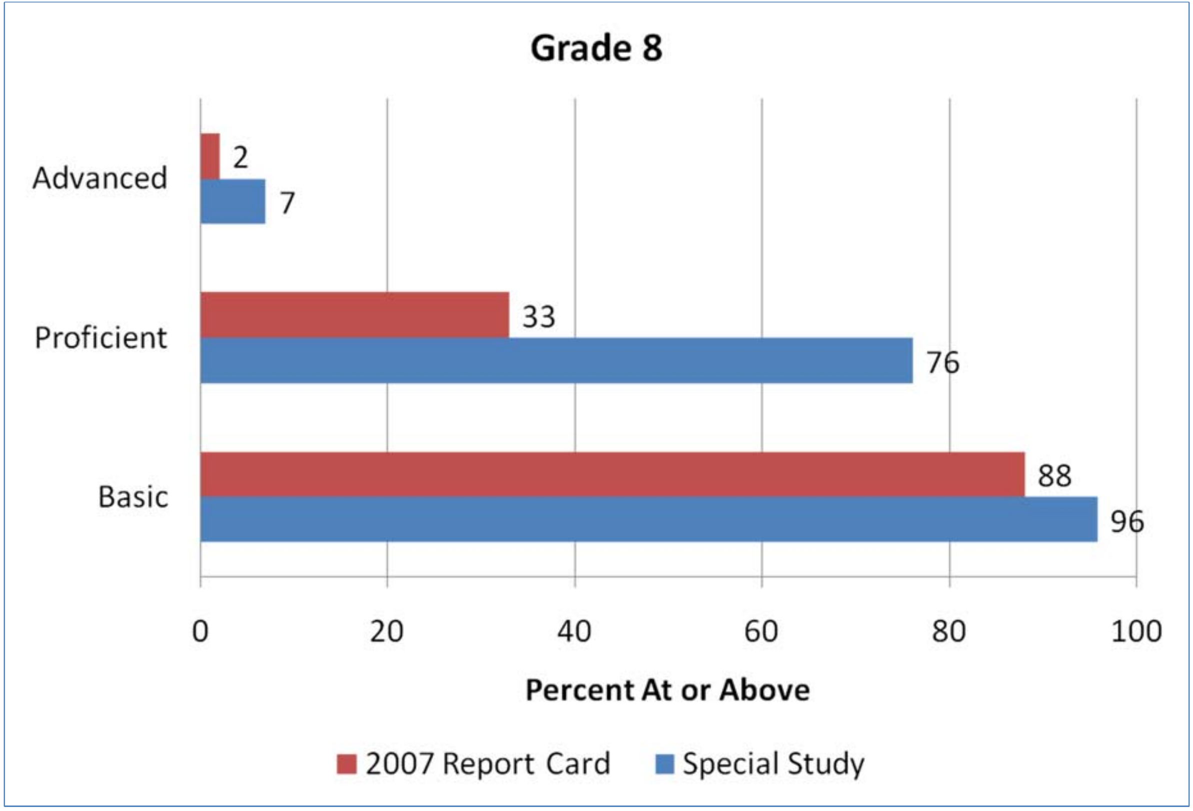
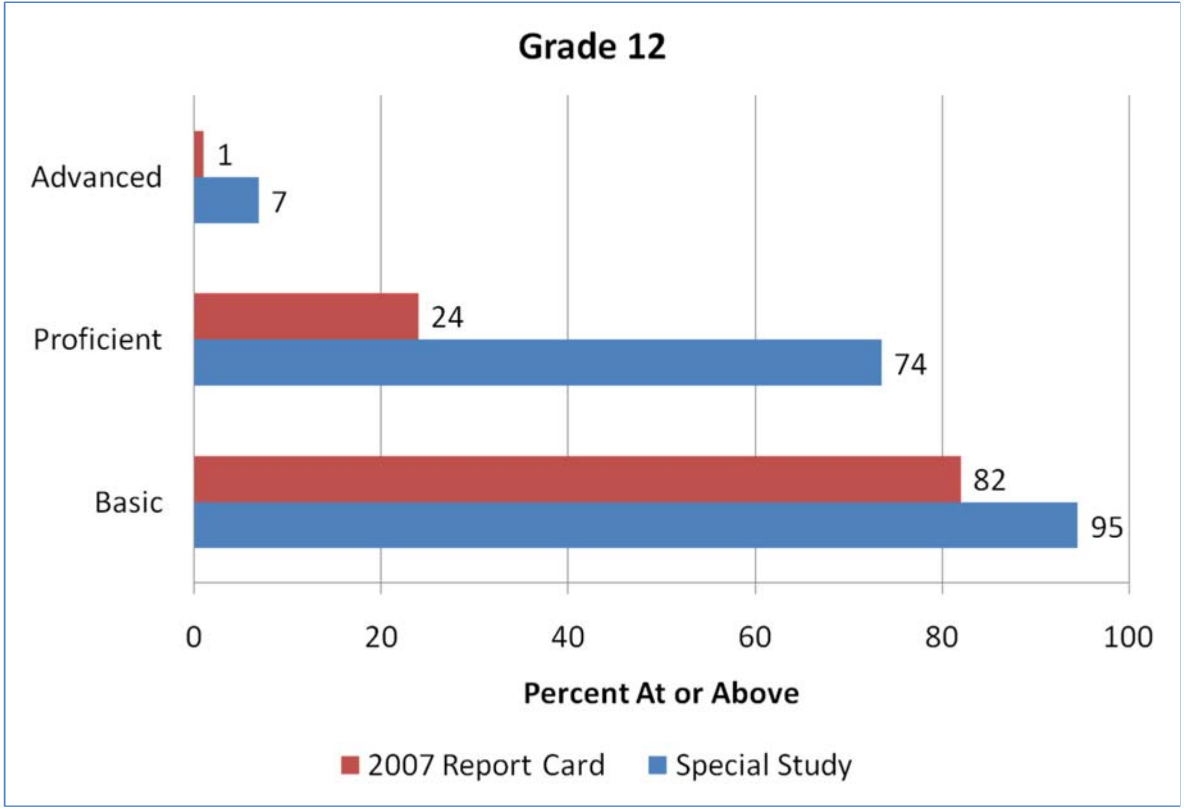


*Figure 30: Initial Special Study Results—Grade 12*



### 4.3.4 Pinpointing Cut Score Computation

During the TACSS meeting immediately prior to the pilot study, discussion ensued on the computation of cut scores from pinpointing. There were two concerns regarding the computation of cut scores using logistic regression. The first specific concern was with regard to the instability of a resulting cut score based on only 15 data points. Additionally, the limited range of scores only exacerbates the instability issue. The recommendation from the TACSS was to include data from Round 2 classifications in the computation. That is, to compute the Proficient cut score for a panelist, one would take the binary classifications of “Below Proficient” (0) or “Proficient or Above” (1) and combine them with the rangefinding classifications where Below Basic and Basic classifications were coded as 0 and Proficient and Advanced classifications were coded as 1. A logistic regression curve was then fitted to this data set with 65 observations (50 rangefinding and 15 pinpointing BoWs around the Proficient cut score).

Inadvertently, the Round 3 cut scores computed on-site were computed with the Round 2 rangefinding classifications plus all of the pinpointing classifications. Thus, each cut score was computed using 95 data points. This issue was found upon return to Measured Progress during a quality assurance check of all results. Note that all Round 3 feedback provided to panelists during the pilot study was based on this unintended way of computing the cut scores.

The trouble with this computation is that it makes an assumption regarding panelists' classifications that might not necessarily be consistent with their judgments. Continuing with the example of computing a panelist's Round 3 Proficient cut scores, all pinpointing BoWs around the Basic cut scores were coded as 0 and all pinpointing BoWs around the Advanced cut score were coded as 1. Although logical, such coding does not take into account the possibility that the panelist might have judged, say, a BoW around the Basic cut score as demonstrating performance at the Proficient level. The rating for that BoW should have then been 1 instead of 0. Table 27 presents the cut scores based on the different data inclusion. Pinpointing cut scores were also computed using only Round 3 classifications with all 45 data points used to compute each cut score, as well as only including the 15 classifications for BoWs selected around a specific cut score. Table 27 presents the corresponding percentages at or above the cut scores. Note that between "BoWTIE" and "Intended" computations, the cut scores do not differ by more than one scaled score, which may easily be attributed to rounding. Additionally, the corresponding percentages at or above do not differ at all.

*Table 27: Round 3 Cut Scores and Consequences Data From the Pilot Study*

| Grade | Achievement Level | BoWTIE Computation | | Intended Computation | |
|---|---|---|---|---|---|
| | | Cut Score | Percent At or Above | Cut Score | Percent At or Above |
| 8 | Basic | 184 | 96 | 184 | 96 |
| | Proficient | 225 | 76 | 225 | 76 |
| | Advanced | 284 | 16 | 285 | 16 |
| 12 | Basic | 510 | 93 | 511 | 93 |
| | Proficient | 553 | 66 | 553 | 66 |
| | Advanced | 628 | 3 | 628 | 3 |

Subsequent to the pilot study, Measured Progress performed additional investigation into the computation of pinpointing cut scores. A simulation study to evaluate the technical merits of computing cut scores using the four different data inclusion rules was performed. A report describing the study and its results is included in Technical Report as an appendix. Results of the simulation study favors computation with more data points.

### 4.3.5 Summary of Outcomes from the Pilot Study

Findings from the pilot study affected several aspects of the operational ALS process? Both numerical and procedural results discussed earlier led to some changes in the implementation of the operational ALS meeting. Some of these changes are discussed in the subsection of this report about the pilot study process. Further, some of the findings directly affected the achievement levels by virtue of a recommendation to review the ALDs, which resulted in modifications.

#### 4.3.5.1 Review and Modification of the Achievement Level Descriptions

Percentages of student performance at or above cut scores from the pilot study (see Table 27) were of magnitudes never seen before in over 20 years of NAEP achievement levels setting. Within the writing assessment, per se, the percentages of student performance at or above the achievement levels were vastly different. For example, student performance at or above Proficient was 33% for grade 8 in 2007, while the percentage based on the Round 3 cut scores from the pilot study was 76%; in grade 12 these numbers were 24% and 66%. Figures 31 and 32 provide the rest of the comparisons between percentages of student performance at or above the achievement levels for the 2007 and 2011 assessments based on pilot study cut scores.

A few conjectures can be made about the data shown in Figures 31 and 32. Note the similarity of the percentages resulting from the special study and the pilot study. Assuming that there was not a monumental change in student performance in writing between 2007 and 2011, the similarities might be attributed to the panelists being consistent in using the ALDs in their BoW classification task.

The magnitude of differences between the percentages reported in the 2007 Report Card and the special study might imply that the 2011 ALDs and the 1998 ALDs, used to set cut scores that were applied in 2007, were really describing different levels of performance. However, the two sets of ALDs were operational definitions of the same levels as defined by the Governing Board's policy definitions (see Chapter 1), which has not changed.

A scrutiny of the process did not indicate that the implementation affected the cut scores. Based on feedback from panelists combined with the observations regarding numerical results from the pilot study and the special study, the TACSS recommended a review of the ALDs.

A review of the ALDs, which was expected to lead to modifications to the ALDs, was proposed to the Committee on Standards, Design and Methodology (COSDAM) during the December 2011 meeting of the Governing Board. The intent of the modifications was not to change the levels, per se, but to better calibrate them to the policy definitions and improve the language to reduce ambiguity and improve parallelism across grades within achievement levels and within each grade across achievement levels. COSDAM, in turn, strongly recommended that if the ALDs were revised, they should be tested with panelists in a small-scale study.

*Figure 31: Pilot Study and Special Study Results Compared to 2007 NAEP Performance—Grade 8*

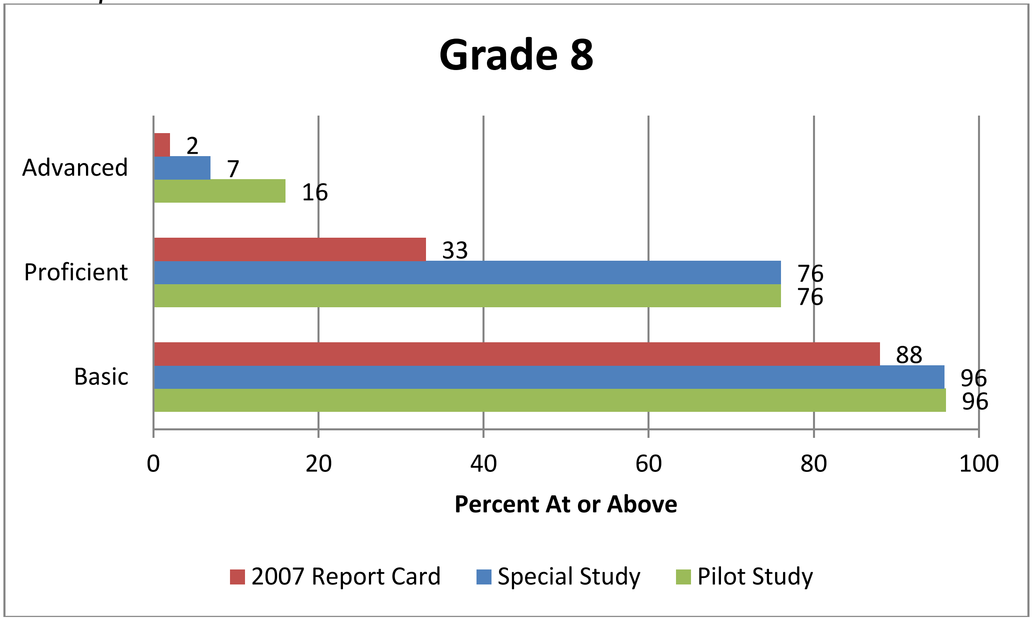
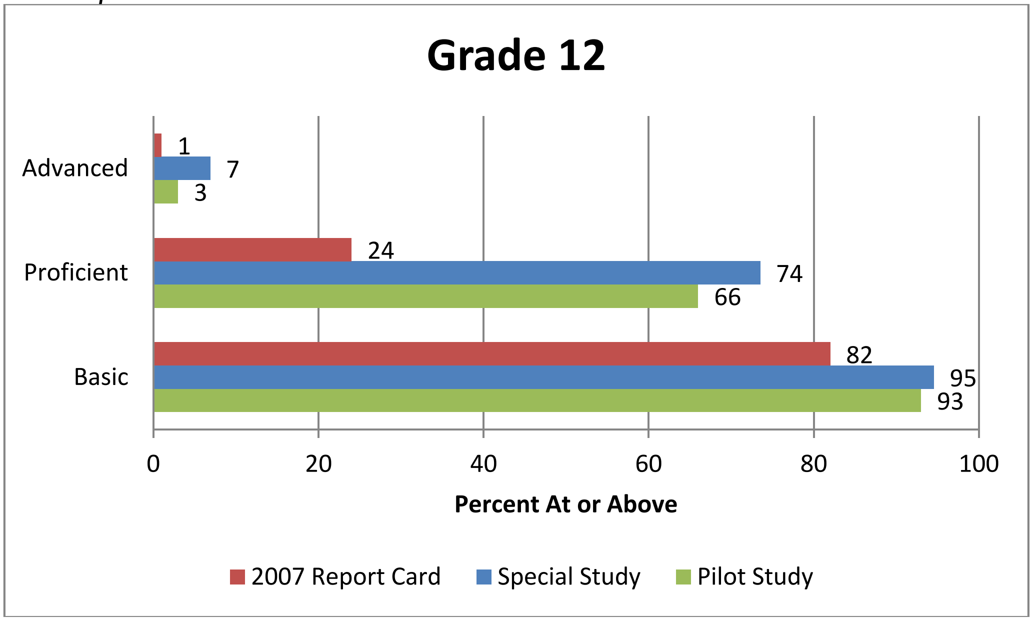


*Figure 32: Pilot Study and Special Study Results Compared to 2007 NAEP Performance—Grade 12*



### *4.3.5.2 Round 3 Classifications*

Pinpointing is the last stage in a traditional implementation of the BoW method. Over the years, standard setters have shied away from implementing this stage in favor of multiple rounds of rangefinding. The reason most cited in the literature is the tremendous logistical challenge associated with selecting and preparing pinpointing student booklets for panelists' use. BoWTIE was developed in response to that challenge. Procedures implemented in the pilot study may be considered proof that technology is key to overcoming the challenge.

After the pinpointing BoWs were selected, prepared, presented to the panelists, and classified into achievement levels, it was concluded that there was no sound way of computing the cut scores. This was supported by the results of the simulation study performed to evaluate cut scores computed using different data inclusion rules. The TACSS then recommended that the third round of classifications be a rangefinding round with a new set of 50 BoWs.

As standard setters and software developers who invested effort in building the pinpointing applications for BoWTIE, we are disappointed that they were not used operationally. As researchers, however, we are encouraged that the process allowed us to remove an obstacle believed by some to hinder the implementation of the BoW method as it was originally designed and intended.

