National Assessment Governing Board

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

Final Submitted: September 5, 2012

Technical Report

Submitted to:

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

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

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

measured progress

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

---

Luz Bay

with

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
Eight-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, D.C.

Table of Contents

List of Tables

List of Figures

Chapter 1—INTRODUCTION

In September 2010, the National Assessment Governing Board (hereafter referred to as the Governing Board) contracted with Measured Progress to conduct research and other activities for setting achievement levels on the 2011 National Assessment of Educational Progress (NAEP) in grade 8 and grade 12 writing. The contract called for a series of reports, including a Technical Report documenting the technical aspects of Measured Progress's contract activities. This Technical Report provides information on the materials and computational procedures used for the Achievement Levels–Setting (ALS) meetings held for this project. Table 1-1 lists all the ALS meetings, including the panel meetings that led up to the operational ALS meeting in February 2012. The primary purpose of each meeting is described in the table.

Table 1-1. Achievement Levels–Setting (ALS) Meetings

Meeting	*Primary Purpose*	*Date*	*Venue*
Field Trial 1	To test logistics involved in using two laptop computers	September 22–23, 2011	Portsmouth, NH
Pilot Study	To implement process designed for operational meeting and evaluate need for change	November 15–18, 2011	St. Louis, MO
Special Study 1	To compare performance on the 2007 and 2011 assessments	November 18–19, 2012	St. Louis, MO
Field Trial 2	To test implementation of modifications based on pilot study findings	January 27, 2012	Dover, NH
Operational ALS Meeting	To implement achievement levels-setting procedures to develop recommendations 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

Field Trial 1 was used to test the logistics of implementing the process. The Pilot Study was used to test the process. Special Study 1 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 descriptors (ALDs) were modified. The modifications to the ALDs 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 focused on a comparison of the 2007 and 2011 NEAP writing assessment results. All of these activities and results are described in detail in *Developing Achievement Levels on the 2011 National Assessment of Educational Progress in Grades 8 and 12 Writing: Process Report* (Bay, 2012; hereafter referred to as the Process Report).

The Body of Work (BoW) methodology used to set the achievement levels for the 2011 NAEP in grades 8 and 12 writing. The BoW method belongs to the holistic family of standard-setting methods in which the panelist rating task consists of reviewing a series of examinee work samples and assigning each sample to one of several performance categories (Hambleton & Pitoniak, 2006). The BoW method (Kingston, Kahl, Sweeney, & Bay, 2001) is the method 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). Traditionally, the BoW method is implemented in two stages. During the rangefinding stage, panelists are provided a set of bodies of work (BoWs) with scores that span the whole range of performance. Scores of the BoWs were not revealed to the panelists so as not to bias their classifications one way or another (S. Kahl, personal communication, August 10, 2012). Panelists classify those BoWs into achievement levels categories. Their classifications are used to compute the cut scores. During the pinpointing stage, panelists are provided a set of BoWs that have scores in the vicinity of the cut scores determined during the rangefinding stage. For each cut score, panelists classify each BoW as below or at or above the achievement level.

For setting cut scores on the 2011 NAEP writing for grades 8 and 12, the BoW method was enhanced by using computer software called Body of Work Technological Integration and Enhancements (BoWTIE), developed by Measured Progress to increase the efficiency and effectiveness of the process. Details of the implementation are described in the Process Report.

1.2. Technical Advice

Throughout the development of the ALS process, the technical procedures implemented were guided by the advice of a Technical Advisory Committee on Standard Setting (TACSS)—a six-member group that collectively represents expertise in standard setting and experience with the NAEP—and the Contracting Officer's Representative Dr. Susan Loomis, the Governing

Board's Assistant Director for Psychometrics. Recommendations from the Committee on Standards, Design, Analysis, and Measurement (COSDAM) of the Governing Board were also followed. Results from all ALS meetings were presented to both the TACSS and COSDAM after each meeting. Extensive discussions with the TACSS were summarized for each TACSS meeting, and recommendations were noted. Appendix A contains the summaries of all the TACSS meetings. Below are the names and affiliations 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, 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, Michigan State University

Additionally, an internal Technical Advisory Group (TAG) at Measured Progress was called upon for guidance with issues particular to the implementation of the BoW method. Specifically, this group was asked to advise on the order of the BoWs presented to the panelists and the computation of cut scores from the pinpointing stage of the BoW method. Because the planned monthly TAG meeting was found to be unnecessary, the group met on an as-needed basis. Recommendations from the TAG were presented to the Contracting Officer's Representative and the TACSS for final recommendations. Below are the names and positions of TAG members.

- Dr. Stuart R. Kahl, Founding Principal
- Mr. Tim Crockett, Senior Vice President, Client Services
- Mr. Thomas Squeo, Chief Information Officer
- Dr. Michael L. Nering, Assistant Vice President, Research and Analysis
- Dr. Phil Robakiewicz, Director, Client Services

1.3. 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 the Process Report.

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.

1.4. Organization of the Report

This report is divided into three main sections: materials and procedures, cut score evaluation, and special study analysis.

1. *Materials and Procedures:* This section describes technical procedures implemented and materials given to the panelists during all the ALS meetings. Technical procedures include the division of panelists into groups and tables, calculation of student ability scores, selection of forms for inclusion in the study, assignment of forms to groups, selection of bodies of work (BoWs) for inclusion in the study, calculation of cut scores, presentation of post-round feedback, and selection of potential exemplar BoWs. Materials provided to panelists that are discussed in this section include both those presented using BoWTIE and those presented on paper.
2. *Cut Score Evaluation:* This section describes how the cut scores resulting from the ALS meeting were evaluated, including estimates of error due to panelist sampling, variability of cut scores, and standard error of panelist estimates.
3. *Special Study Analysis:* This Technical Report also provides information about the technical aspects of a special study Measured Progress conducted during the course of the ALS project. The special study was implemented to provide the Governing Board with information for exploring the relationship between performance on the 2011 assessment and performance on the 2007 assessment. This section describes the results of the special study that was implemented right after the operational ALS meeting.

Chapter 2—MATERIALS AND PROCEDURES

This chapter describes the technical aspect of preparing the materials provided to panelists during the Achievement Levels–Setting (ALS) meeting and the technical procedures implemented during and after the meeting. The 10 subsections of this chapter describe the division of panelists into groups and tables, the calculation of student ability estimates, the division of the writing tasks into two pools, the calculation of cut scores, the selection of bodies of work (BoWs) with high rates of disagreement, the selection of BoWs for pinpointing, the computations and presentations of post-round feedback, the collection of final recommendations through the consequences data questionnaire, the selection of potential exemplar BoWs, and the evaluation of the process.

Other materials prepared for the meeting were sent to panelists in advance. This is consistent with the belief that distributing advanced materials is considered the first step in training panelists (Cizek & Bunch, 2007; Loomis, 2012; Raymond & Reid, 2001). The following materials were sent to panelists in advance:

- meeting agenda
- 2011 NAEP Writing Framework
- achievement levels descriptions (ALDs)
- The Nation's Report Card for 2011
- briefing booklet

The Process Report describes additional communications used to prepare panelists for the ALS process.

2.1. Division of Panelists Into Groups and Tables

For the operational ALS meeting, 55 panelists (27 for grade 8 and 28 for grade 12) were convened. Each group (Group A and Group B) within the grade-level panel consisted of 13 or 14 panelists. Each group was further divided into table groups (Table 1, Table 2, and Table 3) of four or five panelists each for individual work and group discussion. Similarly, for the pilot study, 18 panelists for each grade level were subdivided into two tables for each group (Group A and Group B), with four to five panelists each. The demographic attributes of panelists were

considered when assigning members to groups and tables to maximize their equivalence; otherwise, the assignments were random. The goal was to have groups and tables as equal as possible with respect to panelist type (i.e., teacher, nonteacher educator, or general public), gender, region, and race/ethnicity.

The task pool and panelists were divided into two corresponding sets, A and B, to reduce the number of student responses reviewed by each panelist and thus, minimize the cognitive demand on the panelists. The division also creates a design that allows the reliability of the process to be evaluated (see Reliability section). The division of the task pool is described in subsection 2.3.1. Panel characteristics are described in detail in the Process Report.

2.2. Calculation of *Expected a Posteriori* (EAP) Ability Scores

Students taking NAEP assessments are not given individual-level scores. However, student work must be placed on a score scale in order to implement the BoW standard-setting methodology. Therefore, a score—an *expected a posteriori* (EAP) ability estimate—was calculated for each student (j) used in the standard setting, for the purpose of carrying out the process. An EAP estimate ($\hat{\theta}_j$) is calculated using Equation 2.1.

$$\hat{\theta}_j = \frac{\sum_{q=1}^{Q} X_q \left[L_j \left(X_q \right) \right] A\left(X_q \right)}{\sum_{q=1}^{Q} \left[L_j \left(X_q \right) \right] A\left(X_q \right)} \tag{2.1}$$

where

q indexes the quadrature points (q = 1, 2, 3, … , Q),

X_q represents the quadrature points of a Gaussian distribution,

A represents weights for the prior Gaussian distribution, N~(0,1), and

L_j represents the likelihood function.

The likelihood function (L_j) in Equation 2.1 is a function of the quadrature points (X_q). Generally, the likelihood function is defined as a function of θ, as shown in Equation 2.2,

$$L_j = \prod_{i=1}^{I} P_i\left(u_i \mid \theta \right), \tag{2.2}$$

which is the product of the probability of the response u to the item i conditional on the θ. Note that the generalized partial-credit model for the two responses with the score from 1 to 6 for the writing prompts was utilized in the EAP estimation process. Equation 2.1 is solved for $\hat{\theta}_j$ by substituting Equation 2.2 into Equation 2.1, using the quadrature point values as the values for θ in Equation 2.2. Item Response Theory (IRT) parameters provided by ETS were used in estimating EAP scores.

2.3. Description of Task Pools

Materials used at the ALS meeting included writing tasks, task statistics, and student performance data from the 2011 NAEP grade 8 and grade 12 writing assessments. Each NAEP writing assessment consists of 44 forms. Each form is made up of two writing tasks specifying different communicative purposes. According to the 2011 NAEP Writing Framework, the three communicative purposes for writing are "to persuade; to explain; and to convey experience, real or imagined" (National Assessment Governing Board, 2010, p. vi). In total, there are 22 unique writing tasks. There are also several task types based on included stimuli such as text, visual, audio, and video. The distribution of writing tasks for each writing purpose and task type is presented in Table 2-1. Appendix B provides further information for each of the tasks.

