National Assessment Governing Board

National Assessment of Educational Progress

Judgmental Standard Setting (JSS)

Final Submitted: March 15, 2012

Redacted by the Governing Board to protect the confidentiality of study participants and NAEP assessment items.

TECHNICAL REPORT

Submitted to:
Dr. Susan Loomis
National Assessment Governing Board
800 North Capitol Street, NW, Suite 825
Washington, DC 20002-4233
202.357.6940
Email: Susan.Loomis@ed.gov

This study was funded by the
National Assessment Governing Board
under Contract ED-NAG-10-C-0004.

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

In collaboration with:
WestEd
730 Harrison Street
San Francisco, CA 94107
415.615.3400

measured progress

Measured Progress collaborated with WestEd in the implementation of this study and writing of this report.

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

TABLE OF CONTENTS

CHAPTER 1— INTRODUCTION 1
CHAPTER 2— MATERIALS AND PROCEDURES 4
2.1. Division of Panelists into Rating Groups 4
2.2. Description of Item Rating Pools 4
2.2.1. Division of Item Pool 6
2.2.2. Test Form Administered to Panelists 7
2.3. Ordered Item Book, Constructed Response Ordered Item Book, and Item Map 8
2.3.1. Item Identification Number 11
2.3.2. Computation of Item Scale Values 11
2.3.3. Item Map Values 14
2.4. Setting Bookmarks 14
2.5. Post-Round Feedback 15
2.5.1. Cut Score Results 15
2.5.2. Rater Location Chart 15
2.5.3. Whole Booklet Feedback 16
2.5.4. Consequences Feedback 22
2.6. Consequences Questionnaire 22
2.7. Selecting Potential Exemplar Items 23
2.8. Process Evaluations 24
CHAPTER 3— CUT SCORE EVALUATION 25
3.1. Variability of Cut Scores 25
3.2. Estimates of Standard Errors of Cut Scores 27
3.3. Reliability Analyses 30
CHAPTER 4— SPECIAL ANALYSES 34
4.1. Facilitator Effect Study 34
4.2. Irrelevant Items 36
REFERENCES 37
APPENDICES 38
APPENDIX A CAB DOCUMENTATION
APPENDIX B JSS-TAC SESSION SUMMARIES
APPENDIX C ITEM INFORMATION
APPENDIX D ITEM MAPS
APPENDIX E PILOT STUDY FEEDBACK
APPENDIX F FREQUENCY DISTRIBUTION OF STUDENT PERFORMANCE

LIST OF TABLES

Table 1-1. Operational Workshop Design 2
Table 2-1. Summary of Item Pool by Block—Mathematics 5
Table 2-2. Summary of Item Pool by Block—Reading 6
Table 2-3. Summary of Panel Item Pools—Mathematics 6
Table 2-4. Summary of Panel Item Pools—Reading 7
Table 2-5. Summary of Test Form Administered to Panelists—Mathematics 7
Table 2-6. Summary of Test Form Administered to Panelists—Reading 7
Table 2-7. CROIB Panel A and B Contents—Mathematics 11
Table 2-8. CROIB Panel A and B Contents—Reading 11
Table 2-9. Item Identification Numbers, Scale Values, and Map Values for Easiest and Hardest Items within Item Type—Mathematics Panel A Item Pool 11
Table 2-10. Marginal Content Area Theta Means and Standard Deviations—Mathematics 13
Table 2-11. Marginal Content Area Theta Means and Standard Deviations—Reading 13
Table 2-12. Forms Selected for Whole Booklet Feedback and Summary of an Average Block, by Subject 17
Table 2-13. Summary of Exemplar Items Selection 23
Table 3-1. Mean Absolute Deviation (MAD) by Panel and Round—Mathematics Grade 12 25
Table 3-2. Mean Absolute Deviation (MAD) by Panel and Round—Reading Grade 12 26
Table 3-3. Round to Round Cut Score Changes by Panel—Mathematics Grade 12 26
Table 3-4. Round to Round Cut Score Changes by Panel—Reading Grade 12 27
Table 3-5. Estimates of Standard Error of Cut Scores for NAEP—Mathematics Grade 12 29
Table 3-6. Estimates of Standard Error of Cut Scores for NAEP—Reading Grade 12 30
Table 3-7. Standard Error Estimates and Confidence Intervals for Mean Cut Scores by Post-Secondary Activity—Mathematics 31
Table 3-8. Standard Error Estimates and Confidence Intervals for Mean Cut Scores by Post-Secondary Activity—Reading 32
Table 4-1. *p*-Values from Tests for Facilitator Effects—Mathematics 35
Table 4-2. *p*-Values from Tests for Facilitator Effects—Reading 36

LIST OF FIGURES

Figure 2-1. Virtual OIB (Item List View) 8
Figure 2-2. Virtual CROIB (Item List View) 9
Figure 2-3. Virtual OIB/CROIB (Item Information View) 9
Figure 2-4. Rater Location Chart (Example) 16
Figure 2-5. Example Booklet Score Chart—JSS Operational Session 1 20
Figure 2-6. Example Booklet Score Chart—JSS Operational Sessions 2 and 3 21
Figure 2-7. Cut Scores and Data Consequences Feedback 22
Figure 3-1. Mean Cut Scores and Confidence Intervals by Post-Secondary Activity—Mathematics 32
Figure 3-2. Mean Cut Scores and Confidence Intervals by Post-Secondary Activity—Reading 33

Chapter 1—INTRODUCTION

For over two decades, the National Assessment Governing Board has guided the development and use of the National Assessment of Educational Progress (NAEP) in monitoring the progress of student achievement in the nation across time and content areas. In 2004, the Governing Board began to explore the utility of the NAEP as a tool to predict students' academic preparedness for entry into post-secondary education or job-training programs, forming a Technical Panel on 12th Grade Preparedness Research that was tasked with assisting the Governing Board in planning relevant research and validity studies (National Assessment Governing Board, 2009). The Technical Panel recommended a multi-method approach to exploring the feasibility of reporting post-secondary preparedness on the 2009 Grade 12 NAEP scale for mathematics and reading.

One of the four methodologies proposed included a series of criterion-based judgmental standard-setting (JSS) studies to identify reference points on the NAEP scale that indicate academic preparedness for placement in credit-bearing, entry-level courses of the sort that fulfill general education requirements or eligibility for entry to job-training programs in specified occupations. The JSS study's Process Report (WestEd & Measured Progress, 2011) further describes the JSS studies.

The primary objective of the JSS studies was to obtain cut scores on the NAEP scale that represent academic preparedness for entry into credit-bearing college courses or job-training programs selected by the Governing Board. The Governing Board selected for inclusion in this project the following six post-secondary activities: 1) college, 2) automotive master technician, 3) licensed practical nurse, 4) pharmacy technician, 5) computer support specialist, and 6) heating, ventilation, and air conditioning technician. In order to maximize standardization of the JSS process across the post-secondary activities, the Governing Board developed a Design Document (National Assessment Governing Board, 2010) to guide all aspects of the project's implementation.

The Design Document stipulated the use of a modified bookmark methodology, including the use of whole booklet feedback. A pilot study to evaluate the methodology, materials, and logistics was also mandated. Further, the JSS studies used a replicate panel design in which two replicate panels were convened for each post-secondary activity within a content area to aid in evaluating the reliability of the results. A total of four sessions were held in 2011. The first session was the pilot study that included the college and automotive master technician panels. The other three were operational sessions and the pairings are indicated in Table 1-1. This workshop design ensured that all process facilitators had facilitated a pilot workshop prior to the operational session.

Table 1-1. Operational Workshop Design

Operational Session Number	*Workshop*	*Content Area*	*Content Facilitators*	*Panel A*	*Panel B*
1	College-Preparedness	Reading	Content Facilitator 1	Process Facilitator 1	Process Facilitator 2
		Mathematics	Content Facilitator 2	Process Facilitator 3	Process Facilitator 4
	Automotive Master Technician	Reading	Content Facilitator 3	Process Facilitator 5	Process Facilitator 6
		Mathematics	Content Facilitator 4	Process Facilitator 7	Process Facilitator 8
2	Licensed Practical Nurse (LPN)	Reading	Content Facilitator 1	Process Facilitator 1	Process Facilitator 2
		Mathematics	Content Facilitator 2	Process Facilitator 3	Process Facilitator 4
	Pharmacy Technician	Reading	Content Facilitator 3	Process Facilitator 5	Process Facilitator 6
		Mathematics	Content Facilitator 4	Process Facilitator 7	Process Facilitator 8
3	Computer Support Specialist	Reading	Content Facilitator 1	Process Facilitator 1	Process Facilitator 2
		Mathematics	Content Facilitator 2	Process Facilitator 3	Process Facilitator 4
	Heating, Ventilation, and Air Conditioning Technician (HVAC)	Reading	Content Facilitator 3	Process Facilitator 5	Process Facilitator 6
		Mathematics	Content Facilitator 4	Process Facilitator 7	Process Facilitator 8

In addition, the standard setting process followed the recommendation of the Design Document in using computer software where possible to increase the efficiency and effectiveness of the process. The Design Document specifically identified the use of computers for capturing panelist annotations of knowledge, skills, and abilities (KSA) required to correctly respond to each item or to score at a specific level on constructed response items. Measured Progress developed the Computer-Aided Bookmarking (CAB) software in response to this request. The following key activities were computerized with the development of CAB:

1. KSA annotations
2. Presentation of the Ordered Item Books
3. Bookmark placements
4. Provision of feedback
5. Process evaluation responses
6. Selection of exemplar items

CAB is referred to throughout this report, and its documentation can be found in Appendix A.

This Technical Report serves as a supplement to the Process Report (WestEd & Measured Progress, 2011) by providing a description of the technical procedures implemented before, during, and after the JSS sessions. The technical procedures implemented were guided by the advice of a JSS Technical Advisory

Committee (JSS-TAC)—a five-member group that collectively represents expertise in standard setting, vocational/post-secondary activity education and certification, and experience with the NAEP—and the Contracting Officer's Representative, Dr. Susan Loomis. Reports of the JSS-TAC meetings are included in Appendix B. These reports provide information about key technical decisions made to guide implementation of the process and reporting of results.

This document is divided into three main sections: materials and procedures, cut score evaluation, and special analyses.

1. *Materials and Procedures:* This section describes technical procedures implemented and materials given to the panelists during the JSS sessions. Technical procedures include the division of panelists into rating groups, division of items into rating pools, creation of Ordered Item Books and item maps, setting of bookmarks, presentation of post-round feedback, and selection of potential exemplar items. Materials provided to panelists that are displayed include both those presented using CAB and those presented on paper.
2. *Cut Score Evaluation:* This section describes how the cut scores resulting from the JSS sessions were evaluated including variability, standard error, and reliability analyses.
3. *Special Analyses:* This section contains a special analysis conducted to explore the possibility of facilitator effects across sessions and a special study that looked at the effects of irrelevant items on cut scores.

Chapter 2—MATERIALS AND PROCEDURES

This chapter describes the materials provided to panelists during the judgmental standard-setting (JSS) sessions and the technical procedures implemented before, during, and after the sessions. Eight sub-sections have been developed, including a description of the division of panelists into rating groups, division of items into item rating pools, creation of Ordered Item Books (OIB) and item maps, placement of bookmarks, post-round feedback, a consequences questionnaire, selection of potential exemplar items, and process evaluations.

2.1. Division of Panelists into Rating Groups

For each operational JSS session and subject (mathematics or reading), approximately 20 panelists were convened. Each panel consisted of nine or ten panelists (four to six were present for the pilot study), except for in the operational automotive master technician panels, for which there were seven or eight panelists in each replicate mathematics panel and five or six panelists in each replicate reading panel. Each replicate panel was further divided into table groups of four or five panelists each for individual work and to facilitate group discussion. The demographic attributes (i.e., educator role, gender, geographic region, and race/ethnicity) of panelists were considered when assigning members to replicate panels and table groups to maximize their equivalence. In the Process Report (WestEd & Measured Progress, 2011), panelist characteristics are described in more detail.