# Chapter 5—Field Trial 2

Numerical results from the pilot study, as well as feedback from panelists, indicated the necessity of revisiting the achievement levels descriptions (ALDs). The purpose of the revisit was to determine whether the ALDs were appropriately calibrated with respect to the Governing Board's policy definitions for Basic, Proficient, and Advanced. Further, some statements appeared to be ambiguous and in need of more precision to be useful criteria for the achievement levels–setting (ALS) process. Prior to implementation of the operational ALS meeting in early February 2012, the Governing Board requested implementation of a small-scale study in which panelists would use the modified ALDs to classify student bodies of work (BoWs).

On January 27, 2012, two weeks before the operational ALS meeting, field trial 2 was implemented at the headquarters of Measured Progress in Dover, New Hampshire. The primary purpose of the field trial was to test the modified ALDs. A one-day study was attended by panelists recruited from within 50 miles of Dover. Measured Progress took this opportunity to also try out the Paper Classification Exercise as part of a Body of Work (BoW) implementation.

Field trial 2 was implemented for both grades 8 and 12 and facilitated by the ALS content and process facilitators. The Governing Board's Contracting Officer's Representative (COR) and a member of the Technical Advisory Committee on Standard Setting (TACSS) attended the meeting.

## 5.1 Field Trial 2 Panelists

As described earlier, panelists for field trial 2 were recruited from within a 50-mile radius of the standard-setting site as a sampling of convenience. It was determined that each panel should consist of the typical 55% teachers, 15% nonteacher educators, and 30% general public. In total, 37 panelists participated in field trial 2, which was the result of recruiting 39 panelists, two of whom cancelled due to weather and illness. While there was interest in maintaining appropriate balances of gender and race/ethnicity, the primary criterion centered on panelist type, which is illustrated in Table 28.

*Table 28: Field Trial 2 Panel Composition*

| Criterion | Attribute | Grade 8 | | Grade 12 | | All | |
|---|---|---|---|---|---|---|---|
| | | **n** | **%** | **n** | **%** | **n** | **%** |
| Panelist Type | Teachers | 10 | 52 | 11 | 61 | 21 | 57 |
| | Nonteacher Educators | 3 | 16 | 2 | 11 | 5 | 14 |
| | General Public | 6 | 32 | 5 | 28 | 11 | 30 |

## 5.2 Field Trial 2 Process

Field trial 2 was implemented as a one-day ALS meeting that concluded with an evaluation form and short debriefing immediately after one round of rangefinding. No feedback was provided to the panelists. All appropriate elements of an ALS meeting were implemented. The process was abbreviated by shortening the general orientation and reducing ALDs training to two stages (general session presentation and Paper Classification Exercise) as opposed to three (Section 4.2.3). All the materials panelists used in the process were the same materials used in the operational ALS meeting. The setup of computers and other equipment was the same as in the operational ALS meeting. The BoWs classified by panelists were the same BoWs classified by pilot study and operational ALS panelists during Round 1 rangefinding. To the extent possible, the field trial 2 process was made consistent with the ALS process. A special process evaluation questionnaire focused on the ALDs was used for this process.

Consistent with the secondary purpose of the meeting in Dover, the Paper Classification Exercise was implemented for this project for the first time. For each of the writing tasks marked for release, panelists were provided one sample response for each score level (1–6) based on the scores assigned during the operational scoring. The papers were ordered randomly relative to the scores, but presented by writing task. The content facilitator led the discussion regarding the level of performance demonstrated on each response. Panelists were reminded that their goal was to gain a common understanding of the ALDs, thus, they had to remember during their discussion that consensus was a goal but not a requirement. After each paper was discussed, panelists were then provided with a piece of paper listing the papers and their respective scores. Facilitators were to point out that correspondence between the scores and their achievement level classifications was not necessarily consistent across the different writing tasks.

## 5.3 Field Trial 2 Results

Results of the evaluation indicated that the process went well and as intended. A summary of panelist responses to the evaluation questionnaire is included in Appendix I.

### *5.3.1 Modified Achievement Levels Descriptions*

Cut scores were computed from the rangefinding classifications the field trial 2 panelists provided. These cut scores, along with the percentages at or above each achievement level, are presented in Table 29. The mean absolute deviations (MADs) from the overall cut scores are presented in Figures 33 and 34. Corresponding results from the pilot study are also provided. All cut scores are provided on the pseudo-NAEP scale. Based only on the numerical results, there seemed to be more parallelism on the modified ALDs used in field trial 2, especially at the Advanced levels.

*Table 29: Cut Scores and Percentages At or Above From Field Trial 2 and Pilot Study*

| Grade | Achievement Level | Pilot Study R1 | | Pilot Study R2 | | Pilot Study R3 | | Field Trial 2 | |
|---|---|---|---|---|---|---|---|---|---|
| | | Cut Score | % At or Above | Cut Score | % At or Above | Cut Score | % At or Above | Cut Score | % At or Above |
| 8 | Basic | 189 | 95 | 189 | 95 | 184 | 96 | 198 | 92 |
| | Proficient | 223 | 77 | 229 | 72 | 225 | 76 | 241 | 60 |
| | Advanced | 282 | 17 | 286 | 15 | 285 | 16 | 281 | 18 |
| 12 | Basic | 511 | 93 | 512 | 93 | 511 | 93 | 511 | 93 |
| | Proficient | 556 | 63 | 553 | 66 | 553 | 66 | 543 | 75 |
| | Advanced | 630 | 2 | 632 | 2 | 628 | 3 | 602 | 15 |

*Figure 33: Field Trial 2 MAD—Grade 8*

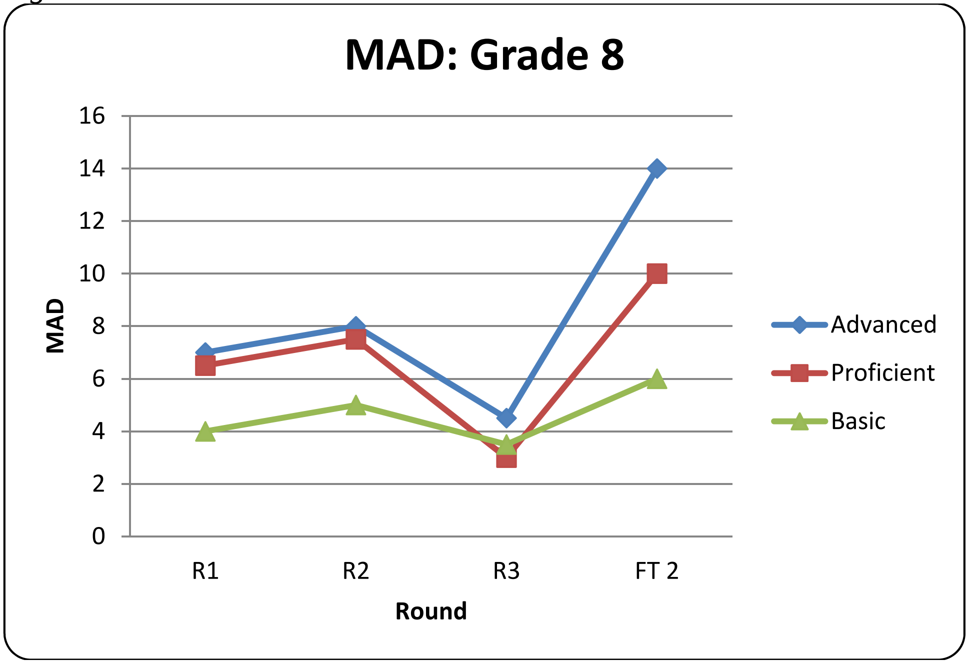
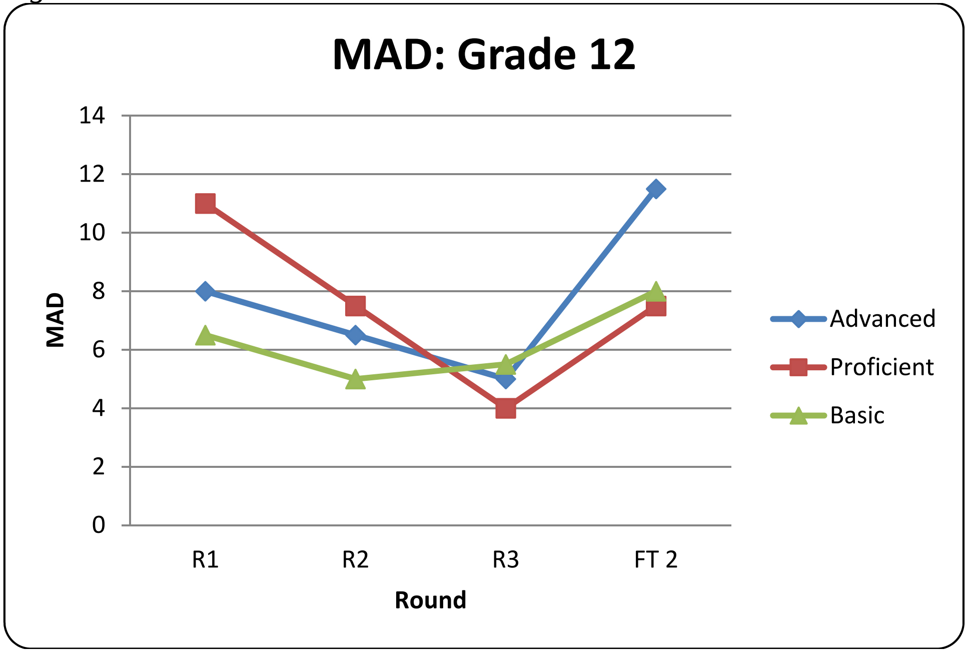


*Figure 34: Field Trial 2 MAD—Grade 12*



Most questions posed to panelists were about the ALDs, using Likert-type questions on an agreement scale. The following summarizes panelist responses:

- All panelists agreed that the description of each achievement level was clear and easy to use, with one or two panelists only somewhat agreeing for each level.
- All panelists agreed that the ALDs represented the Governing Board's Policy Definition of Basic, Proficient, and Advanced level performance.

Panelists were also asked about their perception of the reasonableness of the progressions of the achievement levels. When asked if the ALDs for their grade level provide a reasonable progression of what students should know and be able to do from Basic to Proficient and from Proficient to Advanced, all panelists responded affirmatively to each question. For each achievement level, they were asked if the description for that level for grades 8 and 12 provides a reasonable progression in what students should know and be able to do between these two grades. One panelist responded negatively each of the three times the question was asked. In addition to the Likert questions, panelists were also asked the question "How could the ALDs provide a better description of what students should know and be able to do?" This same question was also asked during the debriefing. Most of panelists felt that the ALDs were already very good and were hard pressed to provide suggestions. Given that most of them were scorers who worked for Measured Progress, they understandably brought the perspective that providing exemplar student responses would enhance their understanding of the descriptions. It was explained to them that that would not be possible since ALS panelists were supposed to be setting the standards based on which exemplars would be selected. Other comments made during the debriefing indicated that some panelists misunderstood the process and their role in the process. For example, panelists commented that they would have done a better job classifying BoWs into achievement levels had they been shown examples of BoWs that exemplify each achievement level. They missed the point that in standard setting, they are to decide what performance in each achievement level should "look like." This might be evidence of the difficulty of covering so much of the process in one day. The panelists might not have received enough instructions to understand that they were not just scoring. Note that most of the field trial 2 panelists had been employed by Measured Progress as scorers for large-scale assessment programs.

The debriefing included discussion between the content facilitators and the panelists on their difficulty in using the ALDs and how the ALDs could be improved upon. Specifically, panelists were asked how the ALDs could be better calibrated with the policy definitions. One topic that arose was the panelists' difficulty in using the Basic achievement level in their classification due to the absence of a Below Basic description. Other discussions were about misuse of some terminologies, such as “partial mastery” versus “partial credit.” The content facilitators left the meeting realizing that the ALDs still required more adjustment to reduce the ambiguity in the language.

### 5.3.2 Paper Classification Exercise

The Paper Classification Exercise was implemented differently in the two grade groups. The grade 12 content facilitator decided that there would not be enough time to discuss 18 papers, so he preselected the papers that he thought would be good to discuss. Also, as papers were being discussed, the panelists were informed what the scores were. The content facilitator even shared his opinion as to what the scores should have been for some papers.

The exercise was implemented as planned in grade 8. For grade 8, the right number of papers was provided for the time allocated for the ALS meeting.

One paper for grade 8 and one paper for grade 12 caused frustration for panelists because they did not agree with the scores assigned to those papers. This issue was presented to the TACSS and they recommended replacing those papers with more suitable ones. Given the purpose of the exercise, the materials presented to panelists should not distract them from the task at hand.

Based on the procedural results from grade 8, the Paper Selection Exercise was implemented in the operational ALS meeting as described in this section. The exercise and its purpose were discussed with the facilitators. The content facilitators participated in clarifying the directions for the exercise and adding them to the facilitator handbook to ensure that the exercise would be implemented as intended.

# Chapter 6—Operational Achievement Levels–Setting Meeting

Sixteen months of preparation led to the operational achievement levels–setting (ALS) meeting. Numerical and procedural results from meetings leading up to the operational meeting were taken into consideration in determining the details of implementation as well as developing the software designed specifically to implement the process as desired. Consultation with the Technical Advisory Committee on Standard Setting (TACSS) and guidance from the Governing Board's Contracting Officer's Representative (COR) led to an implementation considered consistent with the standard settings accomplished by the Governing Board and its previous ALS contractors.

On February 7, 2012, a group of 56 teachers, nonteacher educators, and members of the general public hailing from different states including Alaska and Hawaii gathered at the St. Louis Ritz-Carlton in Clayton, Missouri, to set the writing standards for our nation's youth. These panelists were nominated and selected based on a rigorous process described in Section 2.5. Over the course of four and a half days, the panelists were trained on the 2011 NAEP writing assessment and the achievement levels descriptions (ALDs), and they engaged in the process that resulted in the achievement level recommendations presented to the Governing Board. Immediately after the conclusion of the operational ALS meeting, a special study was held for the purpose of providing the Governing Board information on the relationship of student performance on the 2007 and 2011 NAEP writing assessments.

This chapter describes the panelists, the process, and the results of the operational ALS meeting and the special study. It also describes validity evidence for the achievement levels, touching on procedural validity, internal evidence, and accuracy and consistency of performance classification based on the cut scores.

## 6.1 Panelists

This section describes the panels and panelists associated with the operational ALS meeting. Below are details pertaining to panel composition and panelist accomplishments and awards, illustrating the high caliber of the panelists involved in the operational ALS meeting.

### *6.1.1 Panel Composition*

A total of 55 panelists, 27 for grade 8 and 28 for grade 12, were recruited for the operational ALS meeting. Although the distribution of the selected panelists trended toward the goal percentages, there are three variances worth noting—specifically, gender, race/ethnicity, and representation from the West. First, the percent-ratio of male to female panelists was 25:75 instead of the desired 50:50. Second, 9% of the panelists identified themselves as non-Caucasian compared to a goal of 20%. Third, panelist representation from the West NAEP region was high (36% instead of the intended 20%) while representation from other NAEP regions, most notably the Midwest, was low. Panel composition compared to the demographic distribution required by the Governing Board is presented in Table 30.

*Table 30: Operational ALS Meeting Panel Composition*

| Demographic Variable | Attribute | Grade 8 | | Grade 12 | | All | | Goal |
|---|---|---|---|---|---|---|---|---|
| | | N | % | n | % | n | % | % |
| Panelist Type | Teachers | 16 | 59 | 15 | 54 | 31 | 56 | 55 |
| | Nonteacher Educators | 5 | 19 | 5 | 18 | 10 | 18 | 15 |
| | General Public | 6 | 22 | 8 | 29 | 14 | 25 | 30 |
| Gender | Female | 22 | 81 | 19 | 68 | 41 | 75 | 50 |
| | Male | 5 | 19 | 9 | 32 | 14 | 25 | 50 |
| Race/Ethnicity* | Caucasian | 23 | 85 | 25 | 96 | 48 | 91 | 80 |
| | Non-Caucasian | 4 | 15 | 1 | 4 | 5 | 9 | 20 |
| NAEP Region | Midwest | 6 | 22 | 8 | 29 | 14 | 25 | 35 |
| | Northeast | 5 | 19 | 4 | 14 | 9 | 16 | 20 |
| | South | 6 | 22 | 6 | 21 | 12 | 22 | 25 |
| | West | 10 | 37 | 10 | 36 | 20 | 36 | 20 |

*Two panelists in grade 12 elected not to identify their ethnicity.

Despite these variances, representation on the panel was still fairly broad, which was a clear goal of the recruitment process. Additionally, as illustrated in the Panelist Achievements and Awards subsection, the qualifications of the panelists on this panel were simply excellent.

### 6.1.2 Panelist Achievements and Awards

It has been noted that panelist selection is a critical aspect of any standard-setting study (Cizek & Bunch, 2007; Raymond & Reid, 2001). The panelists in the operational ALS meeting engaged in robust discussion, making keen observations throughout the process, which garnered high praise behind-the-scenes among the staff and observers. In addition to their quality contributions to the study, the panelists' achievements and awards testify to their high qualifications and expertise.