Table 2-1. 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

2.3.1. Selection of Forms and Assignment of Forms to Groups

A total of 11 forms were selected from the available 44 such that all 22 tasks were represented on the 11 forms selected. Further, each group (A and B) was assigned seven forms such that four were unique to the group and three were common to both groups. The first common form contained two tasks that were released to the public—one task for each purpose of

writing. For grade 8, the paired tasks marked for release were a “to convey” task and a “to persuade” task. For grade 12, the paired tasks were “to convey” and “to explain.” The second common form contained the remaining task marked for release that was paired with a task not marked for release. The third common form contained two tasks that were not marked for release. Assignment of tasks to groups occurred such that the average difficulty (i.e., *mean score*) for the tasks was about equal between the two groups. For grade 8, the average mean score was 3.36 for Group A and 3.3 for Group B. For grade 12, the average mean score was 3.72 for Group A and 3.66 for Group B. The assignment of unique forms to groups created a design that allowed for the reliability of the process to be evaluated (see Section 3.3). Table 2-2 presents information on the task and form assignment.

Table 2-2. 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	8_14	
			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_18	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_22	
			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	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

2.3.2. Selection of Bodies of Work

A total of 50 BoWs were selected from the identified forms for inclusion in the rangefinding classification tasks for the first two rounds in the BoW standard-setting process, as described in the Process Report. BoWs were eligible for selection if the student received a total raw score of 3–12, with each response scored 1–6. Additionally, neither of the responses was coded as "non-scorable", "off-task", or "blank". Stratified sampling was used to select the 50 BoWs—seven BoWs were selected from each form (eight from the common form with two tasks marked for release) uniformly across the scale of EAP ability estimates calculated for individual BoWs.

Another set of 50 BoWs was selected for inclusion in the third round of classification following the same procedure. The only additional criterion was that the BoWs selected for the first set were not eligible for selection in the second set. After the pilot study, in which the third round of classification was pinpointing, the TACSS advised that a second set of 50 BoWs be classified by the panelists for the third round of classifications. The issue regarding cut score computation based on pinpointing data led to the TACSS recommendation. A fuller discussion of the issue is found in section 2.6.

2.3.3. Test Forms Administered to Panelists

One of the first tasks panelists performed as part of their training was to take a form of the assessment containing two tasks marked for release. There were three released writing tasks for each grade, one for each writing purpose. For grade 8, the form taken by panelists was one with a "to convey" task and a "to persuade" task. For grade 12, the tasks were "to convey" and "to explain." The form administered to panelists is also the form used as a common form across rating groups from which exemplar BoWs were selected.

2.4. Calculation of Cut Scores

For the operational ALS meeting, BoWs representing the EAP scale range[1] were presented to the panelists, ordered from highest to lowest score, for classification into the four achievement level categories: Below Basic, Basic, Proficient, and Advanced. Cut scores can be determined by modeling the relationship between the panelist's classification of the BoW and the

[1] The EAP scale range represented does not include values resulting from a total raw score of 2.

actual EAP of the BoW. The logistic regression technique, described below, was used to model this relationship and to compute one cut score for each achievement level for each panelist, based on his or her classifications of the 50 BoWs.

In statistics, linear regression is used to model the relationship between a continuous dependent variable (y) and one or more predictor variables (x). The results from a simple linear regression include the slope (b) and intercept (a) of a line that best models the relationship between the dependent variable and the predictor variable. When the results from a linear regression are applied for certain values of the predictor variable, the model gives the expected value of the dependent variable given the predictor variable.

$$E(y) = a + bx \tag{2.3}$$

In this study, regression is applied to the classifications of BoWs into achievement levels given the actual EAP of the BoW. Thus, the dependent variable, classification into a certain achievement level, is no longer continuous but ordinal. When the dependent variable is not continuous, a linear relationship between the dependent variable and the predictor variable cannot be established without some transformation. Logistic regression is a method used for modeling the relationship between an ordinal dependent variable and one or more predictor variables. Logistic regression can be performed using an ordinal dependent variable, but this requires that assumptions be made about the relationship between all levels of the ordinal variable. To avoid having to make these assumptions, the ordinal variable (i.e., achievement level) can be reparameterized as a set of dichotomous variables. The dichotomous variables (z) are defined as classification at or above a certain achievement level or below that achievement level. In applying the logistic regression method to model the relationship between achievement level and the actual EAP of the BoW, reparameterization produces three dichotomous variables (z_1, z_2, and z_3), which represent classification at Basic or above (z_1), Proficient or above (z_2), and Advanced or above (z_3).

Just as with linear regression, the results of the logistic model are used to give the expected value of the dependent variable given the predictor variable. The function representing this relationship in logistic regression is

$$E(z_i) = \frac{e^{(a_i + b_i x)}}{1 + e^{(a_i + b_i x)}}, \tag{2.4}$$

where a_i and b_i are the slope and intercept from the logistic regression of z_i on x. Since the dependent variables (z_1, z_2, and z_3) are dichotomous, the expected value is equivalent to the probability that the value of the dependent variable is 1, or in this case the probability that the BoW is classified at a certain achievement level (i) or above.

$$P(z_i = 1) = \frac{e^{(a_i + b_i x)}}{1 + e^{(a_i + b_i x)}} \tag{2.5}$$

These probability curves can be plotted across different values of the predictor variable (x). The plots take on a familiar S-shaped curve form, with low probabilities for lower values of x that approach 0, and higher probabilities for higher values of x that approach 1. In the application of this method to the BoWs, this means that BoWs with lower actual EAPs are expected to have lower probabilities of being classified above a certain achievement level, and BoWs with higher EAPs are expected to have higher probabilities of being classified at a certain achievement level or above. Substituting values into the equation above shows that the probability

$$\text{is less than 0.5 when } a_i + b_i x < 0,$$

$$\text{is greater than 0.5 when } a_i + b_i x > 0,$$

$$\text{and is exactly 0.5 when } a_i + b_i x = 0.$$

Once the regression coefficients have been estimated, they can be substituted into the above equations. The equations can then be solved to find a range of values for x where the probability is less than 0.5, a range of values for x where the probability is greater than 0.5, and a value for x where the probability is exactly 0.5. Next we consider what could be counted as evidence of the panelists intention. If the panelists were classifying BoWs completely at random (i.e. without actually looking at the BoW) we would consider any classifications they make to be due purely to chance (p=0.5). However we know that they are classifying BoWs with intention. So by comparison, we reason that when the probability that the BoW is classified at a certain achievement level is greater than could be expected by chance, the panelists intention is implied. Thus, we conclude that the panelists intention to classify the BoW at a certain achievement level is implied for the range of x where the probability is greater than 0.5. The cut score is defined as the value of the predictor variable where the probability changes from being what could be expected by chance, to being greater than what could be expected by chance. The intent of a

panelist to classify a BoW at the achievement level or above is implied only when the probability is greater than what could be expected by chance. Thus, three cut scores (x_1*, x_2*, and x_3*) are produced by solving the following three equations:

$$a_1 + b_1 x_1^* = 0 \tag{2.6}$$

$$a_2 + b_2 x_2^* = 0 \tag{2.7}$$

$$a_3 + b_3 x_3^* = 0 \tag{2.8}$$

The regression coefficients (a_1, b_1, a_2, b_2, a_3, and b_3) must be estimated through logistic regression before these equations can be solved to produce cut scores. Although the probabilistic formulations related to logistic regression provide the most intuitive way to understand how the cut scores are produced, further transformation is needed in the actual application of logistic regression to transform the probabilities (which are bounded by 0,1) to an unbounded scale and to establish a linear (not exponential) relationship between the dependent variable and predictor variable(s). This can be achieved by taking the log-odds of the probability:

$$\ln\left(\frac{p(z_i)}{1-p(z_i)}\right) = a_i + b_i x, \tag{2.9}$$

This model is applied to a panelist's classifications three times (once for each cut score) to estimate the regression coefficients (a_1, b_1, a_2, b_2, a_3, and b_3), which are then substituted into the above formulas to find the cut scores (x_1*, x_2*, and x_3*). For example, when the Proficient cut score is computed, the relationship between the EAP of a BoW and the probability that a panelist (j) classifies the BoW at or above Proficient is

$$P(z_{2j} = 1) = \frac{e^{(a_{2j}+b_{2j}x)}}{1+e^{(a_{2j}+b_{2j}x)}}. \tag{2.10}$$

The logistic regression is then run to estimate the panelist's (j) regression coefficients a_{2j} and b_{2j}.

$$\ln\left(\frac{p(z_{2j})}{1-p(z_{2j})}\right) = a_{2j} + b_{2j} x, \tag{2.11}$$

Finally, the equation below is solved to find the panelist's cut score for Proficient (x_{2j}*).

$$a_{2j} + b_{2j}\, x_{2j}^* = 0, \tag{2.12}$$

This method was applied for all three cut scores and for each panelist.

Calculating the three cut scores for each panelist resulted in three distributions of cut scores. A measure of central tendency of each distribution was calculated to produce the panel cut score. In statistics, common measures of central tendency include the mean and the median. Both have their advantages in any application, but the median is always less sensitive to outliers than is the mean. The medians were used as the panel cut scores for this reason. This means that the panel cut score cannot be severely affected by a single panelist who tends to set cut scores that are extremely high or extremely low. Logistic regression was used to determine each panelist's cut score, and then the median was considered the panel cut score. However, when the binary logistic regression model was fitted, the third level nesting structure, due to the fact that the panelists within a group and grade level were given the same BoWs, was not modeled. All the necessary calculations to facilitate this process were built into BoWTIE. Since the classifications varied across rounds, the panel cut scores also varied across rounds.

When results from the operational ALS meeting were presented during a TACSS meeting, it was recommended that the use of the generalized linear mixed model (GLIMM) in computing the overall cut score be investigated. Using GLMMs in the computation of overall cut scores with all panelists' rating data is appropriate in this situation because it takes into consideration that panelists did not rate unique sets of BoWs. A detailed description of the analysis is provided in Appendix C. Table C-3 of Appendix C provides a comparison of the operational cut scores and the cut scores from the GLMMs. Since there are only six cut scores to compare (three cut scores for grade 8 and three for grade 12) and eight values for the percent of students classified at each level (four for grade 8 and four for grade 12), tests of statistical significance of the difference would not have much power in this case. However, the results can still be compared. They show that the cut scores from the GLMMs were always slightly lower than the operational cut scores. Table 2-3 shows a comparison of the cut scores as well as the standard error of the cut scores.