2.2. Description of Item Rating Pools

One of the table group tasks included the development of descriptions of the knowledge, skills, and abilities (KSA) required for responding correctly or receiving the specified number of points for each of the score points in an item pool. To reduce the cognitive load that this task demands, panelists were asked to record KSAs for only a subset of items. Therefore, panelists were assigned one of two subsets of items from their pools. Panelists recorded KSAs for the rest of the items during table group discussions with other panelists assigned to a different group of items. Considerations made when splitting panel item pools in half were similar to those used when assigning items to panel item pools (i.e., item difficulty, type, and content-area representation).

The JSS sessions used items, item statistics, and student performance data from the 2009 NAEP Grade 12 mathematics and reading assessments. Tables 2-1 and 2-2 present summaries of the scored items used in the JSS sessions for the mathematics and reading assessments, respectively. The NAEP assessment items are organized into blocks (12 for mathematics and 13 for reading). These blocks are labeled “MA” through “ML” for mathematics and “RA” through “RM” for reading. For mathematics, there was a total of 164 items, of which 13 or 14 appeared in each block; for reading, there was a total of 131 items, of which 9 to 11 appeared in each block.

For mathematics, 107 items were multiple choice, 15 were dichotomously-scored constructed response, and 41 were polytomously-scored constructed response items. The polytomously-scored items represented a total of 103 score points, or 46% of the points in the grade 12 item pool. Dichotomously-scored items represented 7% of the points, and multiple choice items represented 48% of the points. The total number of points was 226. Table 2-1 shows how the items were distributed by content area and item type.

Table 2-1. Summary of Item Pool by Block—Mathematics

Block	*Total Number of Items*[A]	*Content Area*[B]				*Item Type*[C]			*PCR Points*
		NPO	*M&G*	*DAP*	*ALG*	*MC*	*DCR*	*PCR*	
1	14	2	3	4	5	9	1	4	10
2	14	1	4	3	6	9	0	5	12
3	14	2	5	2	5	9	2	3	8
4	14	2	4	4	4	8	1	5	10
5	13	2	4	3	4	9	1	3	8
6	14	1	4	4	5	10	1	3	8
7	13	1	5	3	4	8	1	4	10
8	13	1	5	4	3	8	1	4	10
9	13	1	4	4	4	10	1	2	6
10	13	1	4	3	5	8	1	4	10
11	14	2	4	3	5	9	3	2	6
12	14	3	4	2	5	10	2	2	5
TOTAL	163	19	50	39	55	107	15	41	103
Actual %		12	31	24	34				46
Target %		10	30	25	35				

[A] Total number of items with item statistics
[B] NPO = Number Properties and Operations; M&G = Measurement and Geometry; DAP = Data Analysis and Probability; ALG = Algebra
[C] MC = Multiple choice; DCR = Dichotomously-scored Constructed Response; PCR = Polytomously-scored Constructed Response

For reading, 76 items were multiple choice, 10 were dichotomously-scored constructed response, and 45 were polytomously-scored constructed response. The polytomously-scored items represented a total of 103 score points, or 54% of the points in the item pool. Dichotomously-scored items represented 5% of the points, and multiple choice items represented 40% of the points. The total number of points was 189. Table 2-2 shows how the items were distributed by content area and item type.

Table 2-2. Summary of Item Pool by Block—Reading

Block	Total Number of Items	Content Area[A]		Item Type[B]			PCR Points
		LIT	INF	MC	DCR	PCR	
1	10	10	-	5	1	4	9
2	10	10	-	6	1	3	7
3	10	-	10	7	2	1	3
4	10	-	10	6	1	3	7
5	9	9	-	4	1	4	9
6	11	11	-	5	2	4	9
7	11	-	11	7	1	3	7
8	10	-	10	6	0	4	9
9	9	-	9	5	0	4	9
10	10	-	10	6	1	3	7
11	10	-	10	6	0	4	9
12	11	-	11	7	0	4	9
13	10	-	10	6	0	4	9
TOTAL	131	40	91	76	10	45	103
Actual %		31	69				54
Target %		30	70				

[A] LIT = Literary; INF = Informational
[B] MC = Multiple Choice; DCR = Dichotomously-scored Constructed Response; PCR = Polytomously-scored Constructed Response

To the extent possible, each item pool was divided into two statistically equivalent item sets for use in the JSS process. The next two sections describe how the item pools were divided.

2.2.1. Division of Item Pool

The item pools were divided into equivalent and overlapping sets (A and B). Items included in both sets are referred to as "common" items. Equivalence was monitored with regard to (a) content area subscale representation, (b) item type representation, and (c) item difficulty. This division also created a design that allowed for the reliability of the process to be evaluated (see Reliability Analysis section). Tables 2-3 and 2-4 present a summary of the item pools by panel and overall for mathematics and reading, respectively.

Table 2-3. Summary of Panel Item Pools—Mathematics

Panel	Number of Items	Percent							Item Difficulty				
		by Subscale[A]				by Item Type[B]							
		NPO	M&G	DAP	ALG	MC	DCR	PCR	Points	Mean	SD[C]	Min	Max
A	81	11	32	24	33	65	10	25	113	195.84	38.70	88.00	300.00
B	82	12	29	27	32	67	9	24	111	194.30	40.29	71.00	300.00
Pool[D]	163	12	31	24	34	66	9	25	225	196.01	39.18	71.00	300.00

[A] NPO = Number Properties and Operations; M&G = Measurement and Geometry; DAP = Data Analysis and Probability; ALG = Algebra
[B] MC=Multiple Choice; DCR=Dichotomously-scored Constructed Response; PCR=Polytomously-scored Constructed Response
[C] SD=Standard Deviation
[D] The pool includes released items, which were excluded from the panel Ordered Item Books

Table 2-4. Summary of Panel Item Pools—Reading

Panel	Number of Items	Percent by Subscale[A]		Percent by Item Type[B]			Item Difficulty				
		LIT	INF	MC	DCR	PCR	Points	Mean	SD[C]	Min	Max
A	69	28	72	59	10	30	96	300.49	52.47	172.00	444.00
B	71	28	72	61	8	31	100	291.91	51.03	158.00	444.00
Pool[D]	131	31	69	58	9	33	186	297.04	50.77	158.00	444.00

[A] LIT=Literary; INF=Informational
[B] MC=Multiple Choice; DCR=Dichotomously-scored Constructed Response; PCR=Polytomously-scored Constructed Response
[C] SD=Standard Deviation
[D] The pool includes released items, which were excluded from the panel Ordered Item Books

2.2.2. Test Form Administered to Panelists

One of the first tasks panelists performed as part of their training was to take a form of the assessment. The assessment form selected for this purpose included released blocks (two blocks each for mathematics and reading). These item blocks were not included in the item pools for which panelists placed their bookmarks. Tables 2-5 and 2-6 present summary information about the test forms that were administered to mathematics and reading panelists, respectively.

Table 2-5. Summary of Test Form Administered to Panelists—Mathematics

Block	Total Number of Items[A]	Content Area[B]				Item Type[C]			PCR Points
		NPO	M&G	DAP	ALG	MC	DCR	PCR	
1	14	1	4	3	6	9	0	5	12
2	13	1	5	3	4	8	1	4	10
Total	27	2	9	6	10	17	1	9	22
Percent[D]		7%	33%	22%	37%	63%	4%	33%	

[A] Total number of items with item statistics
[B] NPO = Number Properties and Operations; M&G = Measurement and Geometry; DAP = Data Analysis and Probability; ALG = Algebra
[C] MC = Multiple Choice; DCR = Dichotomously-scored Constructed Response; PCR = Polytomously-scored Constructed Response
[D] Percents may not sum to 100 due to rounding error

Table 2-6. Summary of Test Form Administered to Panelists—Reading

Block	Total Number of Items	Content Area[A]		Item Type[B]			PCR Points
		LIT	INF	MC	DCR	PCR	
1	11	11	-	5	2	4	9
2	10	-	10	6	0	4	9
Total	21	11	10	11	2	8	18
Percent				52%	10%	38%	

[A] LIT = Literary; INF = Informational
[B] MC = Multiple Choice; DCR = Dichotomously-scored Constructed Response; PCR = Polytomously-scored Constructed Response

2.3. Ordered Item Book, Constructed Response Ordered Item Book, and Item Map

NAEP assessment items included in the JSS studies were organized and presented to panelists in various ways: 1) in an Ordered Item Book (OIBs), 2) in a Constructed Response Ordered Item Book (CROIB), and 3) in an item map. First, the OIB and CROIB are described together. Second, the item map is explained.

The OIB and CROIB are item review tools, both within the Computer-Aided Bookmarking (CAB) software and in paper books. Each mode of item presentation was tailored to include only the items from the NAEP item pool assigned to the replicate panel (Panel A or Panel B). The OIB contained all items, including the constructed-response and multiple-choice items assigned to the panel, while the CROIB only contained constructed-response items. Constructed-response items within the CROIB were organized differently from the OIB. All information about a single constructed-response item was contained together within the CROIB, with items organized by difficulty. Within the OIB, items (and score points) were presented in order of their scale values, from easiest to hardest. The order of items in the OIB and the difficulty of each item on the scale are shown in Appendix C. Items are identified in this appendix by item identification number, scale value, and map value, which are described later. Figures 2-1 through 2-3 are example displays of the virtual OIB and CROIB within CAB. Figures 2-1 and 2-2 display the item lists for the virtual OIB and CROIB, respectively, while Figure 2-3 presents how an item and its information would be displayed within both the virtual OIB and the CROIB.

Figure 2-1. Virtual OIB (Item List View)

Partially Redacted (Item ID column only)

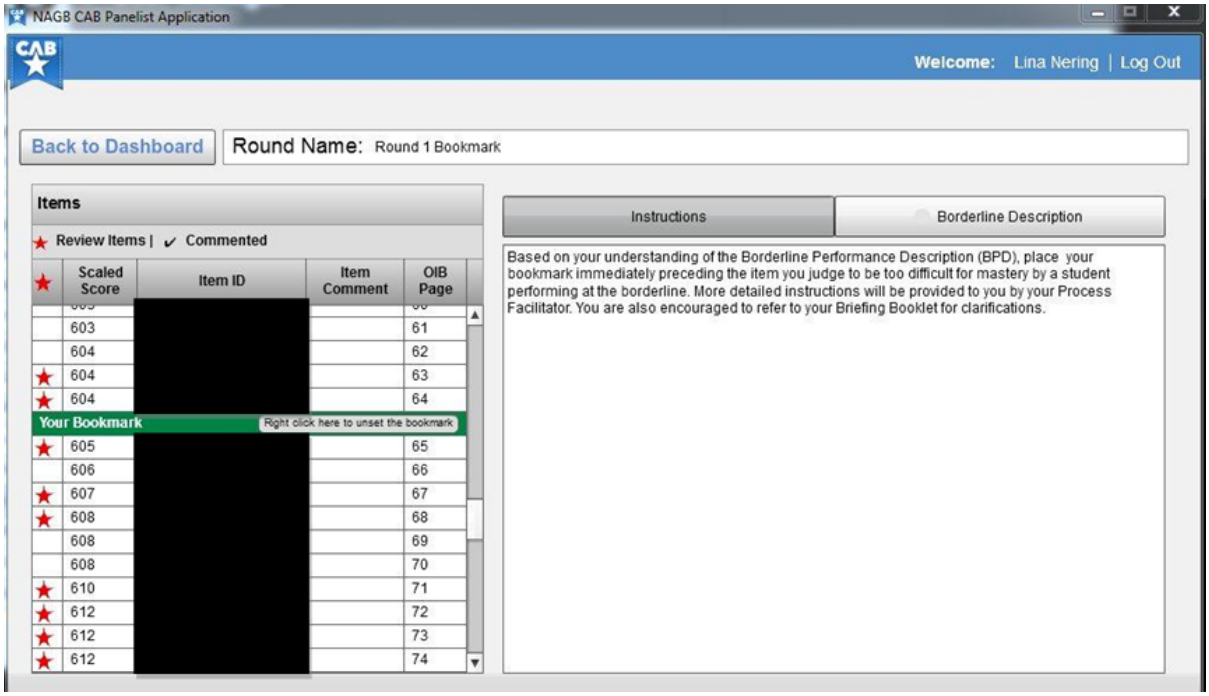


Figure 2-2. Virtual CROIB (Item List View)

Redacted

Figure 2-3. Virtual OIB/CROIB (Item Information View)

Redacted

The paper version of the OIB and CROIB contained the item and an item information box with the item identification number (Item ID), scale (map) value, block, position, domain, and accession number (ACCNUM). Additionally, the paper CROIB contained the following: 1) the page number where the item could be found in the OIB, 2) the scoring rubric, and 3) examples of student responses at each score level, including zero, presented in descending order (i.e., an exemplar for the highest score point was presented first, with the last student response being representative of a score of zero).