Many of the teacher and nonteacher educator panelists are recipients of Teacher of the Year awards and other excellence awards in recognition of their work. Philip Albonetti (Teacher of the Year 2010) and Mary Richards (College Board-Bob Costas Teaching of Writing Award 2007) serve as two of many examples. Among the general public panelists were a number of accomplished and award-winning authors, including, for example, Thomas B. Allen, author of *Remembering Pearl Harbor*, co-author of *National Geographic's Mr. Lincoln's High-Tech War*, and U.S. Naval Institute's Naval History Author of the Year (2004). Table 31 shows a partial list of the panelist achievements and awards, which attest to the high caliber of the panels participating in this achievement levels setting.

*Table 31: Operational Panelists' Achievements and Awards (Partial List)*

| Panelist Type | Name | Achievements and Awards |
|---|---|---|
| Teacher | [redacted] | Marian University Teacher of the Year, Secondary Ed. (2010-2011) |
| | [redacted] | Upper Peninsula Reading Association's Secondary Reading Teacher of the Year (2005)<br>Michigan Reading Association's Secondary Reading Teacher of the Year (2006) |
| | [redacted] | OACHE Educator of the Year (2007) |
| | [redacted] | Teacher of the Year |
| | [redacted] | NMCTE Excellence in English Education Award – High School (2008) |
| | [redacted] | Teacher of the Year: Joint School District #2 – Idaho (2002-2003) |
| | [redacted] | Circle District Secondary Teacher of the Year |
| | [redacted] | College Board-Bob Costas Teaching of Writing Award (2007) |
| | [redacted] | Jewish Education Services Teacher of the Year (2004) |
| Nonteacher Educator | [redacted] | Gaylord College of Journalism at the University of Oklahoma Women, Communication and Leadership Award (2010) |
| | [redacted] | Master Teacher Award, University of Toledo |
| Authors | [redacted] | Works: Toll Road, Scenic Route, Twisted Road Home |
| | [redacted] | Works: Ricochet River<br>Oregon Book Award for Creative Nonfiction (1996)<br>Pacific Northwest Booksellers Award (1993, 1996)<br>Oregon Library Association's 200 Best All-Time Oregon Books (*Ricochet River* and *Voyage of a Summer Sun*) |
| | [redacted] | Works: Sled Dog Wisdom<br>Contribution to Literacy in Alaska Award (2005) |

continued

| Panelist Type | Name | Achievements and Awards |
|---|---|---|
| Authors | [redacted] | Works: Lost in the River of Grass, The Outside of a Horse, Dolphin Sky<br>American Library Association's Schneider Family Book Award, *Dolphin Sky* (2008) |
| | [redacted] | Voice of Youth Advocates Magazine, one of the best Non-Fiction Young Adult Books, *National Geographic's Mr. Lincoln's High-Tech War* (2009)<br>American Library Association Notable Books, *Remember Pearl Harbor* (2001)<br>U.S. Naval Institute's Naval History Author of the Year (2004) |
| | [redacted] | Colorado Authors League Award finalist, *Riddle at the Rodeo* (2011), *Maria's Mysterious Mission* (2008)<br>25th Annual Highlights for Children Fiction Contest winner (2005) |
| | [redacted] | Arizona Author's Association, honorable mention<br>Los Angeles Book Festival, first runner-up, *Write Your Life Story in 28 Days* (2010) |

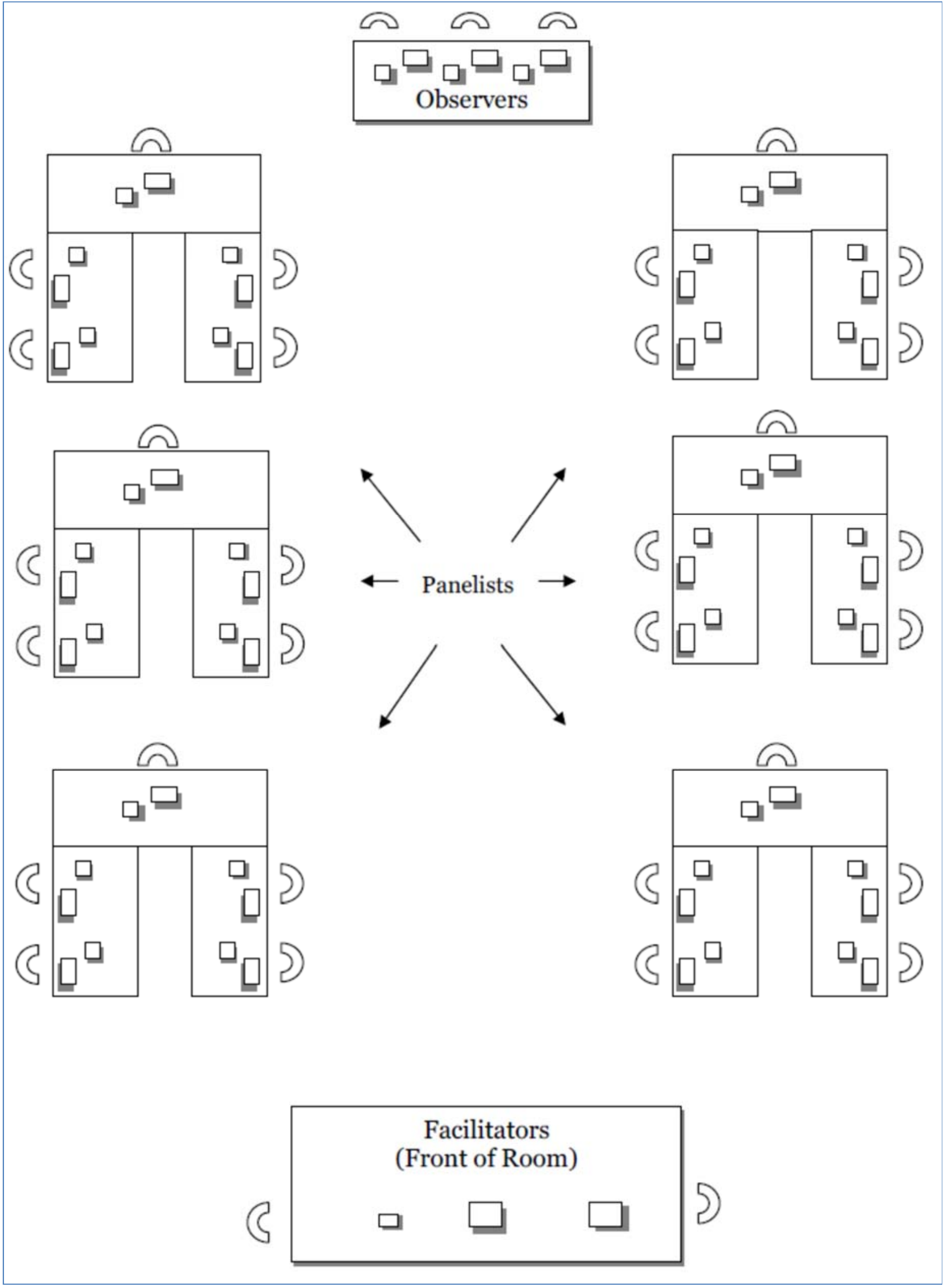
## 6.2 Achievement Levels–Setting Meeting

Fifty-six panelists, 15 staff members, and two observers attended the operational ALS meeting. The process implemented was the culmination of all prior activities. The agenda for the operational ALS meeting is in Appendix A. The details of the implementation were as described in sections 2.9.8 to 2.9.12. The on-site process is also described in brief in the following sections.

### *6.2.1 Preparations*

The operational ALS meeting began on a Tuesday. Staff members arrived in St. Louis as early as Saturday to begin physical preparations. Some computer-based assessment (CBA) laptops were sent to the hotel overnight for a Friday arrival. It was important that the computers were not in transit for too long. The ALS laptops were picked up from a secure storage facility where they had been stored since right after the pilot study. A local secure wired computer network was configured in the hotel in preparation for standard setting. Preparation of the meeting rooms was concluded the day before the ALS meeting. The room configuration for the operational ALS meeting is shown in Figure 35.

*Figure 35:  ALS Room Configuration*



The evening before the operational ALS meeting, the chief of standard setting (CoSS) called a staff meeting attended by all staff members and the Governing Board COR. The first part of the meeting was an overview of the next five days and the expectations of each staff member. The second part of the meeting was a facilitator meeting where the CoSS and the process and content facilitators went over the ALS process with emphasis on parts that were changed relative to the pilot study. Directions to the panelists and the facilitators in the Facilitator Handbook were reviewed, especially the details that had been revised.

During the facilitators' meeting, other staff members worked on the registration table right outside the hotel Amphitheater, registering panelists who arrived earlier that day. About half of the panelists registered then and the other half registered in the morning right before the orientation session. A continental breakfast was provided on the first morning to make sure that panelists arrive at the Amphitheater for the first general session with enough time to get acquainted with other panelists, with whom they would spend many hours in the next few days. At the registration, panelists received hard copies of the 2011 NAEP Writing Framework, the ALDs in the official narrative format and matrix format used for training, the agenda, and the briefing booklet. They had previously received electronic copies of those same materials via e-mail.

### 6.2.2 Panelist Training

The operational ALS meeting began with an orientation at which the CoSS welcomed the panelists and introduced staff and observers. After a few comments on the constitution of the panel and some housekeeping information, the podium was turned over to the Governing Board COR for her welcome remarks. The COR provided information about the Governing Board, NAEP, the writing assessment, and other background information pertinent to setting achievement levels for the 2012 NAEP writing for grades 8 and 12. The podium was then turned back to the CoSS to provide an overview of activities for the next four days. The PowerPoint presentations for the orientation session and the rest of the general sessions are in Appendix E.

The next part of the training was geared at providing panelists familiarity with the assessment. This was accomplished by having them take a form of the assessment and experience how it was to be a student responding to NAEP writing tasks. Panelists used the actual laptop computers used by students during the 2011 NAEP writing administration. With the aid of scoring guides used in operational scoring, panelists reviewed their own responses to gauge what it would take to receive the highest level score in the writing tasks they took.

To continue becoming familiar with the assessment, panelists reviewed the rest of the writing tasks for their grade level, beginning with the ones for which they would be reviewing student responses and classifying bodies of work (BoWs) into achievement levels. It was very important that the panelists see the writing tasks as they were presented to the students, especially those with multimedia stimuli.

In the next stage of the training, panelist were provided an overview of the Body of Work (BoW) method. Details on the rounds of classification and the different feedback information provided between rounds were also given. The Consequences Data Questionnaire and selection of exemplar items were also explained. Panelists were informed about the Body of Work Technological Integration and Enhancements (BoWTIE) software, but were told that a separate general session training would be given prior to using a new application or feature.

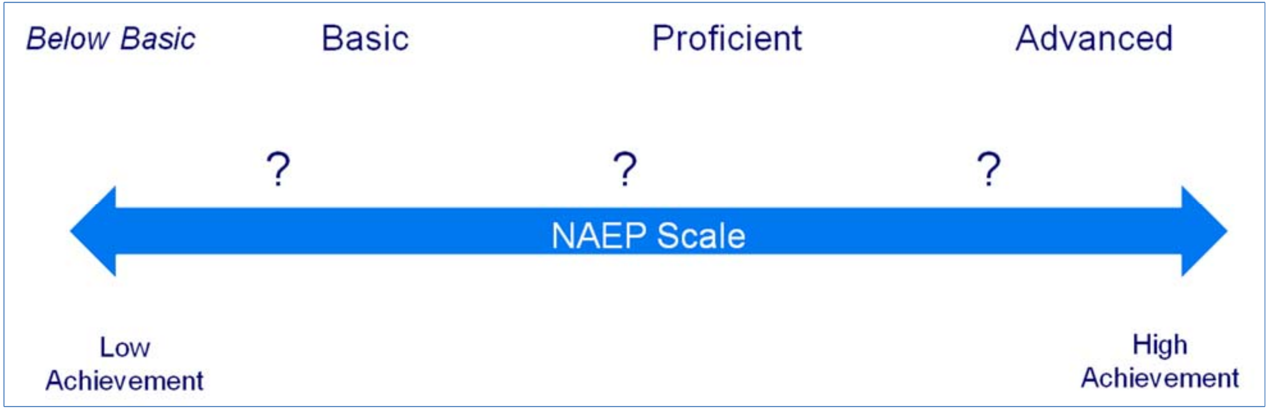
Training on the ALDs began with a general session presentation of the Writing Framework by the content facilitators, who had both been members of the framework steering committees and are deeply familiar with the writing assessment. The facilitators also presented the most recent version of the ALDs, based on the last minor modifications after field trial 2. Both the narrative and matrix versions of the ALDs were presented.

Further training on the ALDs was provided in the grade groups. At the end of the first day, panelists discussed what it meant to perform at each of the achievement levels, taking into consideration the different dimensions of writing. The matrix version of the ALDs was considered a very good tool for discussing and gaining a common understanding of the ALDs. Additionally, the facilitators used some real student responses to show examples of the different dimensions identified in the matrix.

The Response Classification Exercise was the third part of panelists' training on the ALDs. Panelists were provided a response at each score level (1–6) for each of the three writing tasks marked for release, for a total of 18 student responses. For each of the 18 responses, the content facilitator led a discussion in which panelists compared the knowledge, skills, and abilities (KSAs) demonstrated in the response against the KSAs described in each achievement level to determine how to classify each response. In this exercise, panelists were reminded that consensus was a goal but not a requirement. The purpose of the exercise was to engage in discussion that would promote a common understanding of the ALDs.

Prior to the first round of classification, panelists receive general-session training on BoW classification. The training covered both the concept and the mechanics of classifying BoWs into achievement levels. An important concept covered in the presentation was that of a cut score. Graphics presented in Figures 36 and 37 were used to explain the concept to the panelists. In Figure 36, cut scores are scores that delineate between two adjacent achievement levels. In terms of BoW classification, it was explained to the panelists that a cut score is a score associated with a BoW that is the most difficult for them to classify into one of two adjacent levels of achievement. Using the image in Figure 37, it was further explained to the panelists that determining the cut score is akin to asking for the color on any point on the continuum. Wherever they have the hardest time identifying the color on the scale—where they can only answer yellow or red—is the location of the cut score that delineates between yellow and red. Additional information provided in this session included the selection of the BoWs that they would be classifying.

*Figure 36: What are Cut Scores?*



*Below Basic* Basic Proficient Advanced

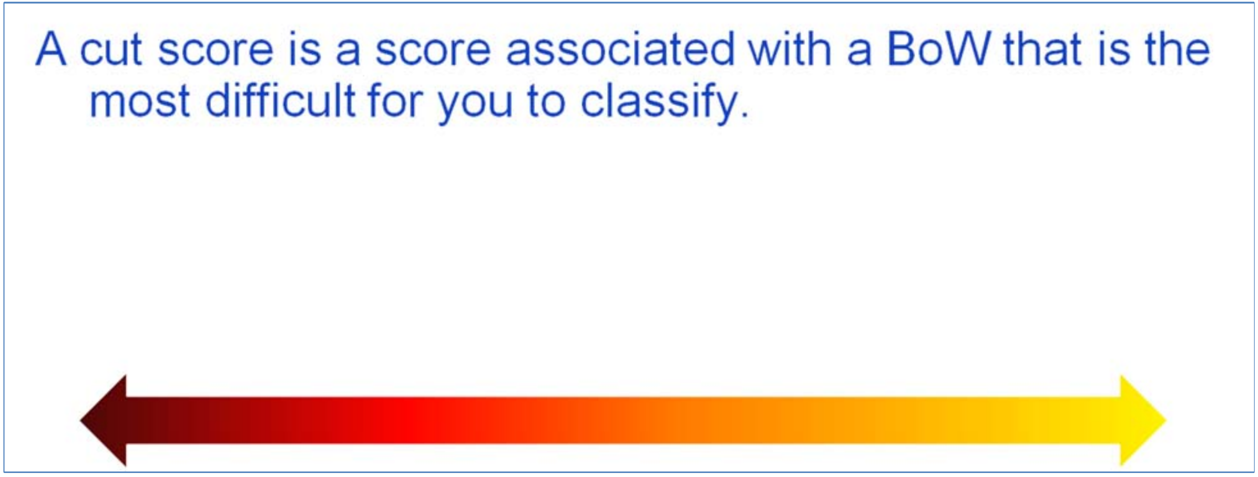
? ? ?

NAEP Scale

Low Achievement

High Achievement

*Figure 37: What is a Cut Score?*



A cut score is a score associated with a BoW that is the most difficult for you to classify.

## 6.3 Rounds of Classification and Feedback[20]

Prior to the first round of classifications, panelists as a group practiced classifying six BoWs into achievement level categories. The process facilitators helped panelists understand the thought process involved in classifying a BoW into an achievement level by ensuring that panelists recognized the following:

- The basis of classification is the match between the KSAs demonstrated in the BoW and the KSAs described in the ALDs.
- Responses to both tasks are considered in the classification.
- It is helpful to keep notes on the reasons for the classifications that panelists make.