Table 2-3. Comparison of Operational and GLMM Cut Scores

Grade	Achievement Level	Scaled Scores			Standard Errors	
		Operational	GLMM	Difference	EmpSE[A]	BootSE[B]
8	Basic	120	119	1	6.84	3.58
	Proficient	173	171	2	5.71	3.18
	Advanced	211	211	0	4.53	3.1
12	Basic	122	121	1	6.08	1.06
	Proficient	173	170	3	5.99	3.52
	Advanced	210	208	2	4.35	2.74

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

The table illustrates that the operational and GLMM cut scores are always well within 1 SE of each other. On average, the cut scores are within 0.3 EmpSEs and 0.6 BootSEs of each other. Considering that roughly two times the standard error is used to establish a 95% confidence interval for the cut score, these differences are very small, meaning that the operational and GLMM scaled scores are very similar.

2.5. Selecting Bodies of Work With High Disagreement Rates

An integral part of the BoW method is the discussion of BoWs to enhance panelists' understanding of the descriptions of the three different achievement levels (Kingston et al., 2001, p. 227; Kahl et al, 1995, p. 5). For the BoW implementation for the NAEP writing assessment, this discussion was conducted twice: once during training and once between classification Rounds 1 and 2.

The BoWs selected for discussion between Rounds 1 and 2 were carefully and deliberately chosen to maximize the effectiveness of the exercise within a minimum amount of time. The TACSS advised that the selection of BoWs for discussion be based on entropy (i.e., degree of disorder) and that BoWs be ranked to indicate the order in which they should be discussed. During the pilot study, the specific criteria for selection were developed by the COR and the CoSS. The criteria represent ways of operationalizing "entropy" for these data. Table 2-4 illustrates the following selection criteria that were created and applied hierarchically:

- 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—classification of a BoW into two or more categories by approximately the same number of panelists
- reversal—classification of a BoW by a majority of panelists was inconsistent with the modal classifications of the BoWs around it

Table 2-4 shows the rank orders of eight BoWs that were selected for discussion. The BoWs were ordered for discussion by the groups. In the event that there was not enough time to discuss all of those selected, the BoWs ordered at the end were considered least informative and least important to discuss. BoWs 14 and 15 have the same spread and highest ranks. However, BoW 15 was classified into more levels; thus, BoW 15 was ranked 1 and BoW 14 was ranked 2. BoWs 4 and 6 were ranked 3 and 4 since they were both categorized in three consecutive categories and thus, ranked lower than BoWs 14 and 15. Each of BoWs 4 and 6 was classified into four categories, but these two BoWs do not have the same rank because BoW 4 is considered to have more spread. BoWs 2, 21, and 22 were all selected for "split," where a 9/9 split was ranked higher than an 8/10 or 10/8 split. Lastly, BoW 11 was selected for "reversal" since the modal classification of all the BoWs around it was Proficient, but its modal classification was Basic.

Table 2-4. Example of a Classification Tally Indicating BoWs Selected for Discussion

BoW ID (Entropic Rank)	*Counts*			
	Below Basic	*Basic*	*Proficient*	*Advanced*
1	15	3	0	0
2 (5)	9	9	0	0
3	5	13	0	0
4 (3)	3	13	2	0
5	4	14	0	0
6 (4)	0	15	2	1
7	0	13	5	0
8	0	13	5	0
9	0	6	12	0
10	0	6	12	0
11 (8)	0	13	5	0
12	0	0	15	3
13	0	0	15	3
14 (2)	3	0	13	2
15 (1)	1	2	12	3
16	0	0	14	4
17	0	0	17	1
18	0	0	12	6
19	0	0	10	8
20	0	0	15	3
21(6.5)	0	0	10	8
22(6.5)	0	0	8	10

2.6. Selecting Bodies of Work (BoWs) for Pinpointing

For the pilot study, panelist cut scores resulting from Round 2 classifications were used to determine the range within which to select BoWs for the pinpointing round (i.e., Round 3). Fifteen BoWs were selected for each achievement level using the following steps:

1. First, the upper and lower scale score bounds were transformed to the theta scale.
2. Next, the theta range was calculated by subtracting the lower bound from the upper bound.
3. Then, the range was divided by 14 to obtain the incremental interval. The 14 intervals created 15 endpoints, which were used as the scores for the 15 BoWs that were to be selected.

BoWs were randomly selected among eligible BoWs nearest each theta increment starting with the lower bound and ending with the upper bound. The selection of the BoWs for pinpointing was built into BoWTIE with an opportunity to examine and replace BoWs not deemed appropriate for any reason. During the pilot study, the content facilitators[2] examined each of the BoWs that were included for the pinpointing round.

For the operational ALS meeting, however, the idea of using the BoWs for pinpointing was abandoned based on extensive discussion by the TACSS. The pilot study revealed that computing the cut scores based on pinpointing data was problematic. Computing a cut score based on 15 BoWs yielded very unstable cut scores. Two alternatives were considered for adding stability to the computed cut scores for Round 3: (1) using classification data for all 45 pinpointing BoWs or (2) combining Round 2 rangefinding and Round 3 pinpointing data. Given that the pinpointing classifications were dichotomous (e.g., below Proficient or Proficient or above), using these data in a logistic regression ignores the possibility that a "below Proficient" classification, for example, might not necessarily mean that the panelist deemed the performance to be Basic, since "below Proficient" might also mean "below Basic." Combining data from Rounds 2 and 3 to compute Round 3 cut scores was also deemed problematic because panelists who might have a dramatically different understanding of the ALDs after Round 2 would not have been able to affect their cut scores appropriately. These issues led the TACSS to instead recommend replacing the Round 3 classifications with a rangefinding round in which a new set of 50 BoWs would be rated.[3] The exact same procedure for selecting BoWs that was adopted for Round 1 (and Round 2) was utilized. The resultant 50 BoWs were verified to be a totally different set from the set used for Rounds 1 and 2.

2.7. Computation and Presentation of Post-Round Feedback

After each round of classifications, feedback was provided to the panelists to inform their judgment for the next round of classifications. Feedback after the first round of classifications included the cut score results, a cut score distribution chart, and cut score location charts. In the

[2] Facilitation of the ALS process implementation occurs mostly in the grade level breakout room. Guided by the chief of standard setting, the process facilitator is primarily in charge of providing information, giving direction, and ensuring that "the train leaves on time" while the content facilitator takes the lead for parts of the process for which knowledge of the assessment is imperative.

[3] A simulation study to evaluate whether implementing pinpointing improves cut score precision was also performed. Findings from the study were presented to the TACSS during an online meeting. The write up for the study is included in Appendix D.

field trial, *p*-value data for the tasks were also provided, but were dropped for the pilot study and the ALS meeting upon advice from the TACSS. The original design for this work (Measured Progress, 2011), specified a graphical version of the Reckase chart (Cizek and Bunch, 2007) was to be provided to the panelists before the second round of classification. The purpose was to provide panelists with information on how tasks vary by difficulty. This feedback was not provided because of the task characteristic curves were so similar that the feedback would not be particularly informative. This decision was reached in consultation with the Contracting Officer's Representative and a TACSS member. The *p*-value feedback and Reckase chart prepared for the field trial are in Appendix E.

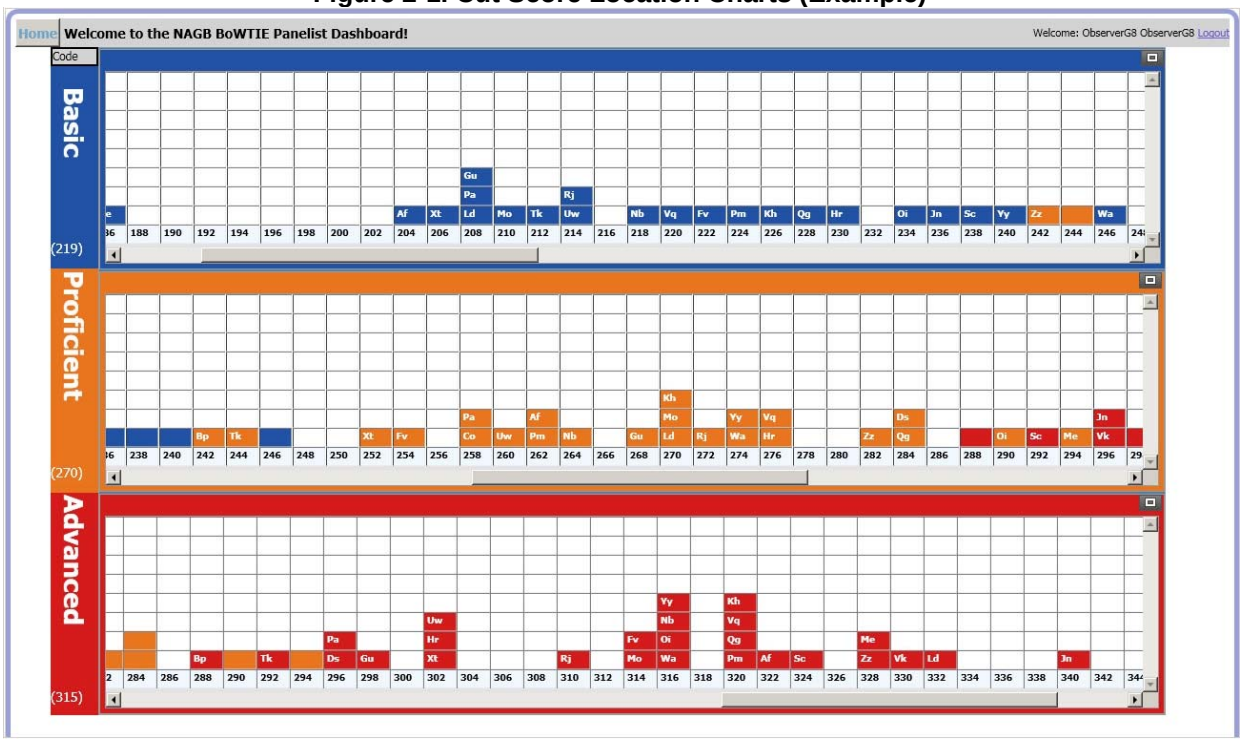
2.7.1. Cut Score Results

After each round of classifications, a cut score for each achievement level was computed, for each grade level panel, as described in an earlier section. The median was used as the grade level cut score. These median cut scores were presented to panelists on a pseudo-NAEP scale after each round. The pseudo-NAEP scale values were derived as a linear transformation of the NAEP scale. This scale was purposely different from the NAEP score scale used for reporting so that panelists would not know the official NAEP results before the results were approved for release by the Governing Board and to prevent panelists from using NAEP cut scores from other assessments to set their cut scores for the 2011writing assessment. Further, a different linear transformation was used for each grade to discourage panelists from comparing their cut scores.

2.7.2. Cut Score Location Chart

The cut score location charts were constructed within BoWTIE. These charts displayed the distribution of cut scores across all panelists for a given round of classifications, thus, providing information about the inter-rater consistency of the panelists' judgments. One chart highlighting each achievement level was generated. Panelist cut scores were identified using "secret" codes to protect confidentiality. Additionally, cut scores were rounded to the nearest even integer to reduce the amount of scrolling required in BoWTIE in order to view the entire score range of the chart. The cut score location charts also displayed the median cut score for the panel along the left-hand side of the screen. An example of the cut score location charts from BoWTIE is shown in Figure 2-1.