During the pilot, the constructed-response items within the CROIB were ordered by item difficulty of the full-credit response, from easiest to most difficult for both the mathematics and reading panels. However, a change resulting from the pilot study was the ordering of the items in the CROIB for reading. Because NAEP reading items are passage-based, ordering the items by difficulty within a passage made the review of constructed-response items much more efficient. This change also made it unnecessary to include the passage that accompanied an item each time the item appeared in the OIB and the CROIB, making the paper materials less cumbersome to use. Tables 2-7 and 2-8 present the contents of the Panel A and Panel B CROIBs for mathematics and reading, respectively. Items appeared in the CROIB in the order listed; however, in the paper version, only the highest score points were shown (e.g., item C3_2 was shown, but C3_1 was not). The items highlighted in yellow were common items (i.e., presented in both Panel A and Panel B CROIBs). Scale and map values were displayed on the NAEP score scale. However, these values were presented on a pseudo-NAEP scale to disguise the NAEP scale, as described later.

The item map for the JSS is a spatially representative display of items ordered by difficulty. Items were ordered on the map from easiest at the bottom to hardest at the top and printed on tabloid-size paper. The score scale at which the item had a 0.67 probability of a correct response was used to locate, or map, the items. In addition to information about the relative difficulty from easiest to hardest, item maps provided information about the actual difference in difficulty between items by placing the ordered items on an interval scale. Items were color-coded to represent the content domain to which they belonged (e.g., literary or informational for reading; geometry, algebra, etc., for mathematics). Items were represented on the item map by an item identification number. Maps used for pilot and operational sessions are shown in Appendix D.

The next three subsections provide more detail related to the construction of the OIB, CROIB, and item maps to include the creation of item identification numbers for items, computation of item scale values, and calculation of item map values.

Table 2-7. CROIB Panel A and B Contents—Mathematics

Redacted

Table 2-8. CROIB Panel A and B Contents—Reading

Redacted

2.3.1. Item Identification Number

An item identification number is a short character string used to represent the item in the OIB and CROIB and on the item map. The first character in the item identification number is "M" if the item is multiple choice and "C" if the item is constructed response. For multiple choice and dichotomously-scored constructed response items, the remaining characters in the item identification number indicate the rank of the item from easiest to hardest, given its scale value, with the easiest item having a rank of 1. Items were ranked separately by item type and item pool. Both mathematics and reading items were divided into two pools of items (Pool A and Pool B). Table 2-9 shows the item identification number, scale values, and map values for the easiest and most difficult items within each item type using the mathematics Panel A item pool as an example. In this example, the multiple-choice item identification numbers ranged from M1 to M53. Three of the dichotomously-scored constructed response items C1, C2, and C28 are included in the constructed response items. Polytomously-scored items have one item identification number for each credited score point. For example, the item identification number C26_2 represents a score of "2" on item C26. Each of these score levels was represented separately and in different locations in the OIB and on the item map and corresponded to its respective scale or map value.

Table 2-9. Item Identification Numbers, Scale Values, and Map Values for Easiest and Hardest Items within Item Type—Mathematics Panel A Item Pool

Redacted

2.3.2. Computation of Item Scale Values

The bookmark method involves rank-ordering items by difficulty and asking panelists to identify the point in the item list at which student performance just good enough to be considered "prepared" has less than a two-thirds chance of correct response. In order to generate this item list, item scale values were needed. The computation of item scale values performed by Measured Progress (for the pilot) and ETS (for the operational sessions) began with the computation of score probabilities conditional on the content areas. Each item in both the mathematics and reading assessments was calibrated separately by ETS to one of the subject-specific content areas shown in Tables 2-10 and 2-11, respectively. Included also in these tables were the slope and intercept for each content area that were used in the scale value calculations and the weights that were determined as part of the framework development.

For multiple-choice and dichotomously-scored items, the item response theory model displayed in equation 2.1 was used to calculate the probability of a correct response (a score of 1) on the item,

$$P\left(U_{ij}=1\middle|\theta_j,\xi_i\right)=P_i\left(\theta_j\right)=c_i+\left(1-c_i\right)\frac{\exp\left[Da_i\left(\theta_j-b_i\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i\right)\right]} \tag{2.1}$$

where
i indexes the items,
j indexes examinees,
θ represents the given ability level for a particular content area subscale,
a represents item discrimination,
b represents item difficulty,
c is the pseudo guessing parameter,
ξ_i represents the set of item parameters (a, b, and c), and
D is a normalizing constant equal to 1.7.

For polytomously-scored items, the item response theory model displayed in equation 2.2 was used to calculate the probability of each possible score on the item,

$$P(U_{ij}=k\mid\theta_j)=P_{ik}\left(\theta_j\right)=\frac{\exp\sum_{v=0}^{k}\left[Da_i\left(\theta_j-b_i+d_{iv}\right)\right]}{\sum_{c=0}^{m_i}\exp\sum_{v=0}^{c}\left[Da_i\left(\theta_j-b_i+d_{iv}\right)\right]} \tag{2.2}$$

where
i indexes the items,
j indexes examinees,
k indexes the score category,
m indexes the maximum score on the given item,
θ represents the given ability level for a particular content area subscale,
a represents item discrimination,
b represents item difficulty,
d represents the category step parameter for score v, and
D is a normalizing constant equal to 1.7.

The composite scale score, η, is related to content area subscale thetas, $\theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ for mathematics and $\theta = \{\theta_1, \theta_2\}$ for reading, through the transformations displayed in equations 2.3 and 2.4,

$$y = A\theta + b \tag{2.3}$$

and

$$\eta = w^t y, \tag{2.4}$$

where A is a diagonal matrix of constants, b is a column vector of constants, and w is a column vector of weights summing to 1. The transformation constants used to create the composite score scale for mathematics and reading used in the JSS sessions can be obtained from Educational Testing Service (ETS).

To obtain the probability of scoring at or above k, conditional on η, a regression procedure based on Donoghue (1997) was used. The integral in equation 2.5 was approximated using numerical integration,

$$P(U_{ij} \geq k|\eta) = \int_{-\infty}^{\infty} P(U_{ij} \geq k|\theta_j) f(\theta_j|\eta)\, \partial\theta_j \tag{2.5}$$

where

$$P(U_{ij} \geq k|\theta_j) = \sum_{h=k}^{m_i} P(U_{ij} = h|\theta_j), \text{ for } k = 1 \text{ or } k = 1, 2, \ldots, m_i.$$

Calculations are conducted under the distributional assumption in 2.6:

$$f(\theta|\eta) \sim N\left(\mu_j + \frac{\sigma_j \rho_{j\eta}(\eta - \mu_j)}{\sigma_j}, \sigma_j^2\left(1 - \rho_{j\eta}^2\right)\right), \tag{2.6}$$

where μ_j and σ_j are the mean and standard deviation of θ_j, μ_η and σ_η are the mean and standard deviation of the composite scale value η, and $\rho_{j\eta}$ is the correlation between θ_j and η which is calculated as shown in 2.7,

$$\rho_{j\eta} = \frac{\mathrm{Co}\ (\theta_j, \eta)}{\sigma_j \sigma_\eta}, \tag{2.7}$$

where $\mathrm{Cov}(\theta_j, \eta)$ is the covariance between θ_j and η and is calculated using equation 2.8,

$$\mathrm{Cov}(\theta_j, \eta) = \sum_{k=1}^{n} w_k A_{kk} \mathrm{Cov}(\theta_j, \theta_k) = \sum_{k=1}^{n} w_k A_{kk} \rho_{jk}\, \sigma_j \sigma_k, \tag{2.8}$$

where n is equal to the number of content areas (i.e., 4 for mathematics and 2 for reading).

The marginal means (μ_j) and standard deviations of the content area subscale thetas (σ_j) are shown in Tables 2-10 and 2-11, respectively, for mathematics and reading. The mean and standard deviation on the mathematics composite score scale (μ_η and σ_η) were 153.30 and 33.61, respectively. For reading, they were 288.27 and 38.28, respectively.

Table 2-10. Marginal Content Area Theta Means and Standard Deviations—Mathematics

Content Area Notation (j)	*Content Area*[A]	*Theta*	
		Mean (μ_j)	*SD (σ_j)*
1	NPO	0.0325	0.9352
2	M&G	0.0290	0.9617
3	DAP	0.0154	1.0115
4	ALG	0.0362	0.9721

[A] NPO = Number Properties and Operations; M&G = Measurement and Geometry; DAP = Data Analysis and Probability; ALG = Algebra

Table 2-11. Marginal Content Area Theta Means and Standard Deviations—Reading

Content Area Notation (j)	*Content Area*[A]	*Theta*	
		Mean (μ_j)	*SD (σ_j)*
1	LIT	0.0509	0.9539
2	INF	0.0495	0.9472

[A] LIT = Literary; INF = Informational

An item scale value was obtained for every score point greater than 0 for the item.

Let η_{ijk} represent the composite scale value of item score k (k > 0) on item i associated with content area subscale j. The value of η_{ijk} was the lowest integer value of η that satisfied the condition in 2.9,

$$P(U_{ij} \geq k|\eta) \geq RP, \tag{2.9}$$

where RP stands for the response probability criterion. For the JSS sessions, an RP of 0.67 was used. If the left side of the equation was less than RP when the composite scale value was equal to the maximum scale score values (η = 300 for mathematics or η = 500 for reading), then η_{ijk} was set to 300 or 500, respectively.

In both the mathematics and reading JSS processes, a constant was added to the item scale value obtained from the previous equation in order to disguise the true scale values from panelists, who were given a copy of the 2009 Grade 12 Reading and Mathematics Nation's Report Card™, which included cut scores for the achievement levels, as well as other data. By design, the constant added varied for each post-secondary activity group and replicate panel. Constants were varied within a subject to deter cross-panel cut score comparisons during the session.

Item scale values varied between the pilot and operational sessions due to the different software programs used to calculate them The JSS Technical Advisory Committee (JSS-TAC) recommended using the ETS values, since ETS developed the process for calculating scale values for NAEP, and Measured Progress and the Governing Board agreed that the ETS computations should be used for the operational studies. Scale values used in the pilot and operational sessions can be found in Appendix C.

2.3.3. Item Map Values

An item map is a spatially representative display of items ordered by difficulty and organized by content. Each panelist had access to an item map on paper and in the CAB. Because of the need to fit all of the items on a single page of a manageable size, all of the scale values could not be presented. To maximize precision, the smallest interval feasible was used. For the JSS sessions all items were mapped to the nearest even-numbered scale value.. Item map values were also displayed in the OIB and CROIBs.