Facilitators also guided the panelists in navigating the functionalities of BoWTIE as they accessed the responses, made annotations, and entered their classification data in the database.

For the first round of classifications, panelists were provided 50 BoWs evenly distributed on the score scale and across forms. The BoWs were presented on BoWTIE, rank-ordered from the highest score to the lowest score, but panelists could classify them in any order. Shortly after the last panelist classified all of the 50 BoWs, the cut scores were computed. Feedback shared with panelists after Round 1 to inform their Round 2 classifications consisted of (a) cut scores, (b) a cut score location chart, and (c) a cut score distribution chart. Feedback was first shared with the panelists in the general session, then discussed in the grade-group session. Samples of this feedback are in Figures 15 and 16, and they are described in Section 2.9.9.2. The cut scores are

[20] All feedback information provided to the panelists after each round of classifications are presented in Appendix H.

shown in Table 34[21]. The associated MADs of the cut scores are shown in Tables 47 and 48[22].

Prior to Round 2 classifications, in which panelists reclassify the same BoWs based on feedback they receive from Round 1 results, panelists were provided a tally of their Round 1 classifications. Separate tallies were provided to panelists for both the BoWs common to work groups and BoWs that were not unique to the group. Tables 32 and 33 provide the tallies for common BoWs. The last column indicates which BoWs were discussed by panelists and the order in which they were discussed. The discussion was for the purpose of enhancing panelists' understanding of the ALDs. The selection of the last column and the ordering were based on criteria of entropy developed during the pilot study. These criteria are discussed in Section 2.9.9.2.

[21] Even though the percentages at or above the levels are presented here for each round, note that these percentages were provided to the panelists after the second and third rounds of classifications only.

[22] MADs are included here as a measure of variability of the cut scores. They were not presented to the panelists.

*Table 32: Tally of Round 1 Classifications for Grade 8*

| BoW ID | Counts | | | | Rank | | Discussion Order |
|---|---|---|---|---|---|---|---|
| | Below Basic | Basic | Proficient | Advanced | Group A | Group B | |
| 2xxxxxxx0 | 0 | 0 | 6 | 21 | 3 | 2 | |
| 2xxxxxxx8 | 1 | 7 | 16 | 3 | 6 | 5 | 1 |
| 2xxxxxxx2 | 0 | 2 | 15 | 10 | 7 | 7 | 4 |
| 2xxxxxxx3 | 0 | 0 | 18 | 9 | 8 | 8 | |
| 2xxxxxxx2 | 0 | 2 | 18 | 7 | 10 | 11 | 5 |
| 2xxxxxxx4 | 0 | 5 | 17 | 5 | 11 | 13 | 2 |
| 2xxxxxxx3 | 0 | 2 | 19 | 6 | 13 | 14 | 8 |
| 2xxxxxxx8 | 0 | 9 | 17 | 1 | 17 | 19 | |
| 2xxxxxxx6 | 0 | 13 | 14 | 0 | 21 | 21 | |
| 2xxxxxxx3 | 0 | 15 | 11 | 1 | 22 | 22 | |
| 2xxxxxxx5 | 3 | 15 | 9 | 0 | 28 | 27 | 3 |
| 2xxxxxxx4 | 1 | 18 | 8 | 0 | 27 | 26 | |
| 2xxxxxxx9 | 9 | 16 | 2 | 0 | 29 | 28 | 6 |
| 2xxxxxxx0 | 7 | 19 | 1 | 0 | 30 | 30 | |
| 2xxxxxxx9 | 11 | 16 | 0 | 0 | 34 | 33 | |
| 2xxxxxxx4 | 15 | 12 | 0 | 0 | 36 | 36 | |
| 2xxxxxxx0 | 22 | 5 | 0 | 0 | 37 | 37 | |
| 2xxxxxxx8 | 12 | 15 | 0 | 0 | 38 | 38 | 7 |
| 2xxxxxxx6 | 25 | 2 | 0 | 0 | 41 | 40 | |
| 2xxxxxxx8 | 20 | 7 | 0 | 0 | 44 | 45 | |
| 2xxxxxxx8 | 23 | 4 | 0 | 0 | 48 | 47 | |
| 2xxxxxxx8 | 22 | 5 | 0 | 0 | 49 | 48 | |

*Table 33: Tally of Round 1 Classifications for Grade 12*

| BoW ID | Counts | | | | Rank | | Discussion Order |
|---|---|---|---|---|---|---|---|
| | Below Basic | Basic | Proficient | Advanced | Group A | Group B | |
| 2xxxxxxx0 | 0 | 2 | 5 | 21 | 3 | 2 | 5 |
| 2xxxxxxx2 | 0 | 0 | 15 | 13 | 5 | 4 | |
| 2xxxxxxx4 | 0 | 4 | 18 | 6 | 7 | 5 | 2 |
| 2xxxxxxx11 | 0 | 1 | 15 | 12 | 9 | 9 | |
| 2xxxxxxx9 | 0 | 0 | 19 | 9 | 11 | 10 | |
| 2xxxxxxx4 | 0 | 4 | 20 | 4 | 13 | 14 | 6 |
| 2xxxxxxx4 | 0 | 7 | 19 | 2 | 15 | 15 | |
| 2xxxxxxx8 | 0 | 11 | 16 | 1 | 16 | 16 | |
| 2xxxxxxx9 | 1 | 16 | 10 | 1 | 19 | 20 | 1 |
| 2xxxxxxx3 | 0 | 10 | 12 | 6 | 22 | 23 | 3 |
| 2xxxxxxx0 | 0 | 9 | 16 | 3 | 23 | 25 | 4 |
| 2xxxxxxx8 | 13 | 14 | 1 | 0 | 26 | 26 | |
| 2xxxxxxx9 | 4 | 18 | 6 | 0 | 29 | 29 | 7 |
| 2xxxxxxx8 | 4 | 19 | 5 | 0 | 31 | 33 | |
| 2xxxxxxx6 | 11 | 16 | 1 | 0 | 34 | 34 | |
| 2xxxxxxx4 | 8 | 18 | 2 | 0 | 35 | 35 | 8 |
| 2xxxxxxx7 | 17 | 10 | 1 | 0 | 39 | 39 | |
| 2xxxxxxx4 | 25 | 3 | 0 | 0 | 40 | 40 | |
| 2xxxxxxx1 | 24 | 4 | 0 | 0 | 43 | 43 | |
| 2xxxxxxx3 | 23 | 5 | 0 | 0 | 45 | 44 | |
| 2xxxxxxx0 | 24 | 4 | 0 | 0 | 46 | 45 | |
| 2xxxxxxx0 | 28 | 0 | 0 | 0 | 49 | 49 | |

For Round 2 classifications, panelists were given an opportunity to reconsider their classifications from Round 1 in light of new information. Their Round 1 classifications were preloaded in BoWTIE Round 2 interface. Panelists were informed that they could change some, all, or none of their Round 1 classifications.

For a future research study, panelists were also asked to indicate their level of confidence in their classification for each of the 50 BoWs. They were asked to provide this information as the last line of their comments provided in BoWTIE.

Feedback provided to the panelists after Round 2 consisted of the cut scores, a cut score location chart, and a cut score distribution chart. Additionally, they were provided consequences data. The consequences data were provided in BoWTIE through an interactive tool described in section 2.9.3. The consequences, or impact, data are the percentages at or above the cut scores. The cut scores and percentages from Round 2 are in Table 34.

For Round 3, panelists were provided a brand new set of 50 BoWs to classify into achievement levels. The 50 new BoWs were selected in the same way as the BoWs classified by panelists in Rounds 1 and 2. Panelists were again asked to indicate their level of confidence in each of their classifications.

All feedback provided after Round 2 was updated to provide feedback for Round 3. The cut scores and percentages from Round 3 classifications are shown in Table 34. In the interest of time, the consequences data feedback for Round 3 was given to the panelists in their grade-group room instead of in the general session, as planned.

*Table 34: Cut Scores and Percentages At or Above the Cut Scores*

| Grade | Achievement Level | Round 1 | | Round 2 | | Round 3 | |
|---|---|---|---|---|---|---|---|
| | | Cut Score | % At or Above | Cut Score | % At or Above | Cut Score | % At or Above |
| 8 | Basic | 120 | 80.36 | 120 | 80.60 | 120 | 80.37 |
| | Proficient | 171 | 28.30 | 174 | 25.60 | 173 | 26.77 |
| | Advanced | 216 | 1.89 | 220 | 1.34 | 211 | 3.01 |
| 12 | Basic | 120 | 80.26 | 122 | 79.06 | 122 | 79.05 |
| | Proficient | 170 | 29.81 | 167 | 32.73 | 173 | 26.83 |
| | Advanced | 214 | 2.31 | 213 | 2.59 | 210 | 3.24 |

## 6.4 Final Cut Score Recommendations

After feedback from the third round of classifications was presented, panelists were asked to respond to the Consequences Data Questionnaire (CDQ). There were three primary questions in the CDQ:

> Given your understanding of student performance at the [Basic/Proficient/Advanced] achievement level, does this percentage reflect your expectation about the proportion of students whose NAEP score would be at or above the Basic cut score?
>
> Having seen the data on the percentages of students whose score on the NAEP was at or above the cut score your panel set for each achievement level, would you change one or more of the achievement levels you have set if you could?
>
> What is your final [Basic/Proficient/Advanced] cut score recommendation to the Governing Board? Please enter a scale value keeping in mind that raising the cut score would lead to a smaller percentage of students

scoring at or above the [Basic/Proficient/Advanced] level and lowering the cut score would lead to a larger percentage of students scoring at or above the [Basic/Proficient/Advanced] level.

For grade 8, between 22 and 25 (81–93%) panelists indicated that the percentages at or above each achievement level reflected their expectations. For grade 12, between 26 and 27 (87–90%) panelists indicated that the percentages at or above each achievement level reflected their expectations. For grades 8 and 12, 8 (32%) and 3 (11%) indicated that they would change one or more cut scores if they could. Responses to the third question were deemed unreliable as there were more panelists (grade 8: 25, grade 12: 28) who provided a response to this question than there were panelists who indicated that they would change one or more cut scores if they could (grade 8: 8 (32%), grade 12: 3 (11%)). Panelists had been instructed to skip the third question if they responded negatively to the second question. Further details of the CDQ response summary are included in Appendix I.

On the last process evaluation questionnaire, panelists were also asked if they would sign a statement recommending the cut scores resulting from the ALS process. Only one grade 8 panelist responded negatively.

## 6.5 Exemplar Responses

Based on the Round 3 cut scores, the 16 BoWs (eight from the set used in Rounds 1 and 2 and eight from the set used in Round 3) from the form with two tasks marked for release were classified into achievement levels. For each grade, two were at Advanced, four at Proficient, and six at Basic. Panelists were asked to judge whether each BoW was illustrative of performance at the achievement level to which it was classified. They were asked to rate each BoW as "Very Good," "Okay," or "Do Not Use." They were also asked to comment on their judgments, especially if they rated a BoW "Do Not Use." The summary of panelists' ratings is in Tables 35 and 36.

*Table 35: Summary of Panelists' Ratings of Exemplar BoWs: Grade 8*

| BoW ID | Level | Do Not Use | | OK | | Very Good | |
|---|---|---|---|---|---|---|---|
| | | n | % | n | % | n | % |
| 2xxxxxxx0 | Basic | 14 | 51.8 | 9 | 33.3 | 4 | 14.8 |
| 2xxxxxxx0 | Basic | 5 | 19.2 | 10 | 38.5 | 11 | 42.3 |
| 2xxxxxxx4 | Basic | 1 | 3.8 | 9 | 34.6 | 16 | 61.5 |
| 2xxxxxxx9 | Basic | 2 | 7.4 | 13 | 48.1 | 12 | 44.4 |
| 2xxxxxxx2 | Basic | 3 | 11.5 | 9 | 34.6 | 14 | 53.9 |
| 2xxxxxxx3 | Basic | 3 | 11.1 | 8 | 29.6 | 16 | 59.3 |
| 2xxxxxxx7 | Proficient | 0 | 0 | 14 | 53.8 | 12 | 46.2 |
| 2xxxxxxx3 | Proficient | 2 | 7.7 | 9 | 34.6 | 15 | 57.7 |
| 2xxxxxxx3 | Proficient | 1 | 3.8 | 5 | 19.2 | 20 | 76.9 |
| 2xxxxxxx8 | Proficient | 6 | 22.2 | 6 | 22.2 | 15 | 55.6 |
| 2xxxxxxx0 | Advanced | 0 | 0 | 11 | 42.3 | 15 | 57.7 |
| 2xxxxxxx5 | Advanced | 2 | 7.4 | 12 | 44.4 | 13 | 48.1 |

*Table 36: Summary of Panelists' Ratings for Exemplar BoWs: Grade 12*

| BoW ID | Level | Do Not Use | | OK | | Very Good | |
|---|---|---|---|---|---|---|---|
| | | n | % | n | % | n | % |
| 2xxxxxxx4 | Basic | 11 | 39.3 | 12 | 42.9 | 5 | 17.9 |
| 2xxxxxxx6 | Basic | 9 | 33.3 | 12 | 44.4 | 7 | 25.9 |
| 2xxxxxxx9 | Basic | 6 | 21.4 | 16 | 57.1 | 6 | 21.4 |
| 2xxxxxxx4 | Basic | 7 | 25 | 8 | 28.6 | 13 | 46.4 |
| 2xxxxxxx0 | Basic | 2 | 7.1 | 13 | 46.4 | 13 | 46.4 |
| 2xxxxxxx2 | Basic | 2 | 7.1 | 11 | 39.3 | 15 | 53.6 |
| 2xxxxxxx9 | Proficient | 6 | 21.4 | 19 | 67.9 | 3 | 10.7 |
| 2xxxxxxx4 | Proficient | 1 | 3.6 | 15 | 53.6 | 12 | 42.9 |
| 2xxxxxxx1 | Proficient | 3 | 10.7 | 5 | 17.9 | 20 | 71.4 |
| 2xxxxxxx0 | Proficient | 3 | 10.7 | 15 | 53.6 | 10 | 35.7 |
| 2xxxxxxx2 | Advanced | 5 | 17.9 | 20 | 71.4 | 3 | 10.7 |
| 2xxxxxxx5 | Advanced | 3 | 10.7 | 11 | 39.3 | 14 | 50 |

Based on discussion with the TACSS, one BoW for each achievement level was selected for each grade. The selected BoWs met the following criteria:

- They were rated “Very Good” by almost 50% of the panelists.
- They were rated “Do Not Use” by very few panelists.

The TACSS also took into consideration the actual student responses and panelists' comments on the responses.

When the achievement levels were presented to the Committee on Standards, Design and Methodology (COSDAM), it was recommended that one of the selected exemplar BoWs be replaced due to some perceived bias. A suitable replacement was found, using the same criteria specified by the TACSS.

## 6.6 Validity Evidence

In an endeavor that relies primarily on informed judgment, validity evidence relies primarily on the design of the process and the fidelity of implementation with respect to the design, as well as indicators of internal consistency of judgments within and among panelists. Procedural validity stems from evidence indicating that procedures are reasonable, were carried out as intended, and were understood by panelists. Internal validity stems from evidence centered on comparisons of results using exactly the same methods on different occasions and on the variability of the cut scores across rounds and groups (Hambleton, Pitoniak, & Coppella, 2012). This section contains a collection of evidence documenting the procedural and internal validity of the ALS results.

### *6.6.1 Procedural Validity*

Procedural validity is the degree to which the entire process is tightly interwoven with strong connections between every component part (Reckase, 2001). For setting achievement levels for the 2011 NAEP writing, evidence of procedural validity lies on process documentation and supported by process evaluation.

#### *6.6.1.1 Process Documentation*

Procedural evidence of validity through process documentation requires

- a sound design document supported by stakeholders
- that the process was implemented as designed
- a thorough documentation of process implementation

(Reckase, M.D., personal communication, October 4, 1993).

The Design Document (Measured Progress, 2011) fully describes the procedures implemented in the ALS process. It represents the best thinking and strategies in process implementation for setting achievement levels for the 2011 NAEP writing for grades 8 and 12. The ALS process described in the Design Document carries with it the rich tradition and rigor of ALS processes implemented in previous NAEP assessments along with state-of-the-art technological enhancements developed for the current process. Public comments on the Design Document were solicited February 10—24, 2011 through notices on the Federal Register and on websites for Measured Progress, the Governing Board, and WestEd.  In addition, a notice on the WestEd Facebook page was directing individuals to a WestEd landing page from which the design document and questions were downloadable. Furthermore, email solicitations were sent to directors of relevant organizations as listed in the Design Document on pages 82—83. These organizations include key organizations that collaborated in the review of the 2011 Writing Framework, Common Core Standards consortia and stakeholders, and others. Comments received were shared and discussed with the TACSS. None of the comments warranted a modification to the Design Document.