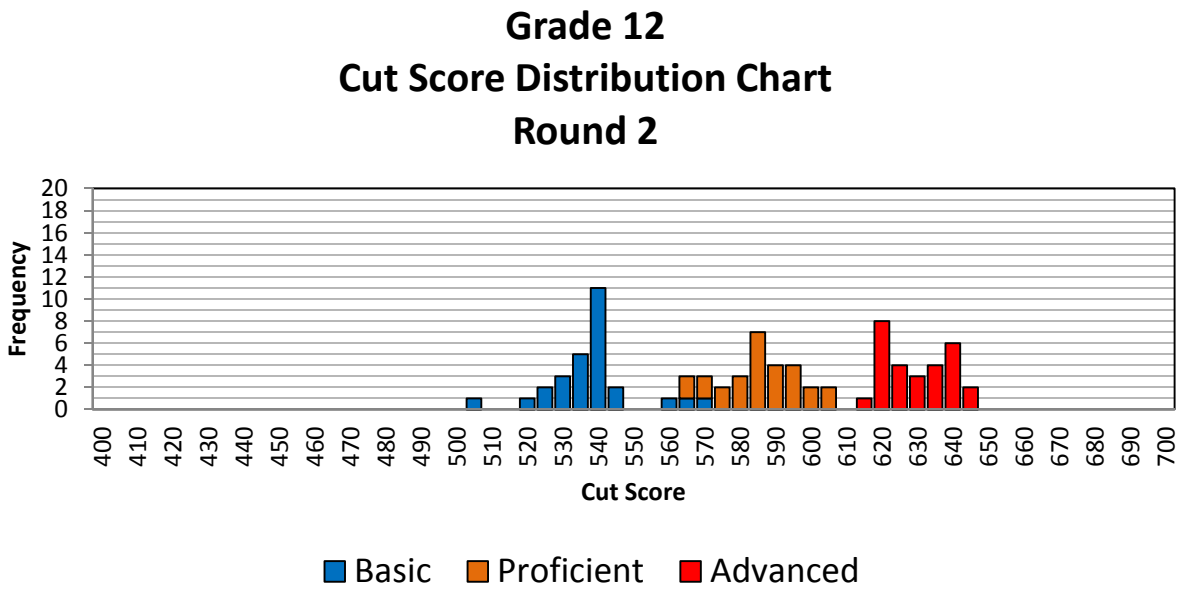
Figure 2-1. Cut Score Location Charts (Example)



2.7.3. Cut Score Distribution Chart

The cut score distribution chart displayed the distribution of cut scores for all panelists for each grade level for a given round of classifications. An example of the cut score distribution chart is shown in Figure 2-2. Cut scores were grouped within 5-point scale score intervals such that the entire scale showing cut scores for all panelists for each achievement level could be observed on a single screen. Cut scores were color-coded by achievement level. Although the exact same information is provided in the cut score location charts and the cut score distribution chart, the latter allowed panelists to more clearly focus on the entire distribution and evaluate the cut score locations.

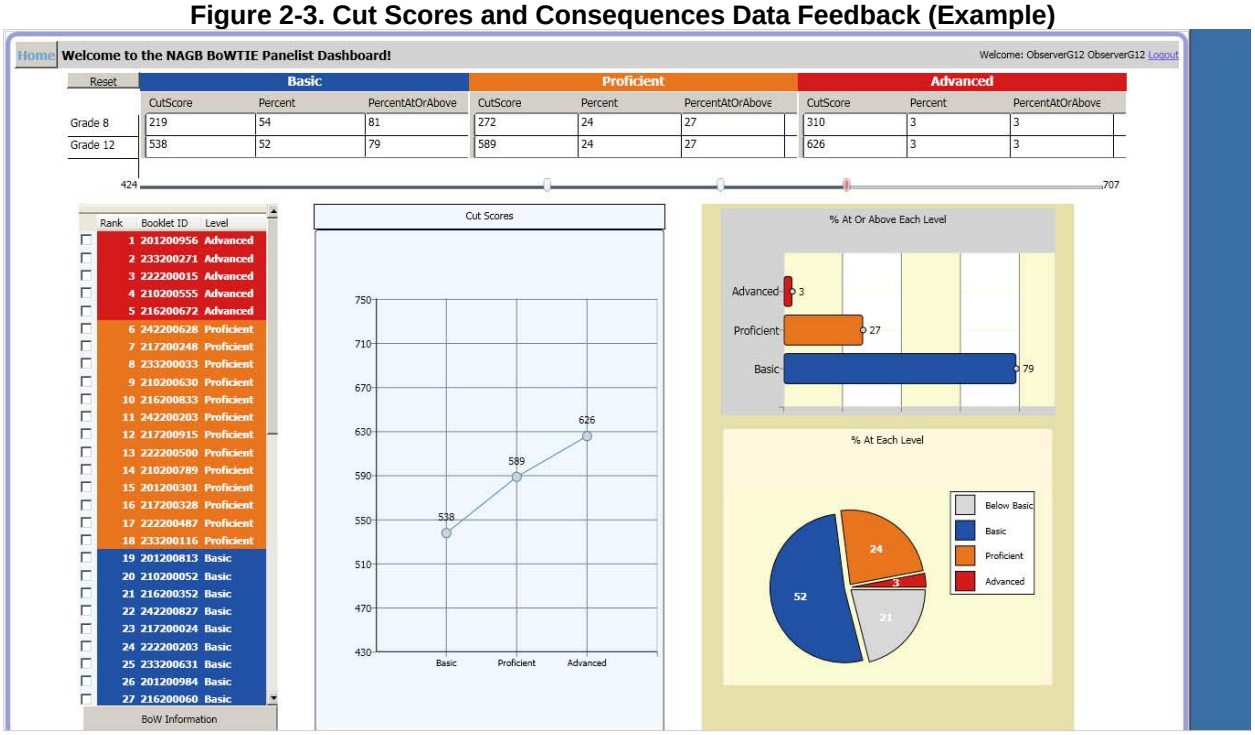
Figure 2-2. Cut Score Distribution Chart (Example)



2.7.4. Consequences Data Feedback

Consequences data feedback consists of the percentages of students performing at or above the grade-level panel cut scores. The frequency distribution of student performances based on the 2011 assessments were provided to Measured Progress by ETS—the Design, Analysis, and Reporting contractor to the National Center for Education Statistics (NCES) for the NAEP program. The frequency distribution tables can be found in Appendix F. BoWTIE displayed the consequences data feedback on an interactive consequences data screen, as pictured in Figure 2-3. Panelists could use the cut score adjuster (located at the top of the screen) to experiment with what the consequences data would look like given alternative cut scores (e.g., a panelist's individual cut scores). The bar graph on the right side of the screen shows the percentages of students scoring at or above the cut scores, and the pie chart on the right side of the screen shows the percentage of students scoring within each achievement level. A BoW list is displayed on the left side of the screen, where BoWs are highlighted based on how they would be classified using the cut scores resulting from the classification round. Classifications are also indicated in the Level column.

Figure 2-3. Cut Scores and Consequences Data Feedback (Example)



2.8. Collection of Final Recommendations

Consequences data feedback was presented to panelists using the BoWTIE application, as described in the previous section and as displayed in Figure 2-3. Consequences data feedback was presented to panelists after Rounds 2 and 3. After Round 3, panelists were asked to complete a Consequences Data Questionnaire indicating whether they felt the proportion of students scoring at or above the panel cut scores seemed reasonable or should be higher or lower, based on their overall knowledge of student performance in writing, as well as their understanding of the achievement levels descriptions—always the primary criterion. This information was collected for reporting to the Governing Board to indicate whether additional changes should be considered for setting the final cut scores. Panelists' responses to the questions are discussed in detail in the process evaluation results section for each grade in the Process Report. The results, shown in Table 2-5, indicated that on average, panelists agreed (% agreement ranges from 87 to 93) that the results of the cut scores matched their expectation for the proportion of students that should be classified as Basic, Proficient, and Advanced.

Table 2-5. Summary of Responses to Consequences Data Question

Given your understanding of student performance at the X achievement level, does this percentage reflect your expectation about the proportion of students whose NAEP score would be at or above the X cut score?

Grade	*Matches Expectation*	*n*	*% Yes*
12 (*N*=30)	Basic	26	87
	Proficient	27	90
	Advanced	27	90
8 (*N*=27)	Basic	22	81
	Proficient	25	93
	Advanced	25	93

Panelists, after having reviewed the consequences data, were also asked whether they would change one or more of the achievement levels if they could (1=Yes, 0=No). Panelists mostly responded no, saying that they would not change the levels, as shown in Table 2-6.

Table 2-6. Cut Score Change Recommendation

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?

Change Achievement Level	*n*	*% No*
Grade 12 (*N*=28)	25	89
Grade 8 (*N*=25)	17	68

Some panelists also recommended specific cut score values (*n*=10). The panel cut score and the recommended cut score are shown in Table 2-7. As can be seen from comparing the panel cut score to the panelist's recommended cut score, the panelists often recommended the panel cut score, indicating that they did not recommend a change. When the recommended score was different from the panel cut score, the changes were mostly small. Moreover, there is no clear indication that the panelists who recommended changes were recommending changes close to their individual cut score (listed in the final column of the table). As can be observed, the recommendation is sometimes closer to the panelist's cut score (if the panelist recommended a lower cut score and the panelist's individual cut score was lower than the panel cut score, or if

the panelist recommended a higher cut score and the panelist's individual cut score was higher than the panel cut score) and sometimes further away (vice versa). Finally, one grade 8 panelist suggested three changes based on the percent of BoWs classified in the different achievement levels. This panelist recommended that the cut score be changed so that the percent of students classified as Basic (80% based on the panel cut score), Proficient (24% based on the panel cut score), and Advanced (3% based on the panel cut score) be changed so that 70% of the BoWs be classified as Basic or above, 25% be classified as Proficient or above, and more than 3% (a specific percent was not recommended by the panelist) be classified as Advanced or above. This represents one recommendation for a large change in the Basic cut score, a small change in the Proficient cut score, and an unknown amount of change for the Advanced cut score.