2.4. Setting Bookmarks

Panelists worked independently to translate the post-secondary activity borderline performance descriptions, as described in the Process Report (WestEd & Measured Progress, 2011), onto the score scale by placing a bookmark to divide the items in their OIB into two groups: 1) items easy enough for two-thirds of students whose performance matches the borderline performance description to answer correctly and 2) items too difficult for that expectation. Based on their understanding of the borderline performance descriptions, panelists reviewed the set of items ordered by difficulty, starting with the easiest, until they came to an item

they judged to be too difficult to match the description of borderline performance. . A bookmark was placed immediately preceding that item to locate the cut score. It is important to note that panelists were instructed not to place their bookmarks immediately upon finding an item that seemed too difficult; rather, they were to continue looking until they encountered mostly items that were too difficult, then to go back within that "range of uncertainty" to locate the last or hardest item that two-thirds of minimally prepared students would answer correctly. Placing the bookmark in CAB (described in the CAB documentation in Appendix A) automatically stored each panelist's selected cut score in the database. Three bookmarking rounds were conducted, and panelists were asked to consider feedback based on each round in the placement of their bookmarks during the subsequent rounds. The next section describes the post-round feedback presented.

2.5. Post-Round Feedback

After each round of ratings, feedback was provided to the panelists to inform their judgment on the next round of ratings. The feedback provided was the same as that for other NAEP ALS procedures using a modified bookmark process; however, for the current meeting all of the procedures and all of the feedback were computerized in CAB. Feedback was also provided for panelists to respond to the Data Consequences Questionnaire, as explained later. Feedback after the first round of ratings included the cut score results, a rater location chart, and whole booklet feedback. In the pilot, *p*-value data for items were also provided. The JSS-TAC advised that these data be omitted from the operational sessions because panelists in the pilot study seemed confused by the p-value data and the relative relationship of those data to the data in the item maps based on response probabilities. Appendix E provides more information about feedback presented during the pilot session. After the second round of ratings, feedback included the cut score results, a rater location chart, and consequences feedback. All cut score results were presented on a pseudo-NAEP scale, which was a linear transformation of the NAEP scale, calculated separately for each replicate panel.

2.5.1. Cut Score Results

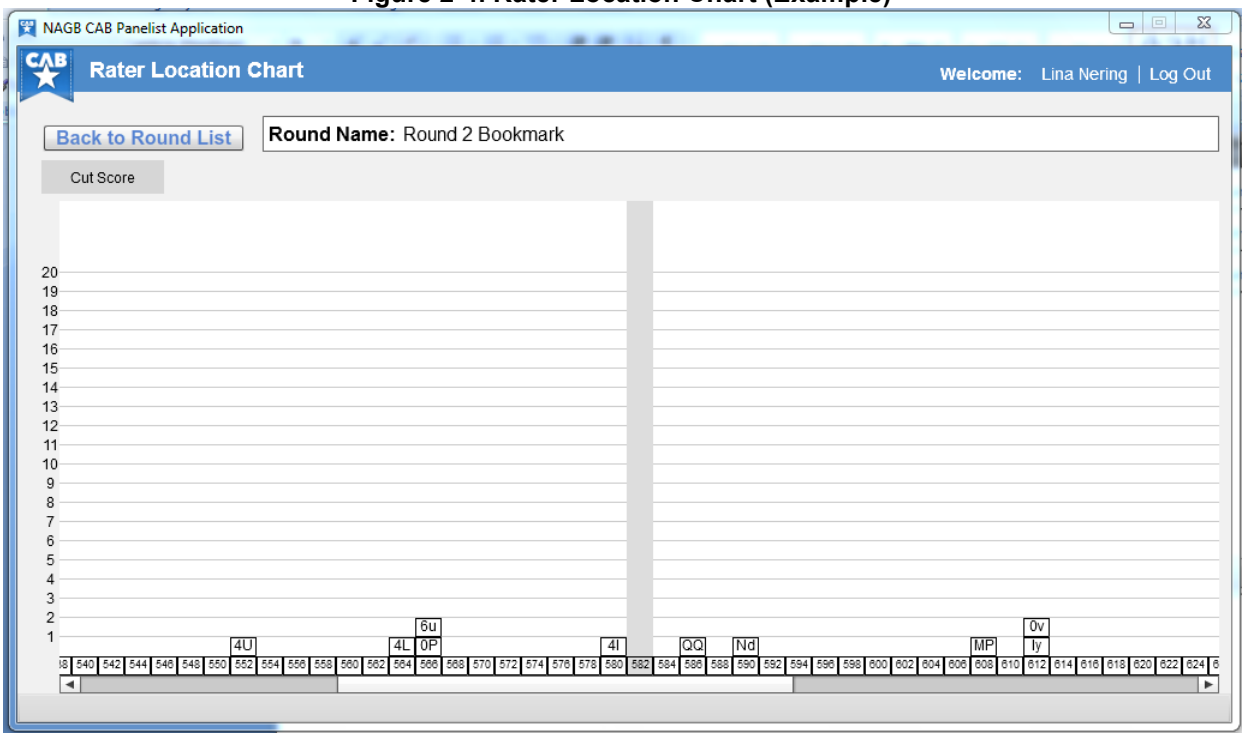
After each round of ratings, the median of the cut scores for each member of the panel (A or B) was determined and reported to panelists within their respective panels. The panelist's cut score was the scale value immediately preceding the placement of the panelist's bookmark.

2.5.2. Rater Location Chart

The rater location chart displayed the distribution of cut scores for all panelists in Panel A or Panel B for a given round of bookmarking, thus providing information on the interrater consistency of the panelists' judgments. Panelists were assigned identification codes to protect confidentiality. Cut scores were rounded to the nearest even integer to enhance the ease of viewing results by reducing the amount of scrolling required in

CAB to view the entire distribution of cut scores. The rater location chart also displayed the median cut score for the panel group. An example of the rater location chart from CAB is shown in Figure 2-4.

Figure 2-4. Rater Location Chart (Example)



2.5.3. Whole Booklet Feedback

After completing round 1 bookmarking, panelists were given feedback in the form of examinee booklets. Six examinee booklets on each of three forms were provided to panelists for each NAEP subject, with each panel reviewing two forms for a total of 12 booklets per replicate panel. Booklets were assigned such that each panelist reviewed one form that was common to both panels A and B. The set of examination booklet from which feedback booklets could be selected were identified prior to the standard-setting sessions. Factors used to select forms included: 1) total number of items within a block, 2) representation of item types (i.e., multiple choice, dichotomously-scored constructed response, and polytomously-scored constructed response), 3) representation of content areas within blocks, 4) mean difficulty (*p*-value), 5) mean discrimination (point biserial), and 6) reliability. Summary statistics are provided at the block level in Table 2-17. Also provided are the averages across all blocks.

From approximately 350 student booklets for each form, 50 were selected for inclusion in the sample booklet pool. Approximately two booklets were selected at each total raw score value, with two exceptions: 1) only one booklet for each score less than 10 was selected and 2) three booklets with raw scores in the middle of the distribution were selected. After the 50 booklets were selected, an electronic version of each booklet was created. This involved the development of a process to perform two main tasks: 1) read student multiple-

choice item responses from a database and mark the student's response within a fabricated electronic version of the form the student took and 2) import images of student responses to constructed-response items provided by the scoring contractor for NAEP into the fabricated booklet. This process was necessary because the actual booklets for the 2009 assessments had already been destroyed.

Table 2-12. Forms Selected for Whole Booklet Feedback and Summary of an Average Block, by Subject

Subject	*Form #/ Type*	*Block ID*	*Total # Points*	*Item Type (# of Points)* [A]			*Content Domain (# of Points)* [B]				*Classical Statistics*		
				MC	*DCR*	*PCR*	*NPO*	*M&G*	*DAP*	*ALG*	*Mean p-value*	*Mean R-BIS*	*Alpha*
Math	AVERAGE		13.58	8.92	1.67	3.00	1.58	4.17	3.25	4.58	0.41	0.66	0.75
	112/ Common	MF	14	10	1	3	1	4	4	5	0.42	0.64	0.73
		MH	13	8	1	4	1	5	4	3	0.42	0.71	0.81
	157/ Panel A	ME	13	9	1	3	2	4	3	4	0.41	0.66	0.77
		MJ	13	8	5	0	1	4	3	5	0.41	0.67	0.76
	142/ Panel B	MI	13	10	1	2	1	4	4	4	0.48	0.67	0.78
		MA	14	9	1	4	2	3	4	5	0.41	0.62	0.72

continued

Subject	Form #/ Type	Block ID	Total # Points	Item Type (# of Points) [A]			Content Domain (# of Points) [B]		Classical Statistics		
				MC	*DCR*	*PCR*	LIT	INF	*Mean p-value*	*Mean R-BIS*	*Alpha*
Reading	AVERAGE		10.08	5.85	0.92	3.31	NA	NA	0.64	0.70	0.74
	56/ Common	RB	10	6	1	3	10	0	0.60	0.70	0.74
		RH	10	6	0	4	0	10	0.64	0.73	0.76
	61/ Panel A	RE	9	4	1	4	9	0	0.58	0.74	0.74
		RJ	10	6	1	3	0	10	0.61	0.66	0.73
	69/ Panel B	RA	10	5	1	4	10	0	0.73	0.69	0.72
		RI	9	5	0	4	0	9	0.59	0.66	0.72

[A] MC = Multiple Choice; DCR = Dichotomously-scored Constructed Response; PCR = Polytomously-scored Constructed Response

[B] NPO = Number Properties and Operations; M&G = Measurement and Geometry; DAP = Data Analysis and Probability; ALG = Algebra; LIT = Literary; INF = Informational

As a result of the differences between the pilot study and the operational sessions, different methods were employed for selecting booklets for whole booklet feedback in the pilot study and subsequent operational studies. The method employed for the pilot session appears in Appendix E. The method used in the operational sessions is described next.

1. Specific student work samples were identified based on the round 1 median cut score for a panel. For each form, six booklets were selected such that they were distributed with respect to the round 1 median cut score. Two booklets from each form scored close to the cut score (one on each side of the median cut score). Two booklets from each form scored within the second quartile of the distribution of panelists' round 1 cut score recommendations (below the median). Two booklets from each form scored within the third quartile of the distribution of panelists' round 1 cut score recommendations (above the median). Booklets identified for review mostly fell between the first and third quartile range of the distribution of panelists' individual cut scores, but they may have extended from the lowest to the highest cut score set by any panelist in the group due to availability of booklets meeting the above criteria.

NAEP calculates plausible values for the purpose of reporting scores. In order to generate individual booklet scores that were independent of demographics, independent of subscale performance, and easier to understand by panelists, scores were estimated for booklets in the sampling pool using the following method:

1. Test Characteristic Curves (TCC) for each of the subscales were constructed individually based on the items in the form the examinee took.
2. Subscale raw scores were calculated for each booklet.
3. Subscale theta scores were assigned based on the appropriate TCCs.
4. Subscale theta scores were then transformed to the subscale reporting metric (i.e., scale score) using the appropriate means and standard deviation for each subscale on the theta scale and on the score scale.
5. Finally, weighted sums of the subscale scale scores from step 4 were calculated based on the appropriate framework weights to derive the booklet's total scale score.

The theta metric was set to range from -4 to +4.

Following the identification of booklets, a booklet score chart (BSC) was generated for each panel to aid panelists in the interpretation of the whole booklets. The BSC maps the scale score (in descending order) to the actual percent of possible points the examinee obtained for his or her responses to the specified form. Booklets are labeled from 1A (lowest scale score) to 6A (highest scale score) for the form common to the replicate panels, and 1B to 6B for the form unique to the panel. Figure 2-1 displays an example BSC. Additional information on the BSC includes: 1) a line to indicate the lowest panelist cut, 2) a highlighted line to indicate the median panel cut, and 3) a line to indicate the highest panelist cut. All are represented on the panel-appropriate pseudo-NAEP scale.