The process implementation described in this Process Report and the accompanying Technical Report are consistent with the original design. Additional details in implementation and any deviation from the design[23] are fully described in the Process and Technical Reports. Decisions and discussions leading to modifications are documented in the TACSS meetings summaries.[24]

#### 6.6.1.2 Process Evaluation

Process evaluation questionnaires were designed to provide feedback on how to improve the process but also provide evidence for evaluating procedural validity. The full text of the relevant survey questions and their response options are located in Appendix I. Most items were scored on a 5-point scale. The response options for each question varied (see Appendix I), but in general scores above 3 indicate stronger support for the validity argument, scores at 3 are somewhat neutral, and scores below 3 suggest a lack of validity. Relevant items on certain topics were selected from the evaluations. Items that were not relevant to the highlighted topics can be found in Appendix I.

*Advanced materials were received and adequate.* Panelists stated that they received advanced materials and that the materials were adequate (Table 37). The average survey response for the question "Materials received" for both the grade 8 and grade 12 panelists was 5, which indicates that panelists received the advanced materials. Also, the average response for the "Materials adequate" for both grade 8 and

[23] The deviations from the design include

1. Allowing at-large nominators for general public panelists
2. Replacing the pinpointing in round 3 with rangefinding using a new set of BoWs
3. BoWs were presented to the operational ALS panelists in rank order from highest to lowest scores

[24] All TACSS meeting summaries are in Appendix A of the Technical Report.

grade 12 panelists was 4, indicating that there was general agreement about the materials being adequate.

*Overview and purpose was clear.* During the workshop, participants were introduced to the method. In order for the panelists to engage in the process, it is important for them to understand the process. Their understanding supports procedural validity. Participants' average response indicated that the method overview was clear (Table 37, Question 11). Participants also indicated that the workshop purpose was somewhat clear at the beginning (Table 37, Question 17) but very clear at the end of the workshop (Table 37, Question 31). The fact that the panelists reported that the workshop purpose was very clear by the end of the meeting indicates that panelists had some understanding of the process. This supports the procedural validity of the ALS process.

Procedural validity is also supported when participants understand the context in which they are to make decisions and perform tasks. During the meeting, panelists were given an introduction to the NAEP writing test and the NAEP writing framework development. Participants reported that both of these elements were clear (Table 37, Questions 6 & 13).

*Table 37: Materials and Overview Evaluation Summary*

| Question* | Evaluation Number | Grade 8 | | Grade 12 | |
|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD |
| Q1. Materials received | 1 | 5.0 | 0.2 | 5.0 | 0.0 |
| Q2. Materials adequate | 1 | 4.0 | 1.0 | 4.0 | 0.9 |
| Q6. NAEP explanation | 1 | 4.1 | 0.6 | 4.3 | 0.5 |
| Q11. Method Overview | 1 | 4.0 | 0.6 | 4.2 | 0.6 |
| Q13. NAEP writing framework development | 1 | 4.1 | 0.6 | 4.2 | 0.6 |
| Q17. Workshop purpose | 1 | 3.0 | 0.5 | 3.0 | 0.7 |
| Q31. Workshop purpose | 5 | 4.7 | 0.5 | 4.8 | 0.4 |

*The actual process evaluation questions may be found in Appendix I: Evaluation Summaries

*Supporting panelists in understanding of their tasks.* To further support the procedural validity claim, it must be clear that the panelists were given clear instructions and that they understood their tasks. The results from the survey show that the panelists were given clear instructions for the task review step (Table 38, Question 9) and that their role in the task review as clear (Table 38, Question 10). Furthermore, at the end of the process, panelists reported that the BoW instructions were clear (Table 38, Question 26).

*Table 38: Instructions and Role Description Evaluation Summary*

| Question* | Evaluation | Grade 8 | | Grade 12 | |
|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD |
| Q9. Task review instructions | 1 | 3.9 | 0.6 | 4.2 | 0.7 |
| Q10. Task review role description | 1 | 4.3 | 0.6 | 4.4 | 0.6 |
| Q26. BoW instructions | 5 | 4.4 | 0.6 | 4.0 | 0.8 |
| Q44. Instructions clear | 5 | 4.7 | 0.5 | 4.6 | 0.5 |

*The actual process evaluation questions may be found in Appendix I: Evaluation Summaries

*Practical experiences during the panel meeting were helpful.* Panelists were provided with several practical experiences to enhance their understanding and skills during the meeting. For example, panelists took the NAEP under realistic testing conditions (Table 39, Questions 7 & 8). Panelists also reviewed student responses (Table 39, Question 1) and practiced classifying students (Table 39, Question 4). To support the procedural validity claim, it should be true that these experiences were helpful to the panelists. The results of the survey show that the panelist found all of these experiences to be helpful. They also reported feeling confident about their roles after taking part in the practical experiences (Table 39, Question 5).

*Table 39: Practical Experiences Evaluation Summary*

| **Question*** | **Evaluation** | **Grade 8** | | **Grade 12** | |
|---|---|---|---|---|---|
| | | **Mean** | **SD** | **Mean** | **SD** |
| Q7. Value of taking NAEP | 1 | 4.7 | 0.8 | 4.7 | 0.4 |
| Q8. Authenticity of taking NAEP | 1 | 4.2 | 0.9 | 4.6 | 0.6 |
| Q1. Reviewing student responses | 2 | 4.9 | 0.4 | 4.7 | 0.5 |
| Q4. Practice classification | 2 | 4.7 | 0.6 | 4.6 | 0.5 |
| Q5. Practice confidence | 2 | 4.6 | 0.8 | 4.5 | 0.6 |

*The actual process evaluation questions may be found in Appendix I: Evaluation Summaries

*Panelists understood ALDs.* A crucial part of procedural validity is missing if it is not clear that the panelists understood the ALDs. In order to perform their tasks in such a way as to produce valid results, the panelists must understand the writing framework and its details (Table 40, Questions 14 & 15). It is also important for the panelists to operate from a common understand of this framework and the accompanying ALDs (Table 40, Questions 2 & 15) and that this understanding remain consistent across the various tasks involved in the meeting (Table 40, Questions 2, 3, & 15). The results from the survey indicate that these points were clear to the panelists and most agreed that there was a common understanding of the ALDs.

*Table 40: Understanding of ALDs Evaluation Summary*

| Question* | Evaluation | Grade 8 | | Grade 12 | |
|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD |
| Q14. Writing framework | 1 | 4.2 | 0.6 | 4.3 | 0.5 |
| Q15. Framework detail | 1 | 4.0 | 0.6 | 4.1 | 0.7 |
| Q2. Panel agreement | 2 | 4.0 | 0.7 | 4.0 | 0.8 |
| Q3. ALD understanding | 2 | 4.6 | 0.6 | 4.4 | 0.7 |
| Q15. Panel agreement | 3 | 4.0 | 0.8 | 3.7 | 0.8 |

*The actual process evaluation questions may be found in Appendix I: Evaluation Summaries

*Panelists understood the method*. The validity of the process is supported when panelists understand the method and its outputs. Panelists indicated that they understood the method overview (Table 41, Question 11). They also indicated that they understood how the group cut scores were produced after each stage in the standard-setting process (Table 41, Questions 1, 5, & 11).

*Table 41: Understanding of Methods Evaluation Summary*

| Question* | Evaluation | Grade 8 | | | Grade 12 | | |
|---|---|---|---|---|---|---|---|
| | | Mean | SD | % at 3 or below | Mean | SD | % at 3 or below |
| Q11. Method overview | 1 | 4.0 | 0.6 | 4<br>(15%) | 4.2 | 0.6 | 3<br>(10%) |
| Q1. Understand group cut scores | 3 | 4.3 | 0.8 | 4<br>(15%) | 4.5 | 0.6 | 1<br>(4%) |
| Q5. Understand group cut scores | 4 | 4.2 | 1.0 | 3<br>(11%) | 4.5 | 0.5 | 1<br>(4%) |
| Q11. Understand group cut scores | 5 | 4.5 | 0.8 | 2<br>(8%) | 4.3 | 0.9 | 3<br>(11%) |

*The actual process evaluation questions may be found in Appendix I: Evaluation Summaries

*Panelists understood tasks.* Panelists must understand their tasks in order to perform them in a valid way. Several survey questions targeted panelists' understanding of specific tasks such as the exemplar task (Table 42, Question 27) and the consequences task (Table 42, Question 21). Questions also addressed whether panelists understood how to access and interpret tools they were given to perform certain tasks. For instance, panelists were required to access and use a panel cut score distribution chart and a panel cut score location chart during several steps in the process. Participants indicated that they understood how to access these tools (Table 42, Question 18) and how to interpret them (Table 42, Question 9, 8, 14, 11, 12, & 17). Because the average of the survey results (means ranged from 4.0 to 4.7) indicated that panelists understood the various tasks they were required to perform and the associated tools, the procedural validity of the ALS process is supported.

*Table 42: Understanding of Tasks Evaluation Summary*

| **Question*** | **Evaluation** | **Grade 8** | | **Grade 12** | |
|---|---|---|---|---|---|
| | | **Mean** | **SD** | **Mean** | **SD** |
| Q27. Perform exemplar task | 5 | 4.6 | 0.6 | 4.3 | 0.8 |
| Q18. Understand how to access panel cut score location chart | 5 | 4.6 | 0.6 | 4.7 | 0.5 |
| Q9. Understand panel cut score distribution chart | 3 | 4.0 | 0.8 | 4.4 | 0.7 |
| Q8. Understand panel cut score distribution chart | 4 | 4.3 | 0.8 | 4.5 | 0.5 |
| Q14. Understand panel cut score distribution chart | 5 | 4.5 | 0.8 | 4.6 | 0.6 |
| Q11. Understand panel cut score location chart | 3 | 4.5 | 0.6 | 4.6 | 0.6 |
| Q12. Understand panel cut score location chart | 4 | 4.4 | 0.9 | 4.5 | 0.5 |
| Q17. Understand panel cut score location chart | 5 | 4.6 | 0.6 | 4.7 | 0.6 |
| Q21. Understand consequences | 5 | 4.3 | 0.8 | 4.4 | 0.7 |

*The actual process evaluation questions may be found in Appendix I: Evaluation Summaries

*Panelists recognized achievement levels for each cut score.* Panelists should also have a clear understanding of the connection between the cut scores produced and the associated achievement levels in order for the ALS process to be valid. Panelists were asked to respond to this question after each standard-setting round (Table 43, Questions 2, 6, & 12). The results show that the panelists' understanding increased after each round, and that by the end of the meeting, they reported clear understanding of the cut scores and their connection to the achievement levels (Table 43, Question 12).

*Table 43: Understanding of Achievement Levels and Cut Scores Evaluation Summary*

| **Question*** | **Evaluation** | **Grade 8** | | **Grade 12** | |
|---|---|---|---|---|---|
| | | **Mean** | **SD** | **Mean** | **SD** |
| Q2. Understand achievement levels and cut scores | 3 | 3.9 | 0.7 | 4.3 | 0.5 |
| Q6. Understand achievement levels and cut scores | 4 | 4.0 | 0.7 | 4.5 | 0.5 |
| Q12. Understand achievement levels and cut scores | 5 | 4.5 | 0.6 | 4.4 | 0.6 |

*The actual process evaluation questions may be found in Appendix I: Evaluation Summaries

*Process produced confident panelists.* The ALS procedures were designed to produce valid cut scores. However the validity claim would not be supported if the designed procedures were not executed as intended. Two major areas of this aspect of procedural validity were addressed by the survey. First, the ALS process was designed to support panelists in their tasks and to produce panelists who can make confident decisions and be confident in executing their tasks. Second, the ALS procedure was designed to produce cut scores that the panelists believed to be meaningful, reasonable, and defensible.

Panelists were asked to report on their confidence level with regard to several tasks involved in the ALS process. Panelists reported confidence in using the panel cut score distribution chart at each stage (Table 44, Questions 9, 9, & 15) and confidence in using the panel cut score location chart (Table 44, Questions 12, 14, & 18). They also reported confidence in the Round 2 (Table 44, Question 2) and Round 3 (Table 44, Question 6) classifications, as well as confidence in using the consequences data (Table 44, Question 22). The confidence level of the panelists (means from 4.1 to 4.7) suggests that the procedures were able to support panelists in their roles and tasks. Panelists were also asked whether they found any particular classification decisions to be difficult. The average panelist was between “disagree” and “somewhat agree,” reporting that basic classification was not difficult or only somewhat difficult (Table 44, Question 2), Proficient classification was not difficult (Table 44, Question 3), and Advanced classification was not difficult (Table 44, Question 4). These results further support the argument that the procedures were carried out as intended to produce confident and competent panelists.

*Table 44: Panelist Confidence Evaluation Summary*

| **Question*** | **Evaluation** | **Grade 8** | | **Grade 12** | |
|---|---|---|---|---|---|
| | | **Mean** | **SD** | **Mean** | **SD** |
| Q9. Confidence in using panel cut score distribution chart | 3 | 4.0 | 0.8 | 4.4 | 0.7 |
| Q9. Confidence in using panel cut score distribution chart | 4 | 4.0 | 0.8 | 4.5 | 0.5 |
| Q15. Confidence in using panel cut score distribution chart | 5 | 4.4 | 0.8 | 4.4 | 0.6 |
| Q12. Confidence using panel cut score location chart | 3 | 4.6 | 0.6 | 4.7 | 0.5 |
| Q14. Confidence using panel cut score location chart | 4 | 4.3 | 0.9 | 4.5 | 0.5 |
| Q18. Confidence using panel cut score location chart | 5 | 4.6 | 0.6 | 4.7 | 0.5 |
| Q2. Basic classification difficult | 5 | 2.8 | 1.0 | 2.7 | 1.1 |
| Q3. Proficient classification difficult | 5 | 2.5 | 0.7 | 2.7 | 0.8 |
| Q4. Advanced classification difficult | 5 | 2.4 | 1.1 | 2.4 | 0.9 |
| Q2. Confidence classification 2 | 4 | 4.1 | 0.8 | 4.5 | 0.5 |
| Q6. Confidence classification 3 | 5 | 4.1 | 0.6 | 4.3 | 0.5 |
| Q22. Confidence using consequences data | 5 | 4.2 | 0.7 | 4.4 | 0.6 |

*The actual process evaluation questions may be found in Appendix I: Evaluation Summaries

*Process produces reasonable, defensible scores.* In going through the ALS process, panelists accumulated experience using the cut scores. In order for there to be strong procedural validity, it should be true that the panelists find each stage, each output, and each tool to be reasonable. Panelists were asked to comment on certain elements of the process in the feedback evaluations. The average response of the panelists shows that panelists agreed that the process resulted in defensible levels (Table 45, Question 32) as well as reasonable levels (Table 45, Question 34). Panelists also reported having used their best judgment (Table 45, Question 37). Finally, the panelists reported that they found the process to be inclusive (Table 45, Question 38). These findings all support the procedural validity claim.

*Table 45: Face Validity Evaluation Summary*

| Question* | Evaluation | Grade 8 | | Grade 12 | |
|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD |
| Q32. Defensible levels | 5 | 4.6 | 0.6 | 4.7 | 0.5 |
| Q34. Reasonable levels | 5 | 4.2 | 0.6 | 4.7 | 0.4 |
| Q37. Best judgment | 5 | 4.7 | 0.6 | 4.7 | 0.6 |
| Q38. Inclusive | 5 | 4.5 | 0.8 | 4.6 | 0.9 |

*The actual process evaluation questions may be found in Appendix I: Evaluation Summaries

### 6.6.2 Internal Validity

The design of the ALS process allows the internal validity of the cut scores to be measured in several ways. First, because panelists received extensive training on the ALDs, they were able to focus their understanding on what students (for a given cut score) should know and be able to do. Thus, from one round to the next there should be less variability among the panelists in the location of their cut scores. This confirmatory approach can be used to establish evidence of internal validity. The second approach compares results from the same procedures when different groups of panelists are used. Subsequent to each meeting, the sets of cut scores from each group were obtained and compared. When similar cut scores are obtained, this suggests that the procedures yield valid and reliable cut scores.

Variability of the scores can be quantified in a few ways. The data can be examined for changes, average deviance of cut scores can be calculated, and the standard error of the cut score can be calculated. Each one of these methods is used to describe the variability of the cut scores across rounds and panels.

A summary of the individual panelist cut score changes between rounds provides preliminary information about the direction in which cut scores varied across rounds. Table 46 reports the number of panelists whose cut scores increased, decreased, or had no change from the previous round for grades 8 and 12. Changes between Rounds 1 and 2 are labeled “R1:R2,” while changes between Rounds 2 and 3 are labeled “R2:R3.”