Table 2-7. Panelist Recommended Cut Scores

Grade	*Panelist*	*Panelist's Recommended Cut Score**			*Panelist's Computed Cut Score*		
		Basic	*Proficient*	*Advanced*	*Basic*	*Proficient*	*Advanced*
8	1	225[A]	277[A]	316[A]	215[C]	276[B]	319[B]
	2	219	272	310	211	293	325
	3	230[A]	272	310	214[C]	279	317
	4	215[A]	270[A]	305[A]	244[C]	281[C]	319[C]
	5	200[A]	250[A]	310	222[C]	270[B]	305
	6	219	272	308[A]	239	269	312[C]
	7	219	272	315[A]	192	234	309
12	8	530[A]			512[B]	551	610
	9		589	626	560	601	616
	10	540[A]	590[A]	630[A]	496[C]	567[C]	628[B]

Note: Grade 8 Cut Scores are 219 (Basic), 272 (Proficient), and 310 (Advanced);
Grade 12 Cut Scores are 538, 589, and 626.

* From Consequences Data Questionnaire

[A] Panelist recommended a cut score different from the panel cut score.

[B] Panelist's recommendation is closer to panelist's individual cut score.

[C] Panelist's recommendation is further from panelist's individual cut score.

2.9. Selection of Potential Exemplar Bodies of Work

After the classification rounds were completed, panelists were asked to make recommendations for exemplar BoWs (i.e., student work illustrative of the knowledge, skills, and abilities representing each achievement level). Potential exemplar BoWs were drawn from the common form containing two tasks marked for public release. A total of 16 BoWs, eight from each rangefinding set, were classified into achievement levels based on cut scores from

Round 3. Those that were classified into Basic, Proficient, or Advanced levels were presented to the panelists for recommendation. Panelists were asked to rate how well each selected potential exemplar illustrated the achievement level it was classified as. Panelists were asked to indicate whether each exemplar should definitely be used ("Very Good"), was okay to use ("Okay"), or should not be used as an exemplar ("Do not use"). The panelists were allowed to discuss potential exemplars with other panelists, but they had to provide their ratings in BoWTIE independently.

A summary of panelists' ratings as well as their comments were presented to the TACSS. One BoW was selected for each achievement level at each grade based on the following criteria:

- 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.

2.10. Evaluation of the Process

Panelists completed a process evaluation form after each major ALS activity (e.g., at the end of each day and after each BoW classification round). Process evaluations were administered using the BoWTIE application. Panelists were asked to evaluate several aspects of the process, including the following:

- receipt and adequacy of pre-meeting materials
- clarity of the overview and purpose of the ALS meeting
- understanding of the NAEP assessment
- clarity of instruction and panelist roles
- utility of the practical experiences provided during the ALS meeting
- understanding of the achievement level descriptions and understanding of the relationship between the cut scores and the achievement levels
- understanding of the method and tasks involved
- confidence in the process and results

Several 5-point Likert items addressed each one of these topics. For each item the mean value for the responses and the standard deviation were calculated. As described in the Process Report, the

summarized evaluations supported the validity of the ALS process. Results from the process evaluations were also used to clarify areas of confusion during the course of the meeting. Open responses were also solicited and used mainly to inform additional aspects of the process that may not have been captured by the other items. A more detailed summary of responses to all process evaluation questions are included in Appendix B of the Process Report.

Chapter 3—CUT SCORE EVALUATION

This chapter describes how the Round 3 cut scores resulting from the Achievement Levels–Setting (ALS) meeting were evaluated in terms of estimates of error due to panelist sampling (3.1), variability of cut scores (3.2), and standard error of panelist estimates (3.3). The estimates of error due to panelist sampling was evaluated using different standard error estimates. The variability of cut scores was evaluated using the mean absolute deviation algorithm along with an analysis of how panelists' cut scores changed from one round to the next. The standard error of panelist estimates was evaluated using results from the two groups (A and B) within each panel.

3.1. Estimates of Error Due to Panelist Sampling

One potential source of error in the cut scores is related to the sampling of the panelists. Estimates of the error related to sampling of the panelists were calculated using two standard error calculation techniques. 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. Because 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. If n is the number of observations and is odd, then the kth moment of the median is given using Equation 3.1,

$$E(median^k) = \frac{n!}{[\left(\frac{n-1}{2}\right)!]^2} \int_{-\infty}^{\infty} x^k \left[F(x)(1-F(x))\right]^{\left(\frac{n-1}{2}\right)} f(x)dx, \quad (3.1)$$

where $f(x)$ is the probability density function of the data, and $F(x)$ is the cumulative distribution function. A similar expression holds when n is even. Applying the transformation $y = F(x)$, Equation 3.1 becomes

$$E(median^k) = \frac{n!}{[\left(\frac{n-1}{2}\right)!]^2} \int_0^1 [\psi(y)]^k \left[y(1-y)\right]^{\left(\frac{n-1}{2}\right)} yd, \quad (3.2)$$

where $\psi(y)=F^{-1}(y)$. Estimating $\psi(y)$ by the observed order statistics results in the following estimator for the kth moment of the median:

$$A_{kn} = \sum_{i=1}^{n} x_{(i)}{}^{k} W_{i,} \tag{3.3}$$

where

$$W_i = \frac{n!}{\left[\left(\frac{n-1}{2}\right)!\right]^2} \int_{\frac{i-1}{n}}^{\frac{i}{n}} y^{\frac{n-1}{2}} (1-y)^{\frac{n-1}{2}} dy \tag{3.4}$$

and $x_{(i)}$ is the ith order statistic. The integral for W_i can be evaluated in closed form with the number of terms in the solution increasing with n, the sample size. The values of n used in the standard setting were above the upper limit for which values of W_i were provided in Maritz and Jarrett (1978). Thus, the values of W_i needed for the study were calculated as part of the study. The value for the square of the standard error for the median was then estimated by the quantity $(A_{2n} - A_{1n}{}^2)$. Similar formulas were solved for when n was even.

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 3-1 through 3-6 present these standard error estimates for grades 8 and 12, respectively, across panelist demographic groupings, tables, and groups (i.e., A and B) for each round. First, these results can be used to illustrate evidence of variation across groups of panelists. Since panelists were arranged into tables and groups in such a way as to minimize differences in the tables and groups, the summary statistics reported for each table and group within a certain grade and round should be comparable to each other. In other words, it should be the case that the mean absolute deviations (MADs) and standard errors (SEs) are similar across tables and groups. Despite efforts to create equivalent tables and groups by minimizing the differences in these groups, some variability in MADs and SEs is seen across tables and groups. Second, it would be expected that these differences across groups would decrease after each

round since the ALS process is used to refine the decisions of the panelists and result in smaller overall MADs and SEs. The results show that the MADs and SEs are usually greatest in earlier rounds and decrease by Round 3 (Tables 3-1through 3-6) as expected. Because of the general pattern of decrease in variability across groups and decrease in overall MADs and SEs, the validity argument is supported.

Table 3-1. Estimates of Standard Error of Cut Scores for NAEP—Writing Grade 8, Round 1, 2011

Table/ Group	Achievement Level	Median Theta	Median NAEP	Standard Error		MAD	Scaled Score		Percent of
				EmpSE	BootSE		Min	Max	
Table 1	Advanced	1.67	209	8.06	7.11	8.0	209	300	3.66
	Proficient	0.40	165	10.31	8.40	7.5	165	208	31.91
	Basic	-1.10	111	8.83	9.14	7.5	111	164	50.80
	Below Basic						0	110	13.63
Table 2	Advanced	2.08	224	5.56	5.25	4.0	224	300	0.91
	Proficient	0.92	182	7.20	6.81	10.0	182	223	17.12
	Basic	-0.40	136	8.97	9.34	6.0	136	181	48.47
	Below Basic						0	135	33.50
Table 3	Advanced	1.96	219	7.74	8.66	2.5	219	300	1.45
	Proficient	0.48	167	7.74	8.10	6.5	167	218	31.43
	Basic	-0.88	119	9.44	8.63	11.0	119	166	48.56
	Below Basic						0	118	18.55
Table 4	Advanced	1.86	216	6.74	6.96	4.0	216	300	2.01
	Proficient	0.42	165	5.95	5.85	9.0	165	215	33.09
	Basic	-0.70	125	8.05	7.63	6.0	125	164	40.81
	Below Basic						0	124	24.09
Table 5	Advanced	2.02	221	17.66	16.95	25.0	221	300	1.19
	Proficient	0.60	171	11.87	12.17	13.0	171	220	27.12
	Basic	-1.16	109	24.49	23.16	33.0	109	170	59.18
	Below Basic						0	108	12.52
Table 6	Advanced	1.84	215	7.62	7.41	8.0	215	300	2.19
	Proficient	0.88	181	10.90	10.82	13.0	181	214	16.80
	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	4.00	4.28	7.0	221	300	1.24
	Proficient	0.61	172	5.11	4.93	10.0	172	220	26.61
	Basic	-0.85	120	6.59	6.04	12.0	120	171	52.51
	Below Basic						0	119	19.64
Group B	Advanced	1.86	216	5.12	5.30	14.0	216	300	2.01
	Proficient	0.60	171	5.21	5.05	12.0	171	215	26.72
	Basic	-0.79	122	7.94	7.95	13.0	122	170	49.86
	Below Basic						0	121	21.41
All	Advanced	1.87	216	6.62	2.87	12.0	216	300	1.89
	Proficient	0.61	171	5.76	3.29	11.0	171	215	26.41
	Basic	-0.85	120	6.36	5.39	12.0	120	170	52.06
	Below Basic						0	119	19.64

Table 3-2. Estimates of Standard Error of Cut Scores for NAEP—Writing Grade 8, Round 2, 2011

Table/ Group	Achievement Level	Median Theta	Median NAEP	Standard Error		MAD	Scaled Score		Percent of
				EmpSE	BootSE		Min	Max	
Table 1	Advanced	1.97	220	21.67	16.67	10.0	220	300	1.39
	Proficient	0.40	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	4.0	222	300	1.15
	Proficient	0.77	177	8.72	7.51	8.0	177	221	21.53
	Basic	-0.60	129	4.61	4.08	6.0	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	5.0	170	218	28.75
	Basic	-1.00	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	4.0	216	300	2.01
	Proficient	0.76	177	5.04	5.39	3.0	177	215	20.67
	Basic	-1.24	106	8.21	7.30	3.0	106	176	66.30
	Below Basic						0	105	11.01
Table 5	Advanced	2.02	221	11.07	9.96	11.0	221	300	1.19
	Proficient	0.71	175	9.66	10.17	11.0	175	220	23.32
	Basic	-0.85	120	14.19	14.08	18.0	120	174	56.10
	Below Basic						0	119	19.40
Table 6	Advanced	1.84	215	3.04	3.05	3.0	215	300	2.10
	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	5.0	221	300	1.24
	Proficient	0.60	171	4.02	3.96	6.0	171	220	27.06
	Basic	-0.83	121	3.19	2.99	8.0	121	170	51.49
	Below Basic						0	120	20.21
Group B	Advanced	1.88	217	2.38	2.28	4.0	217	300	1.84
	Proficient	0.73	176	3.84	3.90	7.5	176	216	21.80
	Basic	-0.87	120	6.96	6.79	14.0	120	175	57.19
	Below Basic						0	119	19.17
All	Advanced	1.98	220	6.35	2.06	5.0	220	300	1.34
	Proficient	0.68	174	5.72	3.21	8.0	174	219	24.26
	Basic	-0.85	120	4.34	2.75	9.0	120	173	55.00
	Below Basic						0	119	19.40