In some cases, there may be a lack of direct correspondence between total possible points and scale scores such that higher a scale score may be associated with lower "% of Total Possible Points" score. The primary explanation for this is that subscale scores were assigned separately and then weighted in the final total score. The weights were determined by content experts who develop the framework. They decided the relative weights of each subscale for each grade level based on content issues in the academic domain. This resulted in unequal weighting of raw score points such that two examinees with the same total raw score may earn different scale scores, depending on the weighting of the subscales. Finally, this allowed for booklets drawn from a particular scale score to represent different raw scores that do not always appear ordered according to the scale scores. An example of this can be observed in Figure 2-1. For Form 112, the intuitive decreasing percent score pattern is observed. However, for Form 157, this is not the case. Here we see that the "% of Total Possible Points" reported in the last column appear to be out of order. These reversals are especially likely when one set of examinees obtain most of their points in highly weighted content domains while other examinees with higher raw scores obtain their points in lower weighted content domains.

After consulting with two JSS-TAC members, for operational JSS sessions 2 and 3, highly aberrant booklets—those with low raw scores and high scale scores, and vice versa, in relationship to the other booklets—were removed from the sample. Figure 2-2 displays an example BSC from the third operational JSS session.

Figure 2-5. Example Booklet Score Chart—JSS Operational Session 1

NAEP JSS						
Grade 12 Mathematics Panel A College Prep						
		Form 112			Form 157	
	Scale	Booklet	% of Total Possible Points		Booklet	% of Total Possible Points
	615					
	614					
	613					
Highest Panelist Cut Score	612					
	611					
	610					
	609					
	608					
	607				6B	50
	606	6A	68			
	605					
	604					
	603				5B	63
	602	5A	66			
	601					
	600					
	599	4A	66		4B	63
Panel Cut Score→	598					
	597	3A	63			
	596				3B	42
	595					
	594					
	593					
	592					
	591				2B	53
	590	2A	58			
	589					
	588					
	587				1B	55
	586					
	585					
	584					
	583					
	582					
	581	1A	53			
	580					
	579					
	578					
	577					
	576					
	575					
	574					
	573					
	572					
	571					
	570					
	569					
	568					
	567					
	566					
	565					
	564					
	563					
Lowest Panelist Cut Score	562					
	561					
	560					
	559					

Figure 2-6. Example Booklet Score Chart—JSS Operational Sessions 2 and 3

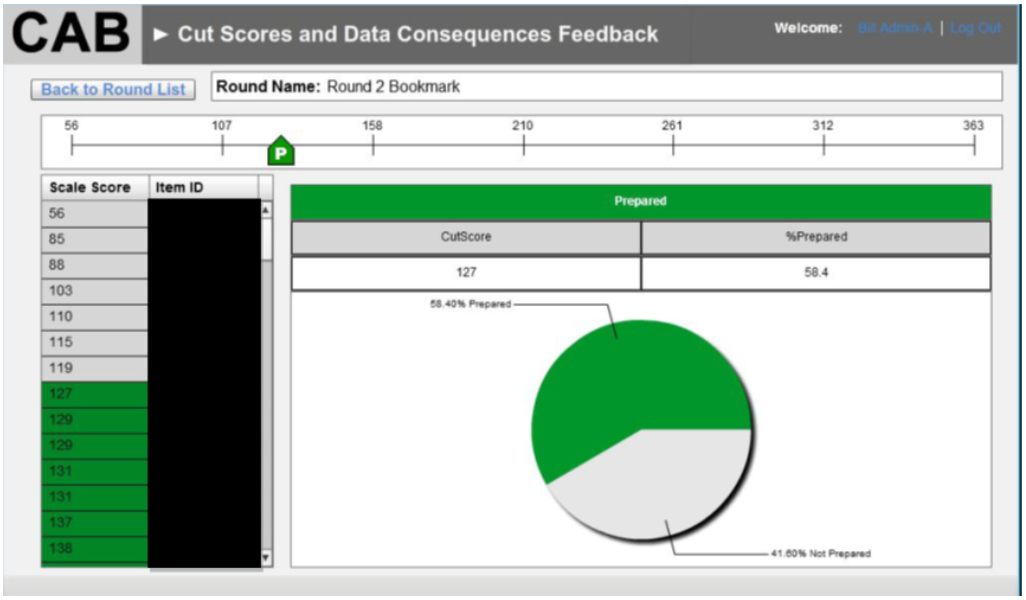
NAEP JSS Grade 12 Reading Panel A HVAC						
		Form 56			Form 61	
	Scale	Booklet	% of Total Possible Points		Booklet	% of Total Possible Points
	713					
	712					
	711					
Highest Panelist Cut Score	710					
	709					
	708					
	707					
	706					
	705					
	704					
	703					
	702					
	701					
	700					
	699					
	698					
	697	6A	62		5B,6B	61,61
	696					
	695	5A	62		4B	61
	694					
	693	4A	62		3B	57
	692					
	691					
Panel Cut Score →	690	3A	59		2B	57
	689					
	688					
	687					
	686	2A	59			
	685					
	684					
	683	1A	55			
	682				1B	54
	681					
	680					
	679					
	678					
	677					
	676					
	675					
	674					
	673					
	672					
	671					
Lowest Panelist Cut Score	670					
	669					
	668					
	667					

2.5.4. Consequences Feedback

Consequences feedback reported the percentage of students who performed at or above the panel cut score. The cumulative frequency distribution of student performances based on the 2009 assessment was provided to Measured Progress by ETS, and these tables can be found in Appendix F. CAB displayed the consequences data feedback on an interactive consequences data screen as pictured in Figure 2-7. Panelists could use the cut score adjuster (located at the top of the screen) to examine how the consequences data changed relative to alternative cut scores. The table and pie chart on the right side of the screen showed the percentage of students who scored at or above the cut score (% Prepared). An item list is displayed on the left side of the screen where items with a scale score at or above the cut score were highlighted in green.

Figure 2-7. Cut Scores and Data Consequences Feedback

Partially Redacted (Item ID Column only)



2.6. Consequences Questionnaire

Consequences data feedback were presented to panelists using the CAB application as described in the previous section and as displayed in Figure 2-7 above. Consequences data feedback were 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 score seemed appropriate or if it should be higher or lower. Panelists' reactions to the consequences data are summarized and presented in the process evaluation results section for each workshop in the Process Report (WestEd & Measured Progress, 2011, pp. 121, 135, 152, 166, 182, 196).

2.7. Selecting Potential Exemplar Items

After the bookmarking rounds were completed, panelists were asked to make recommendations for exemplar items (i.e., items that illustrate the knowledge and skills representing preparedness for entry-level coursework in credit-bearing college courses or occupational job-training programs). Potential exemplar items in the JSS session were drawn from blocks of grade 12 NAEP items that had been released to the public. These items were not included in the pool of items that panelists used to set their bookmarks, but these were the item blocks in the form of the NAEP administered to panelists as part of their training for the process.[1] Items were identified as potential exemplars to be included in the Exemplar Item Questionnaire, dependent upon the individual panel cut score. An item was included in the questionnaire if its scale value was equal to or greater than the panel's median cut score for round 3.

During the exemplar selection task, panelists rated the items as to whether the items should definitely be used, were okay to use, or should not be used as exemplars. They were allowed to discuss potential exemplars with other panelists, but they had to provide their ratings of these items in CAB independently. A full summary of the numerical results of the exemplar selection task can be found in Table 2-13.

Table 2-13. Summary of Exemplar Items Selection

Operational Workshop	*Panel*	*# of Panelists*	*# of Items Presented*	*Median Cut Score*	*# 100% Very Good/OK (Average Scale Value)*	*# at least 75% Very Good/OK (Average Scale Value)*
College-Preparedness	MCA	10	17	201	3 (218)	9 (231)
	MCB	10	21	189	1 (239)	13 (230)
	RCA	10	16	290	4 (307)	14 (328)
	RCB	9	14	304	5 (341)	13 (350)
Automotive Master Technician	MAA	7	29	167	4 (186)	7 (193)
	MAB	8	29	171	2 (172)	6 (197)
	RAA	5	13	308	10 (351)	13 (352)
	RAB	6	16	294	11 (353)	16 (342)
Licensed Practical Nurse (LPN)	MLA	10	26	177	5 (200)	13 (217)
	MLB	10	20	193	1 (198)	6 (238)
	RLA	10	13	307	7 (352)	12 (356)
	RLB	10	18	288	7 (344)	16 (329)
Pharmacy Technician	MPA	9	26	174	1 (198)	8 (206)
	MPB	9	26	176	4 (200)	11 (207)
	RPA	10	10	321	4 (354)	7 (358)
	RPB	9	14	299	6 (322)	12 (341)
Computer Support Specialist	MSA	10	30	165	1 (172)	11 (189)
	MSB	10	23	185	2 (206)	7 (228)

continued

[1] Tables 2-5 and 2-6 presented earlier in this report display the summary statistics for the released blocks.

Operational Workshop	*Panel*	*# of Panelists*	*# of Items Presented*	*Median Cut Score*	*# 100% Very Good/OK (Average Scale Value)*	*# at least 75% Very Good/OK (Average Scale Value)*
Computer Support Specialist	RSA	10	16	292	3 (320)	11 (347)
	RSB	10	13	307	5 (327)	9 (337)
Heating, Ventilation, and Air Conditioning Technician (HVAC)	MHA	10	26	177	0 (N/A)	4 (217)
	MHB	9	29	172	4 (200)	11 (207)
	RHA	10	18	289	5 (328)	15 (324)
	RHB	9	16	292	8 (332)	13 (334)

2.8. Process Evaluations

A process evaluation form was completed by panelists after each major JSS task (e.g., at the end of each day and after each bookmarking round). The Process Report (WestEd & Measured Progress, 2011) contains the JSS agenda, which provides more detail on when evaluations were conducted. Process evaluations were administered using CAB. Panelists were asked to indicate their degree of understanding of process tasks, materials, and instructions. Results from the process evaluations were used both to clarify areas of confusion during the course of the session and to provide evidence of procedural validity. The responses in the process evaluations were on a five-point Likert scale. For each item, the mean value for the responses and the standard deviation were calculated. Open responses were also solicited and used mainly to inform the process.

Chapter 3—CUT SCORE EVALUATION

This chapter describes how the cut scores resulting from the JSS sessions were evaluated. Variability of cut scores was estimated using the mean absolute deviation algorithm along with an analysis of how panelists' cut scores changed from one round to the next. Two standard error estimates were calculated and are reported; and reliability was evaluated using results from the two replicate panels.

3.1. Variability of Cut Scores

Panel cut scores were calculated by obtaining the median cut score within the panel. Therefore, describing variation of the cut scores within a panel using a standard deviation calculation is not appropriate. Instead, variation is described in two ways: 1) mean absolute deviation (MAD) indices and 2) cut scores changes between rounds.

The MAD is the average difference between each panelist's cut score and the median cut score as shown in equation 3.1,

$$MAD = \frac{|x_i - x_{Md}|}{n}, \tag{3.1}$$

where x_i represents a panelist's cut score on the NAEP scale score scale, $x_{Mdn}$is the panel's median cut score, and n is the number of panelists in the panel. Tables 3-1 and 3-2 report MAD for each panel and for each bookmarking round for mathematics and reading, respectively.