*Table 46: Round-to-Round Cut Score Changes by Grade*

| Grade | Round | Achievement Level | Increased | | Decreased | | Change | | No Change | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | N | % | n | % | N | % | N | % |
| 8 | R1:R2 | Advanced | 8 | 29.63 | 6 | 22.22 | 14 | 51.9 | 13 | 48.15 |
| | | Proficient | 12 | 44.44 | 11 | 40.74 | 23 | 85.2 | 4 | 14.81 |
| | | Basic | 12 | 44.44 | 11 | 40.74 | 23 | 85.2 | 4 | 14.81 |
| | R2:R3 | Advanced | 8 | 29.63 | 18 | 66.67 | 26 | 96.3 | 1 | 3.7 |
| | | Proficient | 12 | 44.44 | 15 | 55.56 | 27 | 100 | 0 | 0 |
| | | Basic | 12 | 44.44 | 15 | 55.56 | 27 | 100 | 0 | 0 |
| 12 | R1:R2 | Advanced | 10 | 35.71 | 12 | 42.86 | 22 | 78.6 | 6 | 21.43 |
| | | Proficient | 11 | 39.29 | 12 | 42.86 | 23 | 82.1 | 5 | 17.86 |
| | | Basic | 11 | 39.29 | 12 | 42.86 | 23 | 82.1 | 5 | 17.86 |
| | R2:R3 | Advanced | 8 | 28.57 | 20 | 71.43 | 28 | 100 | 0 | 0 |
| | | Proficient | 14 | 50 | 14 | 50 | 28 | 100 | 0 | 0 |
| | | Basic | 11 | 39.29 | 14 | 50 | 25 | 89.3 | 3 | 10.71 |

Table 46 illustrates that most panelists changed their cut scores from Round 1 to Round 2 and from Round 2 to Round 3. Table 46 illustrates that the proportion of panelists who changed their cut scores was usually greater for R2:R3 than for R1:R2 (except for grade 12, Advanced). Typically we would expect the number of changes to decrease across rounds to support the internal validity argument. However, in our procedures, a new sample of BoWs was drawn for Round 3, meaning the panelists were classifying a completely new set of student work samples. So the increased number of changes in R2:R3 is to be expected.

The tables indicate whether the cut score changed. This indicates that the panelists were recalibrating their cut scores. However, the goal is to achieve some narrowing in the variability of the cut score within each group. The amount of variability in the cut score for each group can also be summarized using a statistic such as the standard deviation. After the standard-setting meetings, cut scores were calculated. Panel cut scores were calculated by obtaining the median panelist cut scores within a panel. Therefore, describing variation of the cut scores within a panel using a standard deviation calculation is not appropriate. Instead, variation is described in terms of mean absolute deviation (MAD) indices.

The MAD is the average difference between each panelist's cut score and the median cut score. Tables 47 and 48 report MAD for each classification round for the grade 8 and grade 12 panels, respectively.

*Table 47: Mean Absolute Deviation (MAD) by Round—Writing Grade 8*

| **Achievement Level** | **MAD** | | |
|---|---|---|---|
| | **Round 1** | **Round 2** | **Round 3** |
| Basic | 12 | 9 | 9 |
| Proficient | 11 | 8 | 7 |
| Advanced | 12 | 5 | 9 |

*Table 48: Mean Absolute Deviation (MAD) by Round—Writing Grade 12*

| **Achievement Level** | **MAD** | | |
|---|---|---|---|
| | **Round 1** | **Round 2** | **Round 3** |
| Basic | 10 | 5.5 | 7 |
| Proficient | 14 | 7.5 | 7 |
| Advanced | 7.5 | 8.5 | 4 |

As the tables show, the variability of cut scores generally decreased from Round 1 to Round 3. For the grade 8 Basic achievement level, the MAD decreased from 12.0 to 9.0. For the grade 8 Proficient Level, the MAD decreased from 11 .0 to 7.0. For the grade 12 Proficient level, the MAD decreased from 14.0 to 7.0. For the grade 8 Advanced level, the MAD decreased from 12.0 to 9.0 overall, although there was a smaller MAD in Round 2. The same overall decrease is seen for grade 12 Basic (10.0 to 7.0), with a smaller Round 2 MAD. For the grade 12 Advanced Level, there was an overall decrease in MAD (7.5 to 4.0) with a larger MAD in Round 2. Although there were some differences in how the MAD decreased across rounds, all of the Round 3 MADs are the smallest MADs for the set. This illustrates that the variability at the end of the process was indeed the smallest. This supports the internal validity claim by showing that the process resulted in less variability among panelists by Round 3.

As noted above, the median was used as the panel cut score in this standard-setting process. Therefore, the usual standard error calculation, which uses the mean, does not give an accurate measure of the variability of the cut score. Since the underlying shape of the distribution of the cut scores is unknown, estimates of variation must be based on approximations. Two approximations are used to calculate the cut score standard error.

The first approximation is based on the Maritz-Jarrett procedure (Maritz & Jarrett, 1978). This procedure provides an empirically estimated standard error for any percentile. The second estimator of the standard error of the median is based on the bootstrap technique (Efron & Gong, 1983). In this procedure, repeated samples with replacement are taken from the original distribution of cut scores, and the median is calculated for each resample. The standard deviation of these medians is then calculated and used as the estimate. In this case, 1,000 samples were created.

Tables 49 through 54 present these standard error estimates for grades 8 and 12, respectively, across panelist demographic groupings, tables, and groups (i.e., A and B) at each round. As described in the Technical Report, panelists were arranged into tables and tables into groups in such a way as to minimize difference in the tables and groups. Thus, the summary statistics reported for each table and group within a certain grade and round should be comparable to each other. The tables are arranged such that the summary statistics are reported for each table, each group (which comprises 3 tables), and for the entire group (All). This information is reported once for each round and grade in a separate table. To give evidence in support of the internal validity of the ALS process, it should be the case that the MADs and standard errors (SEs) are similar across tables and groups. This gives evidence that the variability in the cut score due to choosing specific panelists is minimized. Despite efforts to create equivalent tables and groups by minimizing the difference in these groups, some variability in MADs and SEs is seen across tables and groups. It would be expected that these differences across groups would decrease after each round. The variability in the SEs across tables and groups is greatest in Round 2 (Tables 50 and 53) or Round 1 (Tables 49 & 52) but the variability of the SEs across groups decreases in Round 3 (Tables 51 & 54), as expected. It should also be case that the overall MADs and SEs decrease across rounds. Comparing the MADs and SEs for the group "All" across rounds for grade 8 (Tables 49, 50, & 51) and for grade 12 (Tables 52, 53, & 54) show that the average MADs and SEs over achievement levels generally decreases. The narrowing of the SE band can also be seen visually in Figures 39 and 40, which show the cut scores and their associated SE bands across rounds. Because of the general pattern of decrease in variability across groups and decrease in overall MADs and SEs, the validity argument is supported.

Figure 39: Grade 8 Cut Scores Across Rounds

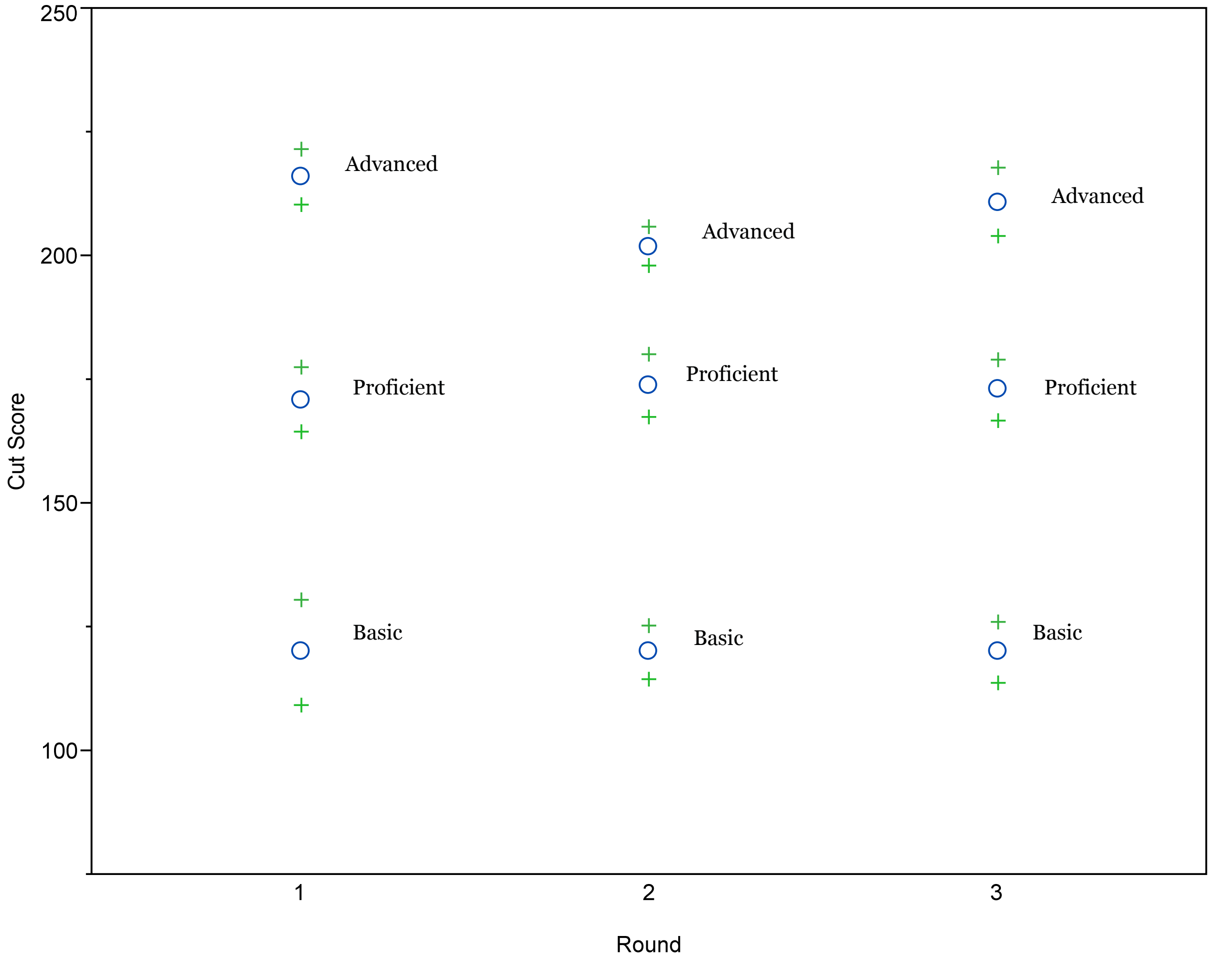


Figure 40: Grade 12 Cut Scores Across Rounds

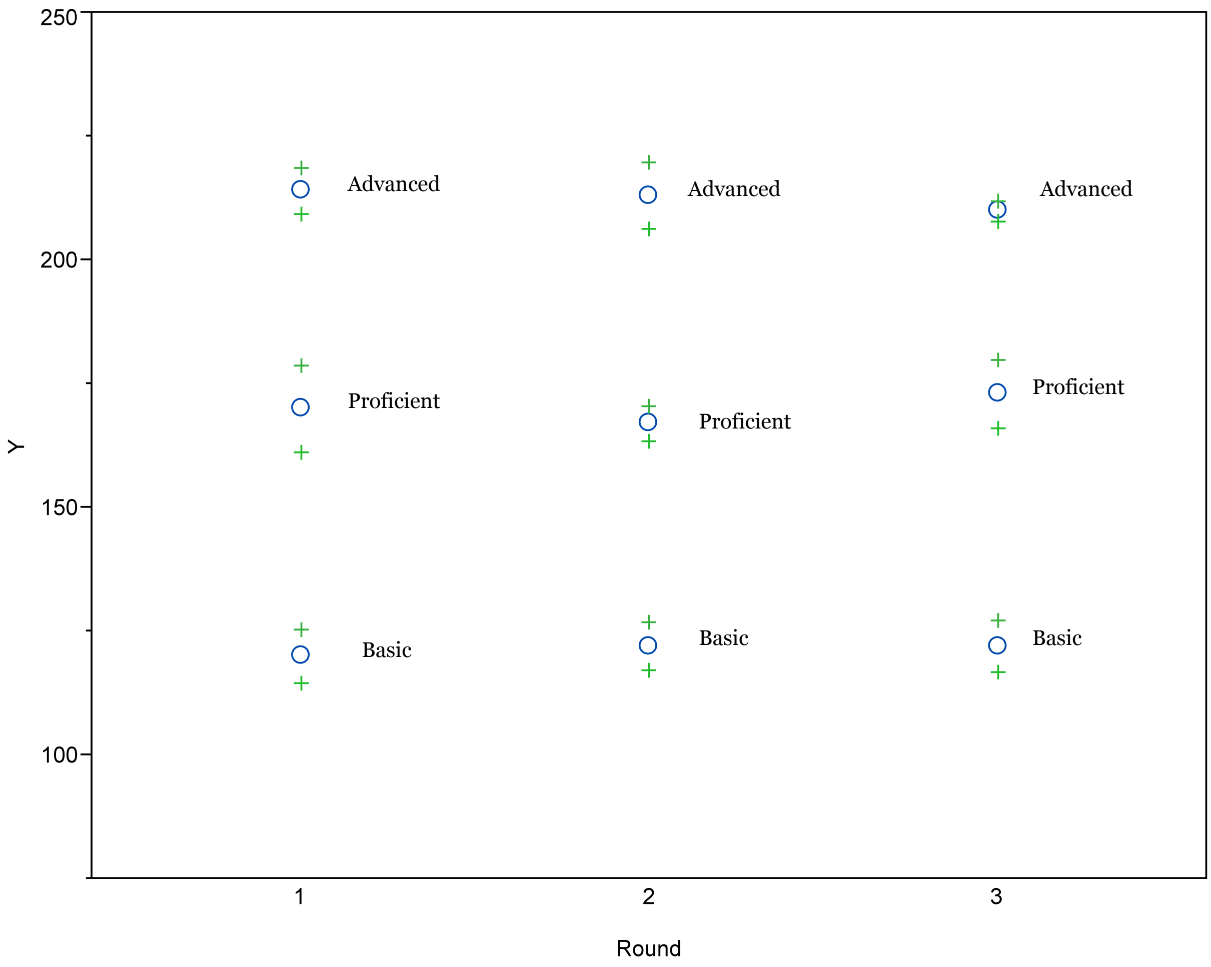


*Table 49: Estimates of Standard Error of Cut Scores for NAEP—Writing Grade 8, Round 1*

| Table/ | Achievement | Median | Median | Standard Error | |  | Scaled Score | | Percent |
|---|---|---|---|---|---|---|---|---|---|
| Group | Level | Theta | NAEP | EmpSE | BootSE | MAD | Min | Max | of |
| Table 1 | Advanced | 1.67 | 209 | 8.06 | 7.11 | 8 | 209 | 300 | 3.66 |
| | Proficient | 0.4 | 165 | 10.31 | 8.4 | 7.5 | 165 | 208 | 31.91 |
| | Basic | -1.1 | 111 | 8.83 | 9.14 | 7.5 | 111 | 164 | 50.8 |
| | Below Basic | | | | | | 0 | 110 | 13.63 |
| Table 2 | Advanced | 2.08 | 224 | 5.56 | 5.25 | 4 | 224 | 300 | 0.91 |
| | Proficient | 0.92 | 182 | 7.2 | 6.81 | 10 | 182 | 223 | 17.12 |
| | Basic | -0.4 | 136 | 8.97 | 9.34 | 6 | 136 | 181 | 48.47 |
| | Below Basic | | | | | | 0 | 135 | 33.5 |
| Table 3 | Advanced | 1.96 | 219 | 7.74 | 8.66 | 2.5 | 219 | 300 | 1.45 |
| | Proficient | 0.48 | 167 | 7.74 | 8.1 | 6.5 | 167 | 218 | 31.43 |
| | Basic | -0.88 | 119 | 9.44 | 8.63 | 11 | 119 | 166 | 48.56 |
| | Below Basic | | | | | | 0 | 118 | 18.55 |
| Table 4 | Advanced | 1.86 | 216 | 6.74 | 6.96 | 4 | 216 | 300 | 2.01 |
| | Proficient | 0.42 | 165 | 5.95 | 5.85 | 9 | 165 | 215 | 33.09 |
| | Basic | -0.7 | 125 | 8.05 | 7.63 | 6 | 125 | 164 | 40.81 |
| | Below Basic | | | | | | 0 | 124 | 24.09 |
| Table 5 | Advanced | 2.02 | 221 | 17.66 | 16.95 | 25 | 221 | 300 | 1.19 |
| | Proficient | 0.6 | 171 | 11.87 | 12.17 | 13 | 171 | 220 | 27.12 |
| | Basic | -1.16 | 109 | 24.49 | 23.16 | 33 | 109 | 170 | 59.18 |
| | Below Basic | | | | | | 0 | 108 | 12.52 |
| Table 6 | Advanced | 1.84 | 215 | 7.62 | 7.41 | 8 | 215 | 300 | 2.19 |
| | Proficient | 0.88 | 181 | 10.9 | 10.82 | 13 | 181 | 214 | 16.8 |
| | Basic | -0.77 | 123 | 12.88 | 11.55 | 12.5 | 123 | 180 | 59.26 |
| | Below Basic | | | | | | 0 | 122 | 21.75 |
| Group A | Advanced | 2.01 | 221 | 4 | 4.28 | 7 | 221 | 300 | 1.24 |
| | Proficient | 0.61 | 172 | 5.11 | 4.93 | 10 | 172 | 220 | 26.61 |
| | Basic | -0.85 | 120 | 6.59 | 6.04 | 12 | 120 | 171 | 52.51 |
| | Below Basic | | | | | | 0 | 119 | 19.64 |
| Group B | Advanced | 1.86 | 216 | 5.12 | 5.3 | 14 | 216 | 300 | 2.01 |
| | Proficient | 0.6 | 171 | 5.21 | 5.05 | 12 | 171 | 215 | 26.72 |
| | Basic | -0.79 | 122 | 7.94 | 7.95 | 13 | 122 | 170 | 49.86 |
| | Below Basic | | | | | | 0 | 121 | 21.41 |
| All | Advanced | 1.87 | 216 | 6.62 | 2.87 | 12 | 216 | 300 | 1.89 |
| | Proficient | 0.61 | 171 | 5.76 | 3.29 | 11 | 171 | 215 | 26.41 |
| | Basic | -0.85 | 120 | 6.36 | 5.39 | 12 | 120 | 170 | 52.06 |
| | Below Basic | | | | | | 0 | 119 | 19.64 |