Table 3-3. Estimates of Standard Error of Cut Scores for NAEP—Writing Grade 8, Round 3, 2011

Table/ Group	Achievement Level	Median Theta	Median NAEP	Standard Error		MAD	Scaled Score		Percent of
				EmpSE	BootSE		Min	Max	
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	2.20	2.26	2.0	113	172	58.30
	Below Basic						0	112	14.93
Table 2	Advanced	1.96	219	10.83	10.71	17.0	219	300	1.43
	Proficient	0.42	165	8.28	6.53	6.0	165	218	33.31
	Basic	-0.58	129	7.92	8.46	4.0	129	164	37.79
	Below Basic						0	128	27.47
Table 3	Advanced	1.40	200	5.44	4.13	2.0	200	300	7.06
	Proficient	0.81	179	5.98	6.83	0.5	179	199	14.34
	Basic	-1.09	112	5.30	5.14	6.5	112	178	64.79
	Below Basic						0	111	13.81
Table 4	Advanced	1.85	216	4.44	4.42	4.0	216	300	2.01
	Proficient	0.65	173	4.51	4.25	6.0	173	215	24.36
	Basic	-0.80	122	10.33	11.07	3.0	122	172	52.81
	Below Basic						0	121	20.83
Table 5	Advanced	1.71	210	8.32	7.65	6.0	210	300	3.20
	Proficient	0.59	171	24.55	23.89	36.0	171	209	25.87
	Basic	-0.76	123	14.33	13.97	22.0	123	170	48.60
	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	11.0	211	300	3.01
	Proficient	0.65	173	5.33	5.40	8.0	173	210	23.76
	Basic	-1.00	115	3.18	2.96	7.0	115	172	57.45
	Below Basic						0	114	15.79
Group B	Advanced	1.74	212	3.36	3.18	7.0	212	300	2.91
	Proficient	0.62	172	3.69	3.70	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	9.0	211	300	3.01
	Proficient	0.65	173	5.71	3.18	7.0	173	210	23.76
	Basic	-0.84	120	4.53	3.10	9.0	120	172	53.60
	Below Basic						0	119	19.64

Table 3-4. Estimates of Standard Error of Cut Scores for NAEP—Writing Grade 12, Round 1, 2011

Table/ Group	Achievement Level	Median Theta	Median NAEP	Standard Error		MAD	Scaled Score		Percent of
				EmpSE	BootSE		Min	Max	
Table 1	Advanced	1.99	220	1.57	1.66	1.0	220	300	1.17
	Proficient	0.98	185	6.80	7.95	3.0	185	219	15.15
	Basic	-0.73	124	9.93	9.74	5.0	124	184	60.75
	Below Basic						0	123	22.93
Table 2	Advanced	2.05	222	11.69	12.03	16.0	222	300	0.96
	Proficient	0.58	171	9.58	10.11	2.0	171	221	28.07
	Basic	-0.97	116	17.38	15.84	10.0	116	170	54.18
	Below Basic						0	115	16.78
Table 3	Advanced	1.73	211	5.40	5.19	6.0	211	300	2.94
	Proficient	0.59	171	10.11	9.77	13.0	171	210	26.09
	Basic	-0.92	118	4.91	5.50	1.5	118	170	53.00
	Below Basic						0	117	17.97
Table 4	Advanced	1.83	215	3.45	3.38	4.0	215	300	2.12
	Proficient	0.36	163	11.03	8.78	7.0	163	214	35.77
	Basic	-0.53	132	6.85	7.40	4.0	132	162	33.12
	Below Basic						0	131	28.99
Table 5	Advanced	1.74	212	4.57	4.40	6.0	212	300	2.72
	Proficient	0.49	167	10.36	10.41	9.0	167	211	29.61
	Basic	-0.83	121	7.89	8.73	9.0	121	166	47.16
	Below Basic						0	120	20.51
Table 6	Advanced	1.47	202	8.92	8.95	10.0	202	300	5.76
	Proficient	0.31	161	10.91	10.84	10.0	161	201	34.11
	Basic	-0.89	119	8.89	9.17	5.0	119	160	41.47
	Below Basic						0	118	18.66
Group A	Advanced	1.97	220	3.50	3.69	4.0	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	10.0	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	12.0	163	211	34.60
	Basic	-0.82	121	4.06	3.96	9.0	121	162	41.73
	Below Basic						0	120	20.73
All	Advanced	1.80	214	6.58	2.47	7.5	214	300	2.31
	Proficient	0.56	170	6.47	4.41	14.0	170	213	27.50
	Basic	-0.85	120	4.34	2.71	10.0	120	169	50.45
	Below Basic						0	119	19.74

Table 3-5. Estimates of Standard Error of Cut Scores for NAEP—Writing Grade 12, Round 2, 2011

Table/ Group	Achievement Level	Median Theta	Median NAEP	Standard Error		MAD	Scaled Score		Percent of
				EmpSE	BootSE		Min	Max	
Table 1	Advanced	1.99	220	3.41	4.05	1.0	220	300	1.17
	Proficient	0.93	183	6.93	7.84	5.0	183	219	16.61
	Basic	-0.73	124	9.90	8.55	2.0	124	182	59.29
	Below Basic						0	123	22.93
Table 2	Advanced	1.85	215	5.91	5.85	7.0	215	300	1.99
	Proficient	0.54	169	5.79	6.40	6.0	169	214	28.55
	Basic	-1.10	111	9.48	7.86	4.0	111	168	55.39
	Below Basic						0	110	14.06
Table 3	Advanced	1.70	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	4.0	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	8.0	161	210	36.48
	Basic	-0.73	124	0.56	0.65	0.0	124	160	37.80
	Below Basic						0	123	22.67
Table 5	Advanced	1.82	214	3.52	2.92	3.0	214	300	2.17
	Proficient	0.48	167	4.40	4.90	3.0	167	213	30.96
	Basic	-0.97	116	3.37	3.26	4.0	116	166	50.08
	Below Basic						0	115	16.78
Table 6	Advanced	1.51	204	2.53	2.83	0.0	204	300	5.10
	Proficient	0.48	167	5.23	5.68	3.0	167	203	27.62
	Basic	-0.79	122	8.45	9.84	5.0	122	166	45.86
	Below Basic						0	121	21.42
Group A	Advanced	1.92	218	4.01	4.21	4.0	218	300	1.51
	Proficient	0.68	174	3.83	3.74	8.5	174	217	24.20
	Basic	-0.80	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	-0.80	122	3.03	3.15	3.5	122	166	46.12
	Below Basic						0	121	21.16
All	Advanced	1.77	213	6.90	3.39	8.5	213	300	2.59
	Proficient	0.49	167	5.14	1.81	7.5	167	212	30.14
	Basic	-0.80	122	4.19	2.44	5.5	122	166	46.33
	Below Basic						0	121	20.95

Table 3-6. Estimates of Standard Error of Cut Scores for NAEP—Writing Grade 12, Round 3, 2011

Table/ Group	Achievement Level	Median Theta	Median NAEP	Standard Error		MAD	Scaled Score		Percent of
				EmpSE	BootSE		Min	Max	
Table 1	Advanced	1.69	210	5.19	5.93	4.0	210	300	3.18
	Proficient	0.71	175	3.00	3.64	0.0	175	209	21.55
	Basic	-0.73	124	5.73	5.86	5.0	124	174	52.34
	Below Basic						0	123	22.93
Table 2	Advanced	1.60	206	10.52	9.62	13.0	206	300	4.19
	Proficient	0.41	165	10.17	8.04	5.0	165	205	31.29
	Basic	-0.80	122	6.06	6.60	0.0	122	164	43.57
	Below Basic						0	121	20.95
Table 3	Advanced	1.60	207	7.17	6.24	7.0	207	300	4.06
	Proficient	0.57	170	8.30	6.87	6.0	170	206	25.75
	Basic	-0.90	118	7.63	7.72	8.5	118	169	51.78
	Below Basic						0	117	18.41
Table 4	Advanced	1.68	209	1.50	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	2.0	210	300	3.13
	Proficient	0.47	167	6.88	6.95	10.0	167	209	30.01
	Basic	-0.93	117	9.84	10.96	3.0	117	166	49.14
	Below Basic						0	116	17.73
Table 6	Advanced	1.71	210	5.42	5.82	2.0	210	300	3.13
	Proficient	0.49	167	11.98	12.47	15.0	167	209	29.20
	Basic	-0.93	117	13.65	13.05	21.0	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	-0.80	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	2.0	210	300	3.18
	Proficient	0.56	170	4.29	4.17	8.0	170	209	27.06
	Basic	-0.86	120	4.30	3.94	6.5	120	169	50.34
	Below Basic						0	119	19.43
All	Advanced	1.69	210	6.08	1.06	4.0	210	300	3.24
	Proficient	0.64	173	5.99	3.52	7.0	173	209	23.59
	Basic	-0.80	122	4.35	2.74	7.0	122	172	52.22
	Below Basic						00	121	20.95

The cut scores and their standard errors were also examined across panelist subgroups. The results for grade 8 and grade 12 panelists are presented by panelist type (i.e., teachers, non-teacher educators, or members of the general public), gender, race, and region in Tables 3-7 and 3-8. Examining the cut scores for the grade 8 panelists by panelist type reveals that the teachers had the lowest cut scores for Basic and Proficient, and the general public had the highest cut scores for Basic and Proficient. For Advanced, teachers also had the lowest cut scores, but the non-teacher educators had the highest cut scores. For grade 12 panelists, the pattern is similar. The teachers had the lowest cut scores across all achievement levels. But for grade 12, the non-teacher educators had the highest cut scores. The grade 8 cut score SEs were smallest for teachers and largest for non-teacher educators. The grade 12 cut score SEs for Basic and Advanced were smallest for teachers, and the grade 12 cut score SEs for all levels were largest for the general public. However, the grade 12 SE for Proficient was remarkably small for non-teacher educators. Given the small number of nonteacher educators, these results should be interpreted with caution.