Table 3-1. Mean Absolute Deviation (MAD) by Panel and Round—Mathematics Grade 12

Post-Secondary Activity	*Panel*	*Round*		
		1	*2*	*3*
Automotive Master Technician	A	13.0	18.4	15.7
	B	10.4	5.6	5.6
College-Preparedness	A	14.9	4.6	3.1
	B	16.8	4.9	4.8
Computer Support Specialist	A	14.4	4.7	6.1
	B	16.9	9.3	7.3
Licensed Practical Nurse (LPN)	A	4.0	4.4	1.7
	B	16.8	7.4	8.5
Heating, Ventilation, and Air Conditioning Technician (HVAC)	A	21.8	11.1	10.9
	B	17.4	5.7	7.2
Pharmacy Technician	A	7.6	6.1	5.2
	B	20.7	5.4	11.0

Table 3-2. Mean Absolute Deviation (MAD) by Panel and Round—Reading Grade 12

Post-Secondary Activity	*Panel*	*Round*		
		1	*2*	*3*
Automotive Master Technician	A	13.8	12.8	15.6
	B	13.7	6.8	6.8
College-Preparedness	A	23.9	10.2	5.9
	B	23.7	3.8	3.6
Computer Support Specialist	A	16.5	9.1	6.5
	B	12.6	4.9	3.6
Licensed Practical Nurse (LPN)	A	19.2	12.9	8.5
	B	17.5	7.7	4.9
Heating, Ventilation, and Air Conditioning Technician (HVAC)	A	7.4	5.2	5.2
	B	14.2	3.9	2.9
Pharmacy Technician	A	11.0	5.0	6.1
	B	12.1	6.8	6.8

A summary of the individual panelists' cut score changes between rounds provides additional information about the direction of how cut scores varied within a panel. Tables 3-3 and 3-4 report the number of panelists whose cut scores increased, decreased, or had no change from the previous round for mathematics and reading, respectively. Changes between rounds 1 and 2 are labeled "R1-R2," while changes between rounds 2 and 3 are labeled "R2-R3."

Table 3-3. Round to Round Cut Score Changes by Panel—Mathematics Grade 12

Post-Secondary Activity	*Panel*	*Round*	*Increased n (%)*	*No Change n (%)*	*Decreased n (%)*
Automotive Master Technician	A	R1-R2	0 (0.0)	1 (14.3)	6 (85.7)
		R2-R3	1 (14.3)	3 (42.9)	3 (42.9)
	B	R1-R2	3 (37.5)	3 (37.5)	2 (25.0)
		R2-R3	0(0.0)	6 (75.0)	2 (25.0)
College-Preparedness	A	R1-R2	5 (50.0)	1 (10.0)	4 (40.0)
		R2-R3	6 (60.0)	2 (20.0)	2 (20.0)
	B	R1-R2	5 (50.0)	1 (10.0)	4 (40.0)
		R2-R3	2 (20.0)	5 (50.0)	3 (30.0)
Computer Support Specialist	A	R1-R2	4 (40.0)	0 (0.0)	6 (60.0)
		R2-R3	1 (10.0)	2 (20.0)	7 (70.0)
	B	R1-R2	6 (60.0)	2 (20.0)	2 (20.0)
		R2-R3	1 (10.0)	3 (30.0)	6 (60.0)
Licensed Practical Nurse (LPN)	A	R1-R2	4 (40.0)	5 (50.0)	1 (10.0)
		R2-R3	0(0.0)	7 (70.0)	3 (30.0)
	B	R1-R2	6 (60.0)	2 (20.0)	2 (20.0)
		R2-R3	2 (20.0)	2 (20.0)	6 (60.0)
Heating, Ventilation, and Air Conditioning Technician (HVAC)	A	R1-R2	6 (60.0)	1 (10.0)	3 (30.0)
		R2-R3	0 (0.0)	4 (40.0)	6 (60.0)
	B	R1-R2	2 (22.2)	0 (0.0)	7 (77.8)
		R2-R3	2 (22.2)	6 (66.7)	1 (11.1)
Pharmacy Technician	A	R1-R2	3 (33.3)	3 (33.3)	3 (33.3)
		R2-R3	0 (0.0)	4 (44.4)	5 (55.6)

continued

Post-Secondary Activity	Panel	Round	Increased n (%)	No Change n (%)	Decreased n (%)
Pharmacy Technician	B	R1-R2	3 (33.3)	1 (11.1)	5 (55.6)
		R2-R3	0 (0.0)	1 (11.1)	8 (88.9)

Table 3-4. Round to Round Cut Score Changes by Panel—Reading Grade 12

Post-Secondary Activity	Panel	Round	Increased n (%)	No Change n (%)	Decreased n (%)
Automotive Master Technician	A	R1-R2	1 (20.0)	3 (60.0)	1 (20.0)
		R2-R3	1 (20.0)	1 (20.0)	3 (60.0)
	B	R1-R2	3 (50.0)	1 (16.7)	2 (33.3)
		R2-R3	0 (0.0)	6 (100.0)	0 (0.0)
College-Preparedness	A	R1-R2	8 (80.0)	0 (0.0)	2 (20.0)
		R2-R3	5 (50.0)	4 (40.0)	1 (10.0)
	B	R1-R2	4 (44.4)	0 (0.0)	5 (55.6)
		R2-R3	0 (0.0)	8 (88.9)	1 (11.1)
Computer Support Specialist	A	R1-R2	3 (30.0)	2 (20.0)	5 (50.0)
		R2-R3	0 (0.0)	8 (80.0)	2 (20.0)
	B	R1-R2	7 (70.0)	0 (0.0)	3 (30.0)
		R2-R3	1 (10.0)	5 (50.0)	4 (40.0)
Licensed Practical Nurse (LPN)	A	R1-R2	5 (50.0)	1 (10.0)	4 (40.0)
		R2-R3	1 (10.0)	7 (70.0)	2 (20.0)
	B	R1-R2	4 (40.0)	2 (20.0)	4 (40.0)
		R2-R3	2 (20.0)	4 (40.0)	4 (40.0)
Heating, Ventilation, and Air Conditioning Technician (HVAC)	A	R1-R2	4 (40.0)	6 (60.0)	0 (0.0)
		R2-R3	0 (0.0)	10 (100.0)	0 (0.0)
	B	R1-R2	6 (66.7)	0 (0.0)	3 (33.3)
		R2-R3	1 (11.1)	6 (66.7)	2 (22.2)
Pharmacy Technician	A	R1-R2	7 (70.0)	0 (0.0)	3 (30.0)
		R2-R3	0 (0.0)	7 (70.0)	3 (30.0)
	B	R1-R2	6 (66.7)	2 (22.2)	1 (11.1)
		R2-R3	0 (0.0)	9 (100.0)	0 (0.0)

3.2. Estimates of Standard Errors of Cut Scores

The median was used as the cut score in this standard-setting process. Therefore, the usual method of calculating the standard error, based on the mean, does not give an accurate measure of the variability of the cut score. Since the underlying shape of the distribution of the cut scores is unknown, estimates of variation must be based on approximations. Two approximations were 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 even, then the kth moment of the median is given using equation 3.2,

$$E[\text{median}]^k = \int x^k \binom{n}{n/2 - 1} \binom{n/2 + 1}{1} \left(F(x)\right)^{n/2 - 1} \left(1 - F(x)\right)^{n/2} f(x)dx, \tag{3.2}$$

where f(x) is the probability density function of the median, and F(x) is the cumulative distribution function. A similar expression holds when *n* is odd. This integral can be transformed to an integral of the beta probability density function using the transformation y = F(x). At the ith ordered cut score, the value of *y* is i/n . Therefore, the integral can be approximated as shown in equation 3.3,

$$\sum_{i=1}^{n} \left(i/n\right)^k \left\{F_\beta\left(1/n, n/2, n/2 + 1\right) - F_\beta\left(i - 1/n, n/2, n/2 + 1\right)\right\}, \tag{3.3}$$

where $F_\beta(x, \alpha_1, \alpha_2)$ is the cumulative distribution function at the point *x* for a beta distribution with parameters α_1 and α_2.

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.

Theoretically, the standard error estimates are only valid for the first round of cut scores, since cut scores for subsequent rounds are influenced by the location of the cut scores for the other panelists and are not truly independent values. Tables 3-5 and 3-6 present these standard error estimates for mathematics and reading, respectively, across tables and groups (i.e., replicate panels) within post-secondary activities.

Table 3-5. Estimates of Standard Error of Cut Scores for NAEP—Mathematics Grade 12

Post-Secondary Activity	Group	Table	N	Round 1 Median	Round 1 EmpSE	Round 1 BootSD	Round 2 Median	Round 2 EmpSE	Round 2 BootSD	Round 3 Median	Round 3 EmpSE	Round 3 BootSD
Automotive Master Technician	A	1	3	201.0	9.37	7.45	167.0	24.56	23.43	168.0	12.87	10.22
		2	4	185.5	9.09	7.34	180.0	10.75	9.04	164.0	16.53	14.03
		Both	7	200.0	9.29	9.24	174.0	11.18	11.56	167.0	10.63	9.70
	B	1	4	170.5	4.96	4.21	168.5	1.83	1.78	168.5	1.83	1.78
		2	4	162.5	10.62	9.03	173.0	6.43	5.27	172.5	6.67	5.50
		Both	8	166.0	5.40	4.63	170.5	2.66	2.49	170.5	2.54	2.39
	All		15	184.0	7.75	7.72	172.0	3.40	2.95	168.0	2.62	2.38
College-Preparedness	A	1	5	177.0	7.59	7.49	201.0	5.74	6.89	201.0	2.96	3.39
		2	5	204.0	3.17	2.96	201.0	3.34	3.94	201.0	2.48	2.87
		Both	10	196.0	8.86	9.32	201.0	1.83	2.38	201.0	1.38	1.64
	B	1	5	206.0	6.70	7.35	192.0	0.90	0.74	191.0	2.71	2.91
		2	5	174.0	6.65	7.02	185.0	4.14	4.23	186.0	4.79	4.99
		Both	10	184.0	10.10	9.75	191.0	2.78	3.09	188.5	2.49	2.55
	All		20	188.0	8.38	8.37	192.0	2.98	2.85	192.0	3.23	3.10
Computer Support Specialist	A	1	5	183.0	11.73	11.60	172.0	4.09	3.41	172.0	5.12	5.37
		2	5	168.0	10.81	8.80	167.0	3.24	2.80	163.0	5.33	5.55
		Both	10	172.5	7.61	7.21	172.0	2.42	2.35	164.5	3.31	3.25
	B	1	5	201.0	13.14	13.96	185.0	12.10	10.86	183.0	5.70	4.80
		2	5	169.0	14.70	13.21	183.0	5.94	5.24	186.0	6.28	6.37
		Both	10	189.5	12.55	12.65	185.0	4.88	4.08	184.5	3.79	3.57
	All		20	178.0	8.55	8.20	177.0	3.66	3.65	172.0	5.22	5.20
Heating, Ventilation, and Air Conditioning Technician (HVAC)	A	1	5	177.0	18.58	20.06	177.0	6.08	5.08	177.0	5.49	5.81
		2	5	198.0	15.90	16.14	200.0	10.14	10.61	197.0	11.06	11.48
		Both	10	180.0	11.33	11.34	185.0	7.43	7.14	177.0	6.58	5.90
	B	1	5	192.0	10.64	10.40	173.0	2.60	2.70	173.0	8.92	6.57
		2	4	168.0	16.89	12.43	170.0	8.12	6.17	170.0	3.58	3.01
		Both	9	177.0	11.91	11.14	172.0	2.54	2.65	172.0	2.75	2.88
	All		19	177.0	9.49	9.34	177.0	3.86	3.50	174.0	2.75	2.47
Licensed Practical Nurse (LPN)	A	1	5	174.0	5.65	4.40	177.0	2.92	2.35	177.0	2.71	2.22
		2	5	177.0	0.87	0.93	179.0	4.70	3.69	177.0	0.47	0.33
		Both	10	177.0	1.30	1.32	177.0	1.96	1.62	177.0	0.46	0.48
	B	1	5	182.0	12.16	12.09	200.0	5.24	5.64	185.0	4.17	3.35
		2	5	175.0	16.37	14.99	192.0	6.47	6.16	200.0	3.53	4.29
		Both	10	179.0	10.03	9.60	198.5	4.56	4.91	192.5	6.59	6.56
	All		20	177.0	2.08	1.89	185.0	4.56	4.37	180.5	3.33	3.26
Pharmacy Technician	A	1	5	174.0	3.85	3.08	174.0	1.87	1.32	174.0	1.78	1.96
		2	4	178.0	10.15	7.99	181.5	9.27	8.01	175.5	8.06	6.80
		Both	9	174.0	3.78	3.25	174.0	3.66	3.04	174.0	1.78	2.05
	B	1	5	206.0	27.09	30.70	202.0	5.06	5.55	176.0	7.10	7.94
		2	4	217.5	5.69	6.15	208.5	2.29	2.29	179.0	11.25	11.84
		Both	9	214.0	9.06	10.89	206.0	2.79	2.87	176.0	6.29	6.93
	All		18	185.5	11.87	11.73	194.5	9.11	9.34	174.0	1.58	1.56