*Table 50: Estimates of Standard Error of Cut Scores for NAEP—Writing Grade 8, Round 2*

| Table/ | Achievement | Median | Median | Standard Error | | | Scaled Score | | Percent |
|---|---|---|---|---|---|---|---|---|---|
| Group | Level | Theta | NAEP | EmpSE | BootSE | MAD | Min | Max | of |
| Table 1 | Advanced | 1.97 | 220 | 21.67 | 16.67 | 10 | 220 | 300 | 1.39 |
| | Proficient | 0.4 | 164 | 6.53 | 5.52 | 5.5 | 164 | 219 | 34.53 |
| | Basic | -0.85 | 120 | 20.47 | 22.89 | 6.5 | 120 | 163 | 44.45 |
| | Below Basic | | | | | | 0 | 119 | 19.64 |
| Table 2 | Advanced | 2.03 | 222 | 4.39 | 4.75 | 4 | 222 | 300 | 1.15 |
| | Proficient | 0.77 | 177 | 8.72 | 7.51 | 8 | 177 | 221 | 21.53 |
| | Basic | -0.6 | 129 | 4.61 | 4.08 | 6 | 129 | 176 | 50.13 |
| | Below Basic | | | | | | 0 | 128 | 27.19 |
| Table 3 | Advanced | 1.96 | 219 | 4.22 | 4.57 | 2.5 | 219 | 300 | 1.45 |
| | Proficient | 0.56 | 170 | 4.37 | 4.39 | 5 | 170 | 218 | 28.75 |
| | Basic | -1 | 115 | 3.27 | 3.27 | 3.5 | 115 | 169 | 53.77 |
| | Below Basic | | | | | | 0 | 114 | 16.02 |
| Table 4 | Advanced | 1.86 | 216 | 5.55 | 5.37 | 4 | 216 | 300 | 2.01 |
| | Proficient | 0.76 | 177 | 5.04 | 5.39 | 3 | 177 | 215 | 20.67 |
| | Basic | -1.24 | 106 | 8.21 | 7.3 | 3 | 106 | 176 | 66.3 |
| | Below Basic | | | | | | 0 | 105 | 11.01 |
| Table 5 | Advanced | 2.02 | 221 | 11.07 | 9.96 | 11 | 221 | 300 | 1.19 |
| | Proficient | 0.71 | 175 | 9.66 | 10.17 | 11 | 175 | 220 | 23.32 |
| | Basic | -0.85 | 120 | 14.19 | 14.08 | 18 | 120 | 174 | 56.1 |
| | Below Basic | | | | | | 0 | 119 | 19.4 |
| Table 6 | Advanced | 1.84 | 215 | 3.04 | 3.05 | 3 | 215 | 300 | 2.1 |
| | Proficient | 0.75 | 177 | 6.43 | 6.07 | 7.5 | 177 | 214 | 20.88 |
| | Basic | -0.48 | 133 | 12.76 | 12.72 | 15.5 | 133 | 176 | 46.26 |
| | Below Basic | | | | | | 0 | 132 | 30.76 |
| Group A | Advanced | 2.01 | 221 | 2.81 | 3.03 | 5 | 221 | 300 | 1.24 |
| | Proficient | 0.6 | 171 | 4.02 | 3.96 | 6 | 171 | 220 | 27.06 |
| | Basic | -0.83 | 121 | 3.19 | 2.99 | 8 | 121 | 170 | 51.49 |
| | Below Basic | | | | | | 0 | 120 | 20.21 |
| Group B | Advanced | 1.88 | 217 | 2.38 | 2.28 | 4 | 217 | 300 | 1.84 |
| | Proficient | 0.73 | 176 | 3.84 | 3.9 | 7.5 | 176 | 216 | 21.8 |
| | Basic | -0.87 | 120 | 6.96 | 6.79 | 14 | 120 | 175 | 57.19 |
| | Below Basic | | | | | | 0 | 119 | 19.17 |
| All | Advanced | 1.98 | 220 | 6.35 | 2.06 | 5 | 220 | 300 | 1.34 |
| | Proficient | 0.68 | 174 | 5.72 | 3.21 | 8 | 174 | 219 | 24.26 |
| | Basic | -0.85 | 120 | 4.34 | 2.75 | 9 | 120 | 173 | 55 |
| | Below Basic | | | | | | 0 | 119 | 19.4 |

*Table 51: Estimates of Standard Error of Cut Scores for NAEP—Writing Grade 8, Round 3*

| Table/ | Achievement | Median | Median | Standard Error | | | Scaled Score | | Percent |
|---|---|---|---|---|---|---|---|---|---|
| **Group** | **Level** | **Theta** | **NAEP** | **EmpSE** | **BootSE** | **MAD** | **Min** | **Max** | **Of** |
| Table 1 | Advanced | 1.85 | 216 | 7.57 | 7.71 | 7.5 | 216 | 300 | 2.01 |
| | Proficient | 0.65 | 173 | 12.33 | 12.29 | 12.5 | 173 | 215 | 24.75 |
| | Basic | -1.05 | 113 | 2.2 | 2.26 | 2 | 113 | 172 | 58.3 |
| | Below Basic | | | | | | 0 | 112 | 14.93 |
| Table 2 | Advanced | 1.96 | 219 | 10.83 | 10.71 | 17 | 219 | 300 | 1.43 |
| | Proficient | 0.42 | 165 | 8.28 | 6.53 | 6 | 165 | 218 | 33.31 |
| | Basic | -0.58 | 129 | 7.92 | 8.46 | 4 | 129 | 164 | 37.79 |
| | Below Basic | | | | | | 0 | 128 | 27.47 |
| Table 3 | Advanced | 1.4 | 200 | 5.44 | 4.13 | 2 | 200 | 300 | 7.06 |
| | Proficient | 0.81 | 179 | 5.98 | 6.83 | 0.5 | 179 | 199 | 14.34 |
| | Basic | -1.09 | 112 | 5.3 | 5.14 | 6.5 | 112 | 178 | 64.79 |
| | Below Basic | | | | | | 0 | 111 | 13.81 |
| Table 4 | Advanced | 1.85 | 216 | 4.44 | 4.42 | 4 | 216 | 300 | 2.01 |
| | Proficient | 0.65 | 173 | 4.51 | 4.25 | 6 | 173 | 215 | 24.36 |
| | Basic | -0.8 | 122 | 10.33 | 11.07 | 3 | 122 | 172 | 52.81 |
| | Below Basic | | | | | | 0 | 121 | 20.83 |
| Table 5 | Advanced | 1.71 | 210 | 8.32 | 7.65 | 6 | 210 | 300 | 3.2 |
| | Proficient | 0.59 | 171 | 24.55 | 23.89 | 36 | 171 | 209 | 25.87 |
| | Basic | -0.76 | 123 | 14.33 | 13.97 | 22 | 123 | 170 | 48.6 |
| | Below Basic | | | | | | 0 | 122 | 22.32 |
| Table 6 | Advanced | 1.74 | 212 | 4.09 | 3.97 | 4.5 | 212 | 300 | 2.91 |
| | Proficient | 0.65 | 173 | 5.88 | 6.21 | 4.5 | 173 | 211 | 23.46 |
| | Basic | -0.65 | 128 | 5.16 | 4.59 | 5.5 | 128 | 172 | 48.04 |
| | Below Basic | | | | | | 0 | 127 | 25.59 |
| Group A | Advanced | 1.74 | 211 | 8.38 | 8.33 | 11 | 211 | 300 | 3.01 |
| | Proficient | 0.65 | 173 | 5.33 | 5.4 | 8 | 173 | 210 | 23.76 |
| | Basic | -1 | 115 | 3.18 | 2.96 | 7 | 115 | 172 | 57.45 |
| | Below Basic | | | | | | 0 | 114 | 15.79 |
| Group B | Advanced | 1.74 | 212 | 3.36 | 3.18 | 7 | 212 | 300 | 2.91 |
| | Proficient | 0.62 | 172 | 3.69 | 3.7 | 7.5 | 172 | 211 | 24.94 |
| | Basic | -0.76 | 123 | 3.49 | 3.43 | 6.5 | 123 | 171 | 49.83 |
| | Below Basic | | | | | | 0 | 122 | 22.32 |
| All | Advanced | 1.74 | 211 | 6.84 | 3.58 | 9 | 211 | 300 | 3.01 |
| | Proficient | 0.65 | 173 | 5.71 | 3.18 | 7 | 173 | 210 | 23.76 |
| | Basic | -0.84 | 120 | 4.53 | 3.1 | 9 | 120 | 172 | 53.6 |
| | Below Basic | | | | | | 0 | 119 | 19.64 |

*Table 52: Estimates of Standard Error of Cut Scores for NAEP—Writing Grade 12, Round 1*

| Table/ | Achievement | Median | Median | Standard Error | | | Scaled Score | | Percent |
|---|---|---|---|---|---|---|---|---|---|
| Group | Level | Theta | NAEP | EmpSE | BootSE | MAD | Min | Max | of |
| Table 1 | Advanced | 1.99 | 220 | 1.57 | 1.66 | 1 | 220 | 300 | 1.17 |
| | Proficient | 0.98 | 185 | 6.8 | 7.95 | 3 | 185 | 219 | 15.15 |
| | Basic | -0.73 | 124 | 9.93 | 9.74 | 5 | 124 | 184 | 60.75 |
| | Below Basic | | | | | | 0 | 123 | 22.93 |
| Table 2 | Advanced | 2.05 | 222 | 11.69 | 12.03 | 16 | 222 | 300 | 0.96 |
| | Proficient | 0.58 | 171 | 9.58 | 10.11 | 2 | 171 | 221 | 28.07 |
| | Basic | -0.97 | 116 | 17.38 | 15.84 | 10 | 116 | 170 | 54.18 |
| | Below Basic | | | | | | 0 | 115 | 16.78 |
| Table 3 | Advanced | 1.73 | 211 | 5.4 | 5.19 | 6 | 211 | 300 | 2.94 |
| | Proficient | 0.59 | 171 | 10.11 | 9.77 | 13 | 171 | 210 | 26.09 |
| | Basic | -0.92 | 118 | 4.91 | 5.5 | 1.5 | 118 | 170 | 53 |
| | Below Basic | | | | | | 0 | 117 | 17.97 |
| Table 4 | Advanced | 1.83 | 215 | 3.45 | 3.38 | 4 | 215 | 300 | 2.12 |
| | Proficient | 0.36 | 163 | 11.03 | 8.78 | 7 | 163 | 214 | 35.77 |
| | Basic | -0.53 | 132 | 6.85 | 7.4 | 4 | 132 | 162 | 33.12 |
| | Below Basic | | | | | | 0 | 131 | 28.99 |
| Table 5 | Advanced | 1.74 | 212 | 4.57 | 4.4 | 6 | 212 | 300 | 2.72 |
| | Proficient | 0.49 | 167 | 10.36 | 10.41 | 9 | 167 | 211 | 29.61 |
| | Basic | -0.83 | 121 | 7.89 | 8.73 | 9 | 121 | 166 | 47.16 |
| | Below Basic | | | | | | 0 | 120 | 20.51 |
| Table 6 | Advanced | 1.47 | 202 | 8.92 | 8.95 | 10 | 202 | 300 | 5.76 |
| | Proficient | 0.31 | 161 | 10.91 | 10.84 | 10 | 161 | 201 | 34.11 |
| | Basic | -0.89 | 119 | 8.89 | 9.17 | 5 | 119 | 160 | 41.47 |
| | Below Basic | | | | | | 0 | 118 | 18.66 |
| Group A | Advanced | 1.97 | 220 | 3.5 | 3.69 | 4 | 220 | 300 | 1.32 |
| | Proficient | 0.73 | 176 | 5.53 | 5.51 | 10.5 | 176 | 219 | 22.48 |
| | Basic | -0.89 | 119 | 4.69 | 4.41 | 10 | 119 | 175 | 57.25 |
| | Below Basic | | | | | | 0 | 118 | 18.94 |
| Group B | Advanced | 1.73 | 212 | 3.12 | 3.22 | 5.5 | 212 | 300 | 2.94 |
| | Proficient | 0.36 | 163 | 5.25 | 5.08 | 12 | 163 | 211 | 34.6 |
| | Basic | -0.82 | 121 | 4.06 | 3.96 | 9 | 121 | 162 | 41.73 |
| | Below Basic | | | | | | 0 | 120 | 20.73 |
| All | Advanced | 1.8 | 214 | 6.58 | 2.47 | 7.5 | 214 | 300 | 2.31 |
| | Proficient | 0.56 | 170 | 6.47 | 4.41 | 14 | 170 | 213 | 27.5 |
| | Basic | -0.85 | 120 | 4.34 | 2.71 | 10 | 120 | 169 | 50.45 |
| | Below Basic | | | | | | 0 | 119 | 19.74 |

*Table 53: Estimates of Standard Error of Cut Scores for NAEP—Writing Grade 12, Round 2*

| Table/ | Achievement | Median | Median | Standard Error | | | Scaled Score | | Percent |
|---|---|---|---|---|---|---|---|---|---|
| Group | Level | Theta | NAEP | EmpSE | BootSE | MAD | Min | Max | Of |
| Table 1 | Advanced | 1.99 | 220 | 3.41 | 4.05 | 1 | 220 | 300 | 1.17 |
| | Proficient | 0.93 | 183 | 6.93 | 7.84 | 5 | 183 | 219 | 16.61 |
| | Basic | -0.73 | 124 | 9.9 | 8.55 | 2 | 124 | 182 | 59.29 |
| | Below Basic | | | | | | 0 | 123 | 22.93 |
| Table 2 | Advanced | 1.85 | 215 | 5.91 | 5.85 | 7 | 215 | 300 | 1.99 |
| | Proficient | 0.54 | 169 | 5.79 | 6.4 | 6 | 169 | 214 | 28.55 |
| | Basic | -1.1 | 111 | 9.48 | 7.86 | 4 | 111 | 168 | 55.39 |
| | Below Basic | | | | | | 0 | 110 | 14.06 |
| Table 3 | Advanced | 1.7 | 211 | 6.57 | 6.32 | 8.5 | 211 | 300 | 3.13 |
| | Proficient | 0.72 | 176 | 4.51 | 4.41 | 5.5 | 176 | 210 | 20.98 |
| | Basic | -0.92 | 118 | 5.01 | 5.22 | 4 | 118 | 175 | 57.92 |
| | Below Basic | | | | | | 0 | 117 | 17.97 |
| Table 4 | Advanced | 1.72 | 211 | 6.43 | 5.39 | 5.5 | 211 | 300 | 3.04 |
| | Proficient | 0.31 | 161 | 6.37 | 6.05 | 8 | 161 | 210 | 36.48 |
| | Basic | -0.73 | 124 | 0.56 | 0.65 | 0 | 124 | 160 | 37.8 |
| | Below Basic | | | | | | 0 | 123 | 22.67 |
| Table 5 | Advanced | 1.82 | 214 | 3.52 | 2.92 | 3 | 214 | 300 | 2.17 |
| | Proficient | 0.48 | 167 | 4.4 | 4.9 | 3 | 167 | 213 | 30.96 |
| | Basic | -0.97 | 116 | 3.37 | 3.26 | 4 | 116 | 166 | 50.08 |
| | Below Basic | | | | | | 0 | 115 | 16.78 |
| Table 6 | Advanced | 1.51 | 204 | 2.53 | 2.83 | 0 | 204 | 300 | 5.1 |
| | Proficient | 0.48 | 167 | 5.23 | 5.68 | 3 | 167 | 203 | 27.62 |
| | Basic | -0.79 | 122 | 8.45 | 9.84 | 5 | 122 | 166 | 45.86 |
| | Below Basic | | | | | | 0 | 121 | 21.42 |
| Group A | Advanced | 1.92 | 218 | 4.01 | 4.21 | 4 | 218 | 300 | 1.51 |
| | Proficient | 0.68 | 174 | 3.83 | 3.74 | 8.5 | 174 | 217 | 24.2 |
| | Basic | -0.8 | 122 | 3.56 | 3.55 | 8.5 | 122 | 173 | 53.34 |
| | Below Basic | | | | | | 0 | 121 | 20.95 |
| Group B | Advanced | 1.65 | 209 | 3.18 | 3.08 | 5.5 | 209 | 300 | 3.61 |
| | Proficient | 0.48 | 167 | 2.93 | 3.13 | 4.5 | 167 | 208 | 29.12 |
| | Basic | -0.8 | 122 | 3.03 | 3.15 | 3.5 | 122 | 166 | 46.12 |
| | Below Basic | | | | | | 0 | 121 | 21.16 |
| All | Advanced | 1.77 | 213 | 6.9 | 3.39 | 8.5 | 213 | 300 | 2.59 |
| | Proficient | 0.49 | 167 | 5.14 | 1.81 | 7.5 | 167 | 212 | 30.14 |
| | Basic | -0.8 | 122 | 4.19 | 2.44 | 5.5 | 122 | 166 | 46.33 |
| | Below Basic | | | | | | 0 | 121 | 20.95 |