Group cut scores were also examined by gender and race. The cut scores for grade 8 female panelists were higher for Basic, lower for Proficient, and lower for Advanced. The opposite was true for grade 12 (lower for Basic, higher for Proficient, higher for Advanced). The standard errors of the cut scores were larger for the male panelists. These results should be interpreted with caution due to the fact that few of the panelists were male. Given the small number of non-White panelists, all non-White panelists were grouped for comparison to White panelists. For grade 8 panelists, the cut scores for non-white panelists were lower for all achievement levels. The same is true for grade 12, except for the Basic cut score, which was the same for non-White and White panelists. The SEs were larger for grade 8 non-White panelists and not computed for grade 12, because of sample size. Differences in cut scores and SEs for White vs. non-White panelists should be interpreted carefully due to the small number of non-White panelists.

Panel group cut scores from the four major geographic regions were also examined. Most of the cut scores were very similar across regions. Notable differences are discussed here. The grade 8 Basic cut scores for Midwest and Northeast panelists were generally lower than those for South and West panelists. The grade 12 Basic cut scores were similar across regions except for

the South, which had much lower average cut scores (103). The Proficient cut scores for grades 8 and 12 were lower for Northeast panelists. The Proficient grade 12 cut score was also lower for West panelists. The Advanced grade 8 cut score was also lower for Northeast panelists. The standard errors associated with the regional group cut scores were similar given the sample size of panelists from each region. Since panelists were not sampled evenly from each region, it would be difficult to substantiate any claims about true differences in cut scores due to the region of the panelist based on this data.

Table 3-7. Cut Scores and Standard Errors by Panelist Type – Grade 8

Subgroup	*N*	*Basic*			*Proficient*			*Advanced*		
		Cut	*EmpSE*	*BootSE*	*Cut*	*EmpSE*	*BootSE*	*Cut*	*EmpSE*	*BootSE*
Teachers	16	115	3.8	3.8	168	4.2	4.3	211	3.9	3.9
Non-Teacher Educators	5	124	7.6	6.9	176	8.6	6.8	219	6.3	6.7
General Public	6	127	5.5	5.1	181	5.4	4.8	216	5.7	5.7
Female	22	121	1.8	2.7	172	0.1	2.9	211		3.4
Male	5	108	11.9	11.7	179	13.6	14.2	216	9.6	9.7
White	23	121	3	3.3	176	2.8	3.4	213	2.5	3.5
Non-White	4	114	7.9	7.3	149	10.5	9.4	203	10	7.6
Midwest	6	115	9.1	8.1	175	9.5	8.4	217	5.1	5.1
Northeast	5	115	9.2	9.5	158	7.2	7.8	202	5.8	5.1
South	6	121	6.4	6	175	7.9	8.2	211	3.9	3.5
West	10	124	4.6	4.8	178	3.1	3.6	215	6.3	6.6

Table 3-8. Cut Scores and Standard Errors by Panelist Type – Grade 12

Subgroup	*N*	*Basic*			*Proficient*			*Advanced*		
		Cut	*EmpSE*	*BootSE*	*Cut*	*EmpSE*	*BootSE*	*Cut*	*EmpSE*	*BootSE*
Teachers	15	122	3.6	3.7	167	3.8	3.6	208	2	2.1
Non-Teacher Educators	5	123	4.9	4.8	175	1.1	1.1	210	2.8	2.2
General Public	8	120	5.5	5.7	171	5	5	210	4.3	4.5
Female	19	120	3.5	3.4	175	2.1	2.3	210	0.9	0.9
Male	9	122	5.4	6.5	165	3.9	4.2	200	4.9	4.7
White	25	122		3.1	174		3.5	210		1
Non-White	1	122			160			193		
Midwest	8	124	5.2	4.9	176	3.1	3.4	210	2.1	1.9
Northeast	4	117	11.8	12.8	166	6.7	6.8	211	5.4	5.9
South	6	103	5.8	5.5	175	12.3	12.6	210	6.7	6.1
West	10	123	3	2.7	167	2.9	2.4	207	3.7	3.9

3.2. Variability of Cut Scores

Another way of examining the change of the cut scores is to examine the variability of cut scores within and across rounds. Because panel cut scores were calculated by obtaining the median cut scores of the panel members, it is not appropriate to use a standard deviation calculation to describe variation of the cut scores within a panel. Instead, variation is described in this section in two ways: (a) mean absolute deviation (MAD) indices and (b) overall summaries of how cut scores changed between rounds

The MAD is the average difference between each panelist's cut score and the panel's median cut score, as shown in Equation 3.5,

$$\text{MAD} = \frac{\sum_{i=1}^{n} |x_i - x_{Mdn}|}{n}, \tag{3.5}$$

where x_i represents a panelist's cut score on the NAEP scale, x_{Mdn} is the panel's median cut score, and n is the number of panelists on the panel. Tables 3-9 and 3-10 report the MAD for each classification round for the grade 8 and grade 12 panels, respectively. 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.

Table 3-9. Mean Absolute Deviation (MAD) by Round—Writing Grade 8

Achievement Level	*MAD*		
	Round 1	*Round 2*	*Round 3*
Basic	12.0	9.0	9.0
Proficient	11.0	8.0	7.0
Advanced	12.0	5.0	9.0

Table 3-10. Mean Absolute Deviation (MAD) by Round—Writing Grade 12

Achievement Level	*MAD*		
	Round 1	*Round 2*	*Round 3*
Basic	10.0	5.5	7.0
Proficient	14.0	7.5	7.0
Advanced	7.5	8.5	4.0

A summary of the individual panelists' cut score changes between rounds provides additional information about how the cut scores varied within a panel. Table 3-11 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 Round 1 and Round 2 are labeled "R1:R2," while changes between Round 2 and Round 3 are labeled "R2:R3." Typically we would expect the number of changes to decrease across rounds. However, in our procedures, a new sample of BoWs was drawn for R3, meaning that panelists were classifying a completely new set of student work samples. So the increased number of changes in R2:R3 is to be expected. Overall the pattern of changes supports the internal validity argument because it matches the expectations about how the cut scores should change across rounds based on the theory of the BoW method and how it was implemented.

Table 3-11. Round-to-Round Cut Score Changes by Grade—2011

Grade	*Achievement Level*	*Round*	*Increased*		*No Change*		*Decreased*	
			N	*%*	*n*	*%*	*n*	*%*
8	Advanced	R1:R2	8	29.63	13	48.15	6	22.22
		R2:R3	8	29.63	1	3.70	18	66.67
	Proficient	R1:R2	12	44.44	4	14.81	11	40.74
		R2:R3	12	44.44	0	0.00	15	55.56
	Basic	R1:R2	12	44.44	4	14.81	11	40.74
		R2:R3	12	44.44	0	0.00	15	55.56
12	Advanced	R1:R2	10	35.71	6	21.43	12	42.86
		R2:R3	8	28.57	0	0.00	20	71.43
	Proficient	R1:R2	11	39.29	5	17.86	12	42.86
		R2:R3	14	50.00	0	0.00	14	50.00
	Basic	R1:R2	11	39.29	5	17.86	12	42.86
		R2:R3	11	39.29	3	10.71	14	50.00

3.3. Standard Error of Panelist Estimates

The standard error of panelists' cut score estimates obtained during a standard-setting meeting is operationally defined in terms of how consistent the cut scores are between groups when using the same standard-setting procedures, assessment, and achievement level descriptions (ALDs).The interpretation of this standard error is such that lower values indicate a more reliable cut score.

Within each panel, each of the two groups (A and B) produced a set of median cut scores. Therefore, there are only two observations—although not entirely independent—for each grade and achievement level. Equation 3.6 is used to calculate the standard error using two observations (Brennan, 2002):

$$\widehat{\sigma}_{\bar{X}} = \frac{|X_1 - X_2|}{2}. \quad (3.6)$$

Tables 3-12 and 3-13 present the standard error estimates for grades 8 and 12, respectively. Also included in the tables are the approximate 95% confidence intervals for each achievement level's mean cut score (using the normal distribution).

Table 3-12. Standard Error Estimates and Confidence Intervals for Mean Cut Scores by Achievement Level—Writing 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 3-13. Standard Error Estimates and Confidence Intervals for Mean Cut Scores by Achievement Level—Writing 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	177.4.0	167.60
Advanced	1	220	212	216.0	4.0	223.84	208.16
	2	174	167	170.5	3.5	177.36	163.64
	3	209	210	209.5	0.5	210.48	208.52

As noted earlier, the standard errors tend to decrease across rounds, with a few exceptions. These include the case where the standard error fluctuates in Round 2 but ends at the same value as it was in Round 1 (Table 3-12 Grade 8 Proficient, Table 3-13 Grade 12 Basic) and the case where the standard error increases (Table 3-12 Grade 8, Basic). Exceptions can also be noted when examining the MADs, where it is sometimes the case that the MAD does not strictly decrease across rounds, but where the Round 3 MAD is smaller than the Round 1 MAD (Table 3-9 Grade 8 Advanced, Table 3-10 Grade 12 Basic, and Table 3-10 Grade 12 Advanced). A decrease in the standard error is desirable because the standard error is inversely related to the reliability of the cut score. So, smaller values for the standard error indicate or higher reliability. The size of the standard error can be evaluated relative to the size of the scale. The reported data in tables 3-12 and 3-13 are small relative to the span of the score scale. This suggests that the standard errors are small, which implies better reliability.

Chapter 4—SPECIAL STUDY ANALYSIS

This chapter describes the technical aspects of a special study implemented immediately following the ALS meeting. This study was conducted to provide the Governing Board information for exploring 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 assessment of writing. To carry out the special study, panelists engaged in a separate classification round to categorize a sample of 50 bodies of work (BoWs) from 2007 using 2011 achievement level descriptions (ALDs).

For the 2007 NAEP writing assessment in each of grades 8 and 12, there were a total of 40 forms and 20 unique writing tasks. The methods of form selection and BoW selection used for the special study were similar to those used for the ALS meeting with the exception that panelists were not assigned to separate groups in each grade level. Ten of the 40 available forms for the 2007 NAEP were selected such that all 20 prompts were represented. Five BoWs were sampled using the stratified sampling method described earlier in this report. Procedures implemented for carrying out this study are further described in the Process Report.

A series of cross-tabular analyses was conducted to understand student performance on the 2007 assessment using the 2007 achievement levels in comparison to performance on the 2007 assessment using the 2011 ALDs. The goal was to compare the achievement level classification of booklets for the 2007 assessment, based on the cut scores established in 1998, to the panelists' classification of booklets for the 2007 assessment, based on their understanding of the 2011 achievement levels descriptions. Because student performance is not actually assigned to an achievement level classification, the comparisons were based on classifications of the first plausible value, as well as the classification that would have resulted had a single *expected a posteriori* (EAP) score been assigned. In addition, the classifications based on the 2011 ALDs were examined using the individual panelist's classifications, as well as the classifications that resulted when cut scores were calculated using logistic regression. The use of logistic regression allowed for all BoWs—not just those selected for the study—to be classified. In addition to tables that include the results based on the 50 selected BoWs, tables were generated that include all BoWs in the entire assessment. These cross-tabular comparisons appear in Tables 4-1 through 4-6 for grade 8 and in Tables 4-7 through 4-12 for grade 12.