Note: EmpSE = Empirical Standard Error; BootSD = Bootstrapped Standard Deviation

Table 3-6. Estimates of Standard Error of Cut Scores for NAEP—Reading Grade 12

Post-Secondary Activity	Group	Table	N	Round 1			Round 2			Round 3		
				Median	EmpSE	BootSD	Median	EmpSE	BootSD	Median	EmpSE	BootSD
Automotive Master Technician	A	1	5	321.0	11.89	10.27	308.0	12.34	9.66	308.0	12.34	9.66
	B	1	6	293.5	8.84	8.08	293.5	5.75	5.13	293.5	5.75	5.13
	All		11	308.0	7.98	8.09	308.0	5.73	6.39	308.0	5.73	6.39
College-Preparedness	A	1	5	297.0	22.96	22.03	299.0	7.05	7.43	291.0	5.83	5.17
		2	5	268.0	7.10	7.11	281.0	7.62	7.63	287.0	3.94	4.36
		Both	10	272.5	10.31	9.65	288.0	5.34	5.35	289.5	2.20	2.12
	B	1	4	313.5	21.57	16.63	305.5	5.46	4.54	305.5	5.46	4.54
		2	5	281.0	7.62	6.54	304.0	1.28	1.48	304.0	1.17	1.42
		Both	9	303.0	12.87	13.03	304.0	1.48	1.57	304.0	1.32	1.43
	All		19	281.0	9.53	9.23	299.0	3.98	4.15	299.0	4.38	4.54
Computer Support Specialist	A	1	5	279.0	6.87	5.75	297.0	8.06	7.76	293.0	3.68	3.63
		2	5	301.0	13.93	12.10	291.0	6.83	6.70	291.0	6.83	6.70
		Both	10	292.5	8.61	8.63	294.0	5.36	4.97	292.0	3.20	3.24
	B	1	5	306.0	16.18	16.27	308.0	0.75	0.58	308.0	0.93	1.14
		2	5	304.0	5.05	5.48	306.0	7.28	6.22	306.0	5.41	5.67
		Both	10	305.0	5.51	6.27	308.0	0.99	1.03	307.0	1.81	2.16
	All		20	300.0	5.93	6.13	306.5	4.09	4.34	297.0	4.23	4.18
Heating, Ventilation, and Air Conditioning Technician (HVAC)	A	1	5	286.0	5.84	6.05	288.0	4.26	4.63	288.0	4.26	4.63
		2	5	289.0	5.42	4.75	289.0	4.70	3.58	289.0	4.70	3.58
		Both	10	288.0	3.00	3.11	288.5	1.78	1.63	288.5	1.78	1.63
	B	1	4	279.5	9.97	8.44	287.0	3.43	3.13	287.5	1.75	1.37
		2	5	283.0	14.77	12.40	292.0	1.69	1.81	292.0	1.69	1.81
		Both	9	283.0	8.96	7.86	292.0	2.57	2.62	292.0	2.32	2.37
	All		19	286.0	3.44	3.28	289.0	1.79	1.75	289.0	1.77	1.73
Licensed Practical Nurse (LPN)	A	1	5	323.0	22.81	22.54	321.0	10.59	8.95	314.0	6.06	5.61
		2	5	308.0	10.34	9.18	303.0	5.77	4.96	302.0	5.12	3.99
		Both	10	308.0	12.22	11.56	311.0	6.20	6.02	307.0	5.15	4.75
	B	1	5	280.0	8.51	7.38	288.0	4.86	3.71	288.0	0.93	0.66
		2	5	306.0	19.31	19.65	287.0	10.02	10.12	283.0	6.96	6.96
		Both	10	282.0	12.18	11.67	288.0	3.22	3.48	288.0	2.04	2.27
	All		20	299.5	8.26	8.23	297.5	4.85	4.71	296.0	4.28	4.18
Pharmacy Technician	A	1	5	328.0	11.61	11.83	323.0	2.88	2.68	323.0	2.88	2.68
		2	5	302.0	2.86	2.78	321.0	6.16	6.20	314.0	6.34	6.01
		Both	10	307.5	6.23	5.28	322.0	2.82	2.84	321.0	3.13	3.31
	B	1	4	305.0	10.02	8.47	305.0	4.38	4.94	305.0	4.38	4.94
		2	5	288.0	5.57	6.52	292.0	5.01	4.03	292.0	5.01	4.03
		Both	9	289.0	5.96	5.39	299.0	5.52	5.44	299.0	5.52	5.44
	All		19	302.0	4.29	4.23	311.0	5.98	5.97	308.0	5.30	5.21

Note: EmpSE = Empirical Standard Error; BootSD = Bootstrapped Standard Deviation

3.3. Reliability Analyses

The reliability of cut scores obtained during a NAEP standard-setting session is thought of in terms of how consistent the cut scores are between replicate panels when using the same standard-setting procedures,

assessment, and borderline performance description. Cut score reliability is evaluated by examining the standard error of the cut score. The interpretation of this standard error is such that lower values indicate a more reliable cut score.

Within a post-secondary activity for the JSS, there were two replicate panels (A and B), each of which produced a median cut score. Therefore, there are two independent observations for job training programs in each post-secondary activity. To calculate the standard error using two observations, equation 3.4 is used (Brennan, 2002),

$$\widehat{\sigma}_{\overline{X}} = \frac{|X_1 - X_2|}{2}. \tag{3.4}$$

Tables 3-3 and 3-4 present these standard error estimates for both mathematics and reading, respectively, for each post-secondary activity in this set of studies. Also included in the tables are the 95% confidence intervals for the mean cut score calculated as the average of the median cut scores for the two replicate panels for each post-secondary activity. Confidence intervals are displayed graphically in Figures 3-1 and 3-2 for mathematics and reading, respectively. In Figures 3-1 and 3-2, the horizontal axis is placed at the NAEP Proficient cut point (i.e., 176 for mathematics and 302 for reading) as a point of comparison. The lower and upper bounds of the vertical axis are set at the Basic (i.e., 141 for mathematics and 265 for reading) and Advanced (i.e., 216 for mathematics and 346 for reading) cut points, respectively.

Table 3-7. Standard Error Estimates and Confidence Intervals for Mean Cut Scores by Post-Secondary Activity—Mathematics

Post-Secondary Activity	*Cut Score*			*Standard Error*	*95% Confidence Interval*	
	Panel A	*Panel B*	*Mean*		*Upper Limit*	*Lower Limit*
Automotive Master Technician	167	171	169.0	2.0	172.9	165.1
College-Preparedness	201	189	195.0	6.0	206.8	183.2
Computer Support Specialist	165	185	175.0	10.0	194.6	155.4
Heating, Ventilation, and Air Conditioning Technician (HVAC)	177	172	174.5	2.5	179.4	169.6
Licensed Practical Nurse (LPN)	177	193	185.0	8.0	200.7	169.3
Pharmacy Technician	174	176	175.0	1.0	177.0	173.0

Table 3-8. Standard Error Estimates and Confidence Intervals for Mean Cut Scores by Post-Secondary Activity—Reading

Post-Secondary Activity	*Cut Score*			*Standard Error*	*95% Confidence Interval*	
	Panel A	*Panel B*	*Mean*		*Upper Limit*	*Lower Limit*
Automotive Master Technician	308	294	301.0	7.0	314.7	287.3
College-Preparedness	290	304	297.0	7.0	310.7	283.3
Computer Support Specialist	292	307	299.5	7.5	314.2	284.8
Heating, Ventilation, and Air Conditioning Technician (HVAC)	289	292	290.5	1.5	293.4	287.6
Licensed Practical Nurse (LPN)	307	288	297.5	9.5	316.1	278.9
Pharmacy Technician	321	299	310.0	11.0	331.6	288.4

Figure 3-1. Mean Cut Scores and Confidence Intervals by Post-Secondary Activity—Mathematics

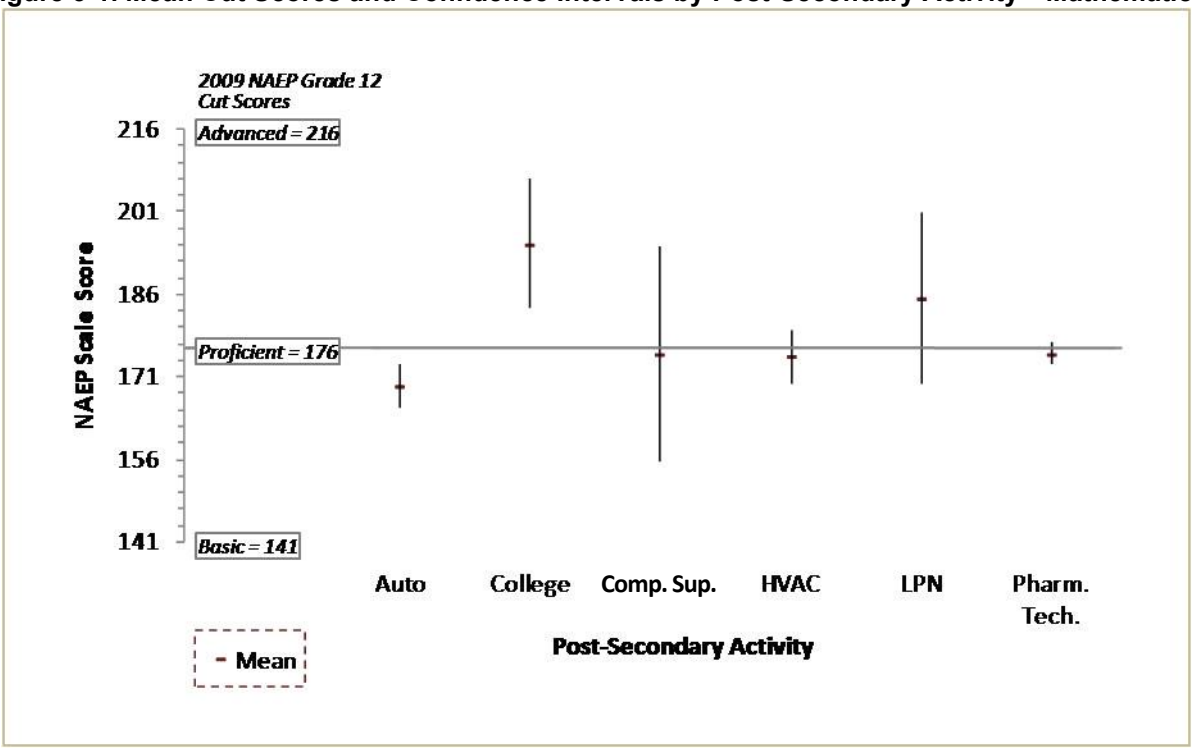
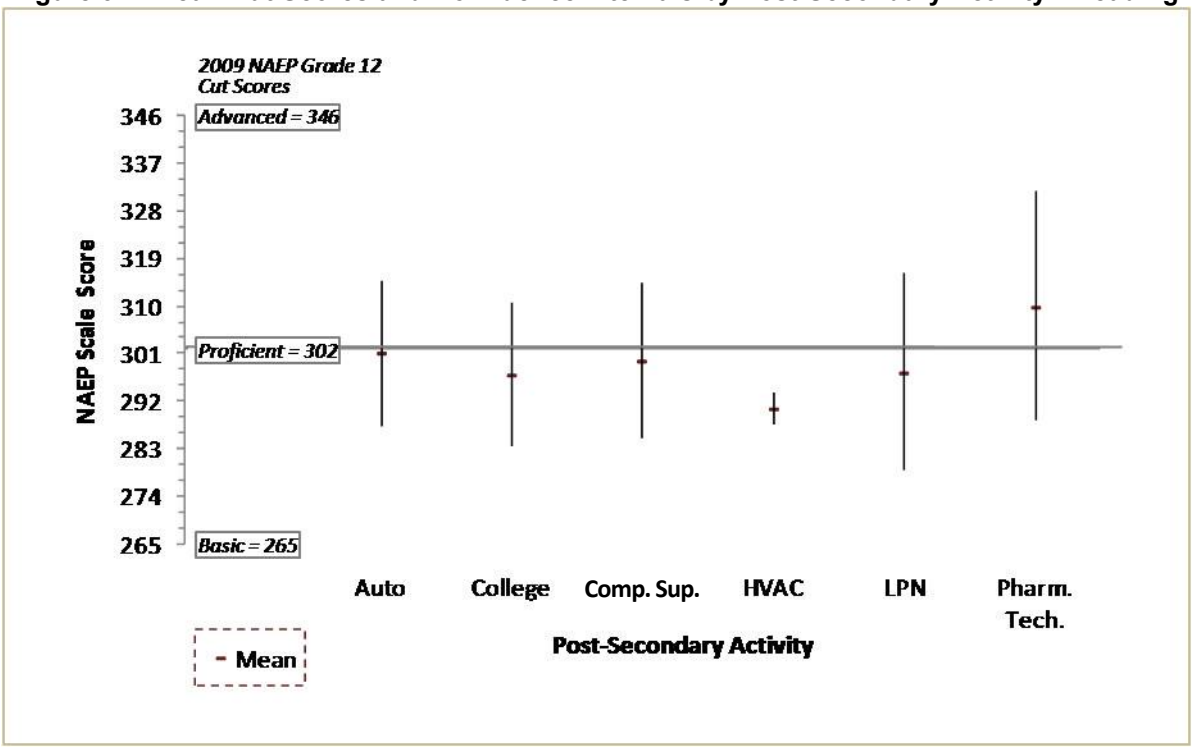