*Table 54: Estimates of Standard Error of Cut Scores for NAEP—Writing Grade 12, Round 3*

| Table/ | Achievement | Median | Median | Standard Error | | | Scaled Score | | Percent |
|---|---|---|---|---|---|---|---|---|---|
| Group | Level | Theta | NAEP | EmpSE | BootSE | MAD | Min | Max | Of |
| Table 1 | Advanced | 1.69 | 210 | 5.19 | 5.93 | 4 | 210 | 300 | 3.18 |
| | Proficient | 0.71 | 175 | 3 | 3.64 | 0 | 175 | 209 | 21.55 |
| | Basic | -0.73 | 124 | 5.73 | 5.86 | 5 | 124 | 174 | 52.34 |
| | Below Basic | | | | | | 0 | 123 | 22.93 |
| Table 2 | Advanced | 1.6 | 206 | 10.52 | 9.62 | 13 | 206 | 300 | 4.19 |
| | Proficient | 0.41 | 165 | 10.17 | 8.04 | 5 | 165 | 205 | 31.29 |
| | Basic | -0.8 | 122 | 6.06 | 6.6 | 0 | 122 | 164 | 43.57 |
| | Below Basic | | | | | | 0 | 121 | 20.95 |
| Table 3 | Advanced | 1.6 | 207 | 7.17 | 6.24 | 7 | 207 | 300 | 4.06 |
| | Proficient | 0.57 | 170 | 8.3 | 6.87 | 6 | 170 | 206 | 25.75 |
| | Basic | -0.9 | 118 | 7.63 | 7.72 | 8.5 | 118 | 169 | 51.78 |
| | Below Basic | | | | | | 0 | 117 | 18.41 |
| Table 4 | Advanced | 1.68 | 209 | 1.5 | 1.13 | 0.5 | 209 | 300 | 3.24 |
| | Proficient | 0.69 | 175 | 3.22 | 3.25 | 3.5 | 175 | 208 | 22.13 |
| | Basic | -0.75 | 124 | 4.36 | 3.34 | 1.5 | 124 | 174 | 52.41 |
| | Below Basic | | | | | | 0 | 123 | 22.22 |
| Table 5 | Advanced | 1.71 | 210 | 2.16 | 1.93 | 2 | 210 | 300 | 3.13 |
| | Proficient | 0.47 | 167 | 6.88 | 6.95 | 10 | 167 | 209 | 30.01 |
| | Basic | -0.93 | 117 | 9.84 | 10.96 | 3 | 117 | 166 | 49.14 |
| | Below Basic | | | | | | 0 | 116 | 17.73 |
| Table 6 | Advanced | 1.71 | 210 | 5.42 | 5.82 | 2 | 210 | 300 | 3.13 |
| | Proficient | 0.49 | 167 | 11.98 | 12.47 | 15 | 167 | 209 | 29.2 |
| | Basic | -0.93 | 117 | 13.65 | 13.05 | 21 | 117 | 166 | 49.94 |
| | Below Basic | | | | | | 0 | 116 | 17.73 |
| Group A | Advanced | 1.67 | 209 | 3.94 | 3.99 | 8.5 | 209 | 300 | 3.33 |
| | Proficient | 0.69 | 175 | 4.21 | 4.39 | 9.5 | 175 | 208 | 22.04 |
| | Basic | -0.8 | 122 | 4.07 | 4.24 | 6.5 | 122 | 174 | 53.68 |
| | Below Basic | | | | | | 0 | 121 | 20.95 |
| Group B | Advanced | 1.69 | 210 | 0.88 | 0.89 | 2 | 210 | 300 | 3.18 |
| | Proficient | 0.56 | 170 | 4.29 | 4.17 | 8 | 170 | 209 | 27.06 |
| | Basic | -0.86 | 120 | 4.3 | 3.94 | 6.5 | 120 | 169 | 50.34 |
| | Below Basic | | | | | | 0 | 119 | 19.43 |
| All | Advanced | 1.69 | 210 | 6.08 | 1.06 | 4 | 210 | 300 | 3.24 |
| | Proficient | 0.64 | 173 | 5.99 | 3.52 | 7 | 173 | 209 | 23.59 |
| | Basic | -0.8 | 122 | 4.35 | 2.74 | 7 | 122 | 172 | 52.22 |
| | Below Basic | | | | | | 210 | 300 | 3.18 |

### 6.6.3 Summary

In this section, procedural and internal validity evidence of the ALS results are discussed. Evidence for procedural validity was drawn from the evaluation surveys given to respondents after each round. Internal validity was discussed in terms of a decrease in the variability of cut scores over rounds. Based on the accumulation of multiple types of validity evidence, the information reviewed in this section supports the validity of the ALS.

# Chapter 7—Recommendations to the Governing Board

Four sets of recommendations are included in this report. The first set of recommendations refers to the main deliverable for this contract. The rest are recommendations for future standard-setting efforts per the Governing Board's request.

## 7.1 Achievement Levels

With the support of the Technical Advisory Committee on Standard Setting (TACSS), achievement levels were recommended to the Governing Board for reporting results of the 2011 NAEP writing in grades 8 and 12. The recommended achievement levels had three parts: (a) the achievement levels descriptions (ALDs), (b) cut scores, and (c) exemplar bodies of work (BoWs). The recommended ALDs are presented in this report as Figures 2–4. The recommended cut scores are those resulting from Round 3 of the operational achievement levels–setting (ALS) meeting, as presented in Table 34. The modified set of exemplar BoWs, at the request of the Committee on Standards, Design and Methodology (COSDAM), were recommended as the third part of the achievement levels. The packet submitted to the Governing Board when the achievement levels were recommended during the May 2012 quarterly meeting is included as Appendix L.

## 7.2 Recruiting Procedures

Panelist recruitment was a challenging task. Recruitment may be improved if the Governing Board adopts the following recommendations.

First, allow the use of multiple panelists from a single nominator, provided the qualifications and NAEP diversity criteria are met. This will allow recruitment contractors to partner with nomination-rich nominators, particularly those with nationwide reach, without changing the panel-defining criteria established by the Governing Board.

Second, allow use of at-large nominations, provided the qualifications and NAEP diversity criteria are met. Acceptance of this recommendation will allow recruitment contractors to make use of nomination-rich resources even though no "partner" nominator is defined, and will improve the contractors' ability to reach general public nominees directly or at the recommendation of other general public nominees who are acquainted with those in their field.

## 7.3 Achievement Levels–Setting Procedures

Two procedural recommendations are being made. One is a global recommendation, and the other is smaller in scope.

### *7.3.1 Computerization of Standard Setting*

The Governing Board, in its latest standard-setting efforts, has again raised the bar for standard setting by computerizing standard-setting methods. Through the use of the Body of Work Technological Integration and Enhancements (BoWTIE) software in this current project and Computer-Aided Bookmarking (CAB) in its academic preparedness research (Measured Progress & WestEd, 2012), the Governing Board has shown that use of technology in standard setting brings efficiency and effectiveness to the process. Further, technology helps bring to light some issues previously hidden by logistical difficulties. The Governing Board has substantially advanced the field of educational assessment in the area of standard setting and it should continue to do so. It is thus recommended that future standard settings use similar technology-assisted approaches.

### *7.3.2 Computation of Overall Cut Scores in Body of Work Standard Setting*

The computation of the overall cut scores for the current project might not have been the most technically sound. Computing individual cut scores and then determining the median is not consistent with the data structure. This issue is discussed in the Technical Report. Measured Progress now recommends use of a model that is more appropriate for  computing cut scores (e.g., generalized linear mixed model) for the overall cut scores, while still using the traditional logistic regression for the purpose of providing variance components used for calculating consistency. Analysis performed after the operational study revealed that overall cut scores computed using the generalized linear mixed models are not materially different from the official operational cut scores.

## 7.4 Validity Evidence

This last set of recommendations is for research studies that might be included in the collection of validity evidence of the cut scores. Data are currently available for these studies. With the cut scores having been set using the BoW method, answers to the two questions below can provide information regarding internal evidence of validity.

- Did the panelists classify the BoWs holistically based on the two responses? Or were their classifications dominated by one of the two responses?
- Do panelists' computed cut scores correspond to their conceptual cut scores?

The first question addresses the issue of correspondence between the panelists' classification models and the scaling model, whereas the second question addresses the issue of correspondence between the panelists' classification model and the logistic regression model.

Findings from the studies will provide “lessons learned” for BoW standard setting. Information from each investigation will be useful in determining enhancements to the BoW method in terms of additional rigor in training and instructions that could only lead in better implementations that may provide more evidence in support of the validity of the cut scores. Both of the research questions above are important to standard setting and each is especially important for an assessment such as NAEP writing with only two scorable performances.

### *7.4.1 Body of Work Classification*

The first research question is related to the classifications that panelists provided in the BoW method. There are two underlying assumptions in the classifications: (1) the classifications are based on responses to two tasks; and (2) the classification decision model is compensatory. Investigating the relationship between classifications and the scores of the two tasks will reveal whether there is evidence in support of the two assumptions or if it appears that one or both assumptions were violated. Results of a study on the 1998 NAEP writing ALS provided some indication that panelists use the compensatory decision-making model when classifying student performances (Bay, 2000).

The analysis will involve examining the relationships between the achievement levels classifications and the scores on the two responses. In a way, this analysis will determine whether the classifications were “dominated” by one of the responses, where “dominant” response can be any of the following:

- Response with higher score
- Response with lower score
- First response
- Second response

- Longer response
- Shorter response

This investigation will reveal whether the panelists' judgment model match the compensatory nature of the NAEP scaling model. A match between the models will provide evidence of internal validity.

#### *7.4.2 Relationship Between Conceptual and Computational Cut Scores*

Conceptually, a panelist's cut score represents the point of greatest ambivalence for the panelist in classifying BoWs. This may be the score of an actual BoW, or it may be the score between the two most similar and least differentiable BoWs –those at the borderline of adjacent achievement levels. For the second and third rounds of classifications in the ALS process, panelists were asked to indicate their level of confidence on each BoW classification. If panelists were making classification decisions based on their understanding of the achievement levels descriptions and the knowledge, skills, and abilities demonstrated in each BoW as they relate to the descriptions, the expectation is that their conceptual cut score for each level is where their confidence in their classification dips to the lowest level. In standard setting there is an implicit assumption that panelists' conceptual cut score is represented by the computed cut scores. In this case, the investigation will reveal whether the panelists' judgment model matches the logistic regression model. A match between the models will provide internal validity evidence of resulting cut scores.

# References

ACT. (2007). *Developing achievement levels on the 2006 National Assessment of Educational Progress in grade twelve economics: Process report*. Iowa City, IA: ACT.

Bay, L. G. & Loomis, S. C. (1995). *Validation of the 1994 NAEP achievement levels: booklet classification study*. National Conference on Large Scale Assessment, Phoenix, AZ.

Bay, L. (2000). *Setting Achievement levels on the 1998 National Assessment of Educational Progress in writing: Performance profiles study*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans.

Bay, L. (2012). *Developing achievement levels on the 2011 National Assessment of Educational Progress in grades 8 and 12 writing: Technical report*. Dover, NH: Measured Progress.

Berk, R.A. (1986). *A consumer's guide to setting performance standards on criterion-referenced tests*. Review of Educational Research, 56, 137-172.

Brennan, R. L. (2002, October). *Estimated standard error of a mean when there are only two observations*. (CASMA Tech. Note No. 1). Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement and Assessment.

Chen, W., Loomis, S.C., & Fisher, T. (2000). *Developing achievement levels on the 1998 NAEP civics and writing: Technical report*. Iowa City, IA: ACT, Inc.

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), pp. 36–48.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. Brennan (Ed.), *Educational measurement*, 4th ed. (pp. 433–470). Washington, DC: American Council on Education and Westport, CT: Praeger Publishers.

Hambleton, R. K., Pitoniak, M. J., & Coppella, J. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability and validity of results In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 47-106). New York, NY: Routledge.

Hanson, B. A. & Bay, L. G. (April, 1999). Classifying student performance as a method for setting achievement levels for NAEP Writing. Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada.

Hanson, B. A., Bay, L. G., & Loomis, S. C. (April, 1998). Booklet classification study. Paper presented at the annual meeting of the National Council on Measurement in Education. San Diego, CA.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification Consistency indexes estimated under alternative strong true score models. Journal of Educational Measurement, 27, 345–359.

Huynh, H. (1976). On the reliability of domain-referenced testing. *Journal of Educational Measurement*, 13, 253–264.

Kane, M. T., & Bay, L. G. (1996). *The intersection of score interpretation, assessment design, and standard-setting methodology on NAEP*. National Council on Measurement in Education, New York, NY.

Kingston, N. M., Kahl, S. R., Sweeney, K. P., & Bay, L. (2001). Setting performance standards using the body of work method. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219–248). Mahwah, NJ: 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.

Loomis, S. C., & Hanick, P. L. (2000). *Developing achievement levels for the 1998 NAEP in writing: Final report*. Iowa City, IA: ACT.

Maritz, J. S., & Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, *73*(361), 194–196.

Martineau, J. A. (2007). An expansion and practical evaluation of expected classification accuracy. Applied Psychological Measurement.

Measured Progress. (2011). *Developing Achievement Levels on the National Assessment of Educational Progress for Writing Grades 8 and 12 in 2011: Design Document*. Dover, NH: Author.

Measured Progress & WestEd. (2012). *National Assessment of Educational Progress Judgmental Standard Setting (JSS): Technical report*. Dover, NH: Authors.

National Assessment Governing Board. (2010). Work statement for developing achievement levels on the National Assessment of Educational Progress for developing achievement levels for writing grades 8 and 12 in 2011 and grade 4 in 2013. Washington, DC: Author.

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

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to Perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159–174). Mahwah, NJ: Lawrence Erlbaum.

Reckase, M.D. (2012). The role, format, and impact of feedback to standard setting panelists. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 149-164). New York, NY: Routledge.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. Practical Assessment, Research & Evaluation, 7(14).

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment,* Research & Evaluation, 10(13).

Skorupski, W. (2012). Understanding the cognitive processes for standard setting panelists. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 135-148). New York, NY: Routledge.

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, *13*, 265–276.

U.S. Census Bureau. (2011, January). *Poverty*. Retrieved January 14, 2011 from the U.S. Census Bureau website: www.census.gov/hhes/www/poverty/about/overview/index.html

U.S. Department of Education, National Center for Education Statistics, & Institute of Education Sciences. (2010a, August). *Common Core of Data*. Retrieved February 4, 2011 from the National Center for Education Statistics website: http://nces.ed.gov/ccd/ccddata.asp

U.S. Department of Education, National Center for Education Statistics, & Institute of Education Sciences. (2010b, August). *Local Education Agency (School District) Universe Survey Data*. Retrieved February 4, 2011 from the National Center for Education Statistics website: http://nces.ed.gov/ccd/pubagency.asp

U.S. Department of Education, National Center for Education Statistics, & Institute of Education Sciences. (2010c, August). *Private School Universe Survey Data*. Retrieved February 4, 2011 from the National Center for Education Statistics website: http://nces.ed.gov/surveys/pss/pssdata.asp

U.S. Department of Education, National Center for Education Statistics, & Institute of Education Sciences. (2010d, August) *Public Elementary/Secondary School Universe Survey Data*. Retrieved February 4, 2011 from the National Center for Education Statistics website: http://nces.ed.gov/ccd/pubschuniv.asp

U.S. Department of Education, National Center for Education Statistics, & Institute of Education Sciences. (2011, November 9). *Interpreting Reading Results: NAEP Reporting Groups*. Retrieved 22 December, 2010 from the National Assessment of Educational Progress website: http://nces.ed.gov/nationsreportcard/reading/interpret-results.asp#repgroups

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational Measurement*, *37*, 141–162.

# Appendices