Table 4-1. Correspondence Between 2007 Achievement Levels Based on EAP and Panelists' Classifications Based on 2011 ALDs—Writing Grade 8

Actual 2007 Classifications (EAP Estimates)	*Special Study Panelists' Classifications (2011 ALDs)*			
	Below Basic (n=272)	*Basic (n=247)*	*Proficient (n=183)*	*Advanced (n=98)*
Below Basic (n=192)	0.89	0.11	0.00	0.00
Basic (n=288)	0.34	0.58	0.08	0.00
Proficient (n=304)	0.01	0.19	0.52	0.27
Advanced (n=16)	0.00	0.00	0.06	0.94

Table 4-2. Correspondence Between 2007 Achievement Levels Based on Plausible Values and Panelists' Classifications Based on 2011 ALDs—Writing Grade 8

Actual 2007 Classifications (Plausible Values)	*Special Study Panelists' Classifications (2011 ALDs)*			
	Below Basic (n=272)	*Basic (n=247)*	*Proficient (n=183)*	*Advanced (n=98)*
Below Basic (n=176)	0.85	0.14	0.02	0.00
Basic (n=368)	0.33	0.45	0.19	0.03
Proficient (n=208)	0.01	0.26	0.47	0.26
Advanced (n=48)	0.00	0.02	0.29	0.69

Table 4-3. Correspondence Between 2007 Achievement Levels Based on EAP and Panelists' Classifications Based on 2011 ALDs After Logistic Regression (50 BoWs)—Writing Grade 8

Actual 2007 Classifications (EAP Estimates)	*Special Study Panelists' Classifications (2011 ALDs): After Logistic Regression*			
	Below Basic (n=15)	*Basic (n=17)*	*Proficient (n=11)*	*Advanced (n=7)*
Below Basic (n=12)	1.00	0.00	0.00	0.00
Basic (n=18)	0.17	0.83	0.00	0.00
Proficient (n=19)	0.00	0.11	0.58	0.32
Advanced (n=1)	0.00	0.00	0.00	1.00

Table 4-4. Correspondence Between 2007 Achievement Levels Based on Plausible Values and Panelists' Classifications Based on 2011 ALDs After Logistic Regression (50 BoWs)—Writing Grade 8

Actual 2007 Classifications (Plausible Values)	*Special Study Panelists' Classifications (2011 ALDs): After Logistic Regression*			
	Below Basic (n=17)	*Basic (n=20)*	*Proficient (n=7)*	*Advanced (n=6)*
Below Basic (n=11)	1.00	0.00	0.00	0.00
Basic (n=23)	0.26	0.74	0.00	0.00
Proficient (n=13)	0.00	0.23	0.54	0.23
Advanced (n=3)	0.00	0.00	0.00	1.00

Table 4-5. Correspondence Between 2007 Achievement Levels Based on EAP and Panelists' Classifications Based on 2011 ALDs After Logistic Regression (All BoWs)—Writing Grade 8

Actual 2007 Classifications (EAP Estimates)	*Special Study Panelists' Classifications (2011 ALDs): After Logistic Regression*			
	Below Basic (n=18,372)	*Basic (n=86,370)*	*Proficient (n=27,796)*	*Advanced (n=2,735)*
Below Basic (n=17,735)	1.00	0.00	0.00	0.00
Basic (n=80,009)	0.07	0.93	0.00	0.00
Proficient (n=39,972)	0.00	0.17	0.79	0.05
Advanced (n=2,194)	0.00	0.00	0.00	1.00

Table 4-6. Correspondence Between 2007 Achievement Levels Based on Plausible Values and Panelists' Classifications Based on 2011 ALDs After Logistic Regression (All BoWs)—Writing Grade 8

Actual 2007 Classifications (Plausible Values)	*Special Study Panelists' Classifications (2011 ALDs): After Logistic Regression*			
	Below Basic (n=29,963)	*Basic (n=77,614)*	*Proficient (n=27,761)*	*Advanced (n=4,572)*
Below Basic (n=11,904)	1.00	0.00	0.00	0.00
Basic (n=86,921)	0.15	0.85	0.00	0.00
Proficient (n=35,311)	0.00	0.25	0.69	0.06
Advanced (n=1,137)	0.00	0.00	0.00	1.00

Table 4-7. Correspondence Between 2007 Achievement Levels Based on EAP and Panelists' Classifications Based on 2011 ALDs—Writing Grade 12

Actual 2007 Classifications (EAP Estimates)	*Special Study Panelists' Classifications (2011 ALDs)*			
	Below Basic (n=300)	*Basic (n=298)*	*Proficient (n=216)*	*Advanced (n=86)*
Below Basic (n=270)	0.82	0.18	0.00	0.00
Basic (n=324)	0.24	0.61	0.14	0.01
Proficient (n=306)	0.00	0.17	0.56	0.27
Advanced*	NA	NA	NA	NA

* Of the 2007 NAEP writing forms selected for the Special Study, none of the BoWs have EAP scores at the Advanced range of the 2007 NAEP writing scale.

Table 4-8. Correspondence Between 2007 Achievement Levels Based on Plausible Values and Panelists' Classifications Based on 2011 ALDs—Writing Grade 12

Actual 2007 Classifications (Plausible Values)	*Special Study Panelists' Classifications (2011 ALDs)*			
	Below Basic (n=300)	*Basic (n=298)*	*Proficient (n=216)*	*Advanced (n=86)*
Below Basic (n=252)	0.89	0.11	0.00	0.00
Basic (n=360)	0.21	0.60	0.18	0.01
Proficient (n=288)	0.00	0.19	0.53	0.28
Advanced*	NA	NA	NA	NA

* Of the 2007 NAEP writing forms selected for the Special Study, none of the BoWs have EAP scores at the Advanced range of the 2007 NAEP writing scale.

Table 4-9. Correspondence Between 2007 Achievement Levels Based on EAP and Panelists' Classifications Based on 2011 ALDs After Logistic Regression (50 BoWs)—Writing Grade 12

Actual 2007 Classifications (EAP Estimates)	*Special Study Panelists' Classifications (2011 ALDs): After Logistic Regression*			
	Below Basic (n=20)	*Basic (n=12)*	*Proficient (n=16)*	*Advanced (n=2)*
Below Basic (n=15)	1.00	0.00	0.00	0.00
Basic (n=18)	0.28	0.67	0.06	0.00
Proficient (n=17)	0.00	0.00	0.88	0.12
Advanced*	NA	NA	NA	NA

* Of the 2007 NAEP writing forms selected for the Special Study, none of the BoWs have EAP scores at the Advanced range of the 2007 NAEP writing scale.

Table 4-10. Correspondence Between 2007 Achievement Levels Based on Plausible Values and Panelists' Classifications Based on 2011 ALDs After Logistic Regression (50 BoWs)—Writing Grade 12

Actual 2007 Classifications (Plausible Values)	*Special Study Panelists' Classifications (2011 ALDs): After Logistic Regression*			
	Below Basic (n=16)	*Basic (n=15)*	*Proficient (n=17)*	*Advanced (n=2)*
Below Basic (n=14)	1.00	0.00	0.00	0.00
Basic (n=20)	0.10	0.75	0.15	0.00
Proficient (n=16)	0.00	0.00	0.88	0.13
Advanced*	NA	NA	NA	NA

* Of the 2007 NAEP writing forms selected for the Special Study, none of the BoWs have EAP scores at the Advanced range of the 2007 NAEP writing scale.

Table 4-11. Correspondence Between 2007 Achievement Levels Based on EAP and Panelists' Classifications Based on 2011 ALDs After Logistic Regression (All BoWs)—Writing Grade 12

Actual 2007 Classifications (EAP Estimates)	*Special Study Panelists' Classifications (2011 ALDs): After Logistic Regression*			
	Below Basic (n=5,722)	*Basic (n=15,081)*	*Proficient (n=5,940)*	*Advanced (n=343)*
Below Basic (n=3,832)	1.00	0.00	0.00	0.00
Basic (n=18,078)	0.10	0.83	0.06	0.00
Proficient (n=5,108)	0.00	0.00	0.95	0.05
Advanced (n=68)	0.00	0.00	0.00	1.00

Table 4-12. Correspondence Between 2007 Achievement Levels Based on Plausible Values and Panelists' Classifications Based on 2011 ALDs After Logistic Regression (All BoWs)—Writing Grade 12

Actual 2007 Classifications (Plausible Values)	*Special Study Panelists' Classifications (2011 ALDs): After Logistic Regression*			
	Below Basic (n=7,079)	*Basic (n=12,789)*	*Proficient (n=7,211)*	*Advanced (n=782)*
Below Basic (n=5,025)	1.00	0.00	0.00	0.00
Basic (n=16,016)	0.13	0.80	0.07	0.00
Proficient (n=6,558)	0.00	0.00	0.92	0.08
Advanced (n=262)	0.00	0.00	0.00	1.00

There is a high level of correspondence between the two classifications for the 2007 and 2011 NAEP writing assessments in both grades 8 and 12, as indicated in Tables 4-1 through 4-12. The level of correspondence between actual classifications of BoWs and the classifications provided by the special study panelists indicates that performance relative to the two sets of ALDs was approximately the same. This implies that the operationalization of policy definitions for Basic, Proficient, and Advanced for the 2011 NAEP writing assessment was fairly similar to that for the 1998 assessment.

The special study was first conducted after the pilot study. Findings from the pilot study warranted some necessary changes to the ALDs. The changes to the ALDs rendered the results of the original special study moot. Thus, the special study was implemented again after the operational ALS meeting. The special study, which explored the relationship between performance on the 2011 assessment, based on the new writing framework, and performance on the 2007 assessment, provided evidence of a relationship between the performance of students on the two assessments.

REFERENCES

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

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.

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

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

Hambleton, R. K., & Pitoniak, M. 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.

Kahl, S. R., Crockett, T. J., DePascale, C. A., & Rindfleisch, S. L. (1995, April). *Setting standards for performance levels using the student-based constructed-response method.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

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

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

Measured Progress (2011). *Developing achievement levels on the National Assessment of Educational Progress for writing grades 8 and 12 in 2011 and grade 4 in 2013: Design document*. Dover, NH: Author.

National Assessment Governing Board (2010). *Writing framework for the 2011 National Assessment of Educational Progress*. Washington, DC: Author.

APPENDICES