Figure 3-2. Mean Cut Scores and Confidence Intervals by Post-Secondary Activity—Reading



Chapter 4—SPECIAL ANALYSES

This chapter presents two special analyses. The first was conducted by JSS-TAC member Ed Haertel to explore the possibility of facilitator effects across sessions. The second was conducted to determine if the presence of items identified by panelists as "irrelevant" significantly impacted the placement of cut scores. This second study is described in more detail in the Process Report (WestEd & Measured Progress, 2011).

4.1. Facilitator Effect Study

Considerable variability was observed in the JSS workshops' resulting cut scores despite efforts to have the replicate panels be equivalent and to standardize the process to be the same for each of the eight panels in each workshop. Additionally, initial examination of cut scores and percentages seems to indicate that there might have been a facilitator effect. Two sets of analyses were conducted to address this concern. The first set of analyses (i.e., cut-score location) examined two dependent variables: the cut scores and the rank order of those cut scores for all panels within a subject area for a JSS session (ignoring the distinction between the two post-secondary activities within each session).For this analysis, the cut score used was the mean of the two table cut scores within a replicate panel. These results appear in Tables 4-1 and 4-2 in the column labeled "NAEP Cut Score Rank w/in Session."

The second set of analyses (i.e., cut-score convergence) examined three dependent variables: the cut score, the empirical standard error (EmpSE), and the bootstrapped standard error (BootSE). For this analysis, the cut score used was the median rather than the mean. The two versions of the cut score (mean and median) for the two analyses differ by less than one NAEP-scale score point. These minor discrepancies mean that the results do not agree exactly, but they are close. Results for each study are summarized such that the one-way ANOVAs include and exclude the pilot round results. Statistical significance (*p*-values) for facilitator effects are summarized in Tables 4-1 and 4-2.

Two kinds of analyses were conducted separately for mathematics and reading. First, separately for each round (as the observations across rounds within a session are not statistically independent), a one-way ANOVA was conducted to look for differences between means for the four facilitators within a subject area. Thus, each ANOVA had either 16 (including the pilot) or 12 (excluding the pilot) observations. These analyses did not entail any serious violations of statistical assumptions, but they had low power due to a very small number of data points. This same analysis was also performed to examine the difference between the round 1 and round 3 cut scores (labeled "Convergence" in Tables 4-1 and 4-2). Second, a regression model was run in which data across rounds were combined. For these analyses, the pilot data were excluded. Thus, there were 36 observations in each analysis. Dummy variables were entered for post-secondary activity, round, and the post-secondary activity by round interaction, using up 17 degrees of freedom and leaving 18 degrees of freedom to represent the contrasts between the two facilitators within each round by post-secondary activity combination. A second regression, with all of the previous variables together with dummy

variables for facilitators, was also run, and the R-square change (representing the main effect of a facilitator) was investigated for significance. These analyses violated statistical assumptions, because the rounds were still not statistically independent; entering dummy variables to remove the "main effect" of "round" did not fix this problem. Final-stage regressions using data from just one round at a time were rerun for the "standard error" analyses. For these analyses, there were 12 observations; after removing post-secondary activity there were six degrees of freedom remaining, and the test for the facilitator effect was an F with just three and three degrees of freedom. These are legitimate but have low power.

The results indicate that there are no detectable facilitator effects for the location of the final cut score unless those effects are in the direction of convergence. For reading, F ratios are extremely small (*p*-values are above 0.95). While it is possible that facilitators might compare results with one another and make suggestions to panelists that would cause them to move toward agreement, no such actions were reported by observers in any sessions. The process facilitators for replicate panels were generally paired across companies, and it seems highly unlikely that this sort of manipulation took place.

With regard to within-panel variability (reflected in the "EmpSE" and "BootSE" dependent variables), the pattern is a bit less clear. Here, pooling the regressions over rounds (which violates statistical assumptions) shows effects so strong that they cannot be given any credence. The within-round regressions, which have very low power, show effects at $p < .10$, however, especially for math and for round 2 in reading. The possibility of facilitator effects on the degree of panelist convergence might warrant further investigation, although the question is of little importance in the overall scheme of the conceptual and statistical issues surrounding JSS.

Table 4-1. *p*-Values from Tests for Facilitator Effects—Mathematics

	NAEP Cut Score Rank w/in Session	*Mean Cut Score (NAEP Scale)*	*Median Cut Score (NAEP Scale)*	*EmpSE*	*BootSE*
Round 1 ANOVA (Pilot round included)	0.2604	0.0624			
Round 2 ANOVA (Pilot round included)	0.0278*	0.0512			
Round 3 ANOVA (Pilot round included)	0.1187	0.2169			
Round 1 ANOVA (Pilot round excluded)	0.5957	0.3386	0.3489	0.4590	0.4116
Round 2 ANOVA (Pilot round excluded)	0.2170	0.1172	0.1153	0.0435*	0.0811
Round 3 ANOVA (Pilot round excluded)	0.2004	0.2285	0.2529	0.3528	0.3908
Convergence (Round 1 to Round 3)			0.4375	0.3926	0.4059
Regression Analysis (Pilot excluded) all rounds	0.1119	0.0522	0.0542	0.0005*	0.0005*
Round 1 (F test with 3 and 3 df)			0.6805	0.0676	0.1094
Round 2 (F test with 3 and 3 df)			0.1743	0.0877	0.0969
Round 3 (F test with 3 and 3 df)			0.7544	0.0759	0.0779

* Indicates statistical significance. Note: EmpSE = Empirical Standard Error; BootSE = Bootstrapped Standard Error

Table 4-2. *p*-Values from Tests for Facilitator Effects—Reading

	NAEP Cut Score Rank w/in Session	*Mean Cut Score (NAEP Scale)*	*Median Cut Score (NAEP Scale)*	*EmpSE*	*BootSE*
Round 1 ANOVA (Pilot round included)	0.6525	0.6937			
Round 2 ANOVA (Pilot round included)	0.4937	0.8777			
Round 3 ANOVA (Pilot round included)	0.5372	0.9168			
Round 1 ANOVA (Pilot round excluded)	0.8417	0.7734	0.7680	0.3665	0.2811
Round 2 ANOVA (Pilot round excluded)	0.9000	0.9692	0.9739	0.7207	0.6576
Round 3 ANOVA (Pilot round excluded)	0.9106	0.9441	0.9474	0.8437	0.8739
Convergence (Round 1 to Round 3)			0.6536	0.5553	0.5282
Regression Analysis (Pilot excluded) all rounds	0.9737	0.9438	0.9415	0.0043*	0.0076*
Round 1 (F test with 3 and 3 df)			0.9887	0.4963	0.4338
Round 2 (F test with 3 and 3 df)			0.9963	0.0570	0.0246*
Round 3 (F test with 3 and 3 df)			0.9868	0.2282	0.1770

* Indicates statistical significance. Note: EmpSE = Empirical Standard Error; BootSE = Bootstrapped Standard Error

4.2. Irrelevant Items

Based on feedback from panelists in prior JSS sessions and observations that panelists had difficulty with the content of the assessments, Computer Support Specialist and HVAC panelists participated in a special study to explore the utility of an alternative item map format within the context of these studies, . Panelists in prior JSS sessions seemed to consider many items as irrelevant for students to be minimally prepared for their training program or coursework, and some panelists identified entire content domains as irrelevant. Since, in previous bookmark-based standard setting studies for NAEP, the items had been grouped by content area on the item maps, item maps that grouped items by content were used in this special study.

Using reconfigured item maps, panelists participated in an exercise in which they identified where they would set their cut scores if given the opportunity to place a bookmark for each content domain (e.g., for mathematics, Number Properties and Operations, Measurement, Geometry) separately, as well as to identify items in their rating pools that they considered to be irrelevant for their training programs. Item maps were modified so that items from different content domains were differentiated by color and separated into columns within the item maps. When marking items that they considered irrelevant for their training programs, panelists were instructed to distinguish these items from those that assess relevant content at a more advanced level than required for a minimal level of preparedness to enter a job training program in this occupation. The details, including methodology and results, of this study are presented in the JSS Process Report (WestEd & Measured Progress, 2011).

REFERENCES

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.

Donoghue, J. R. (1997, March). *Item mapping to a weighted composite scale.* Paper presented at the session of the American Educational Research Association, Chicago, IL.

Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, Vol. 37, No. 1, pp. 36–48.

Maritz, J. S. & Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, Vol. 73, No. 361, pp. 194–196.

National Assessment Governing Board. (2009). *Making new links, 12th grade and beyond: Technical panel on 12th grade preparedness research final report.* Washington, D.C.: U.S. Department of Education.

National Assessment Governing Board. (2010). *Design document for 12th grade NAEP preparedness research judgmental standard setting studies*. Washington, D.C.: U.S. Department of Education.

WestEd & Measured Progress. (2011). *National Assessment of Educational Progress Grade 12 preparedness research project judgmental standard setting (JSS) studies: Process report.* San Francisco, CA: Authors.

APPENDICES

Redacted