National Assessment
Governing Board

National Assessment of Educational Progress Grade 12 Preparedness Research Project Judgmental Standard Setting (JSS) Studies

Submitted: November 28, 2011

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

PROCESS REPORT

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

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

Submitted by:
WestEd
730 Harrison Street
San Francisco, CA 94107
Phone: 415.615.3400

WestEd

Table of Contents

Executive Summary 1
- Overview 1
- Standard-Setting Studies 2
  - Postsecondary Areas 2
  - Panelists 3
  - Standard-Setting Process 3
- Standard-Setting Outcomes 5
- Procedural Validity 7
- Recommendations 8
  - Recruiting Procedures 8
  - Advance Webinars in the Development of BPDs 11
  - Facilitator Training 11
  - Standard-Setting Procedures 11
  - Handling of Irrelevant Items 13

Introduction 14
- Governing Board's Approach to Preparedness 15
- Judgmental Standard-Setting (JSS) Studies 16

Project Overview 21
- Study Staff and Observers 21
  - JSS-TAC. 21
  - Project Management Staff. 22
  - Content Facilitators. 22
  - Process Facilitators. 24
  - Observers 25
- Panelist Recruitment Plan 26
  - Panelist Recruitment for College-Preparedness Panels 27
  - Panelist Recruitment for Career-Preparedness Panels 32
- Judgmental Standard-Setting Process 39
  - Development of Borderline Performance Descriptions 40
  - Advance Materials 42

Computer-Aided Bookmarking (CAB) 44
Replicate Panels 45
Item Pool Division 46
NAEP-like Scales 48
Facilitator Training 48
Panelist Training 50
Bookmark Standard-Setting Design 51
Process Evaluations 68
Differences in Implementation Across JSS Sessions 68
Data Analysis 69
Pilot Study 72
Pilot Study Panelists 72
Pilot Study JSS Process 80
Pilot Study BPDs and Numerical Results 80
College-Preparedness Workshop Results 80
Automotive Master Technician Workshop Results 84
Pilot Study Exemplar Item Ratings 88
Pilot Study Process Evaluation Results 89
College-Preparedness Workshop Process Evaluation Results 90
Automotive Master Technician Workshop Process Evaluation Results 96
Operational Session 1 102
College-Preparedness Operational Workshop 107
College-Preparedness Operational Workshop Panelists 107
College-Preparedness Operational Workshop Numerical Results 111
College-Preparedness Operational Workshop Exemplar Item Ratings. 115
College-Preparedness Operational Workshop Process Evaluation Results 116
Automotive Master Technician Operational Workshop 122
Automotive Master Technician Operational Workshop Panelists 122
Automotive Master Technician Operational Workshop Numerical Results 125
Automotive Master Technician Operational Workshop Exemplar Item Ratings 129
Automotive Master Technician Operational Workshop Process Evaluation Results 130
Operational Session 2 136
LPN Operational Workshop 139

LPN Operational Workshop Panelists 139
LPN Operational Workshop Numerical Results 142
LPN Operational Workshop Exemplar Item Ratings 145
LPN Operational Workshop Process Evaluation Results 146
Pharmacy Technician Operational Workshop 153
Pharmacy Technician Operational Workshop Panelists 153
Pharmacy Technician Operational Workshop Numerical Results 156
Pharmacy Technician Operational Workshop Exemplar Item Ratings. 160
Pharmacy Technician Operational Workshop Process Evaluation Results 161
Operational Session 3 167
Computer Support Specialist Operational Workshop 169
Computer Support Specialist Operational Workshop Panelists 169
Computer Support Specialist Operational Workshop Numerical Results 172
Computer Support Specialist Operational Workshop Exemplar Item Ratings 175
Computer Support Specialist Operational Workshop Process Evaluation Results 176
HVAC Operational Workshop 183
HVAC Operational Workshop Panelists 183
HVAC Operational Workshop Numerical Results 186
HVAC Operational Workshop Exemplar Item Ratings 190
HVAC Operational Workshop Process Evaluation Results 191
Summary and Conclusions 197
Cut Scores and Percentages at or Above Cut Scores 198
Distribution of Cut Scores 202
Reliability Estimates for Cut Scores 203
Summary of Process Evaluations 207
Understanding of Tasks. 207
Understanding of the Borderline Performance Description 208
Comfort and Confidence. 210
Independence of Judgment 215
Helpfulness of Software 216
Special Study 219
Recommendations 225
Recruiting Procedures 225

Advance Webinars in the Development of BPDs 228
Facilitator Training 228
Standard-Setting Procedures 228
Handling of Irrelevant Items 230
References 231
Appendix A: JSS Session Agendas A-1
Appendix B: College-Preparedness Postsecondary Sample Nomination Email B-1
Appendix C: College-Preparedness Postsecondary Sample Nomination Form C-1
Appendix D: College-Preparedness Sample Recruitment Informational Form D-1
Appendix E: College-Preparedness Secondary-Level Sample Nomination Email E-1
Appendix F: College-Preparedness Secondary-Level Sample Nomination Form. F-1
Appendix G: Job-Training Sample Nomination Email and Additional Detail G-1
Appendix H: Job-Training Sample Nomination Form H-1
Appendix I: Job-Training Sample Recruitment Informational Form I-1
Appendix J: JSS Session Briefing Booklets J-1
Appendix K: JSS Session Facilitator Handbooks K-1
Appendix L: JSS Session Orientation Presentation Slides L-1
Appendix M: Pilot Study Process Evaluation Results M-1
Appendix N: U.S. Census Bureau Census Regions N-1
Appendix O: Borderline Performance Descriptions O-1
Appendix P: College-Preparedness Process Evaluation Results P-1
Appendix Q: Automotive Master Technician Process Evaluation Results Q-1
Appendix R: LPN Process Evaluation Results R-1
Appendix S: Pharmacy Technician Process Evaluation Results S-1
Appendix T: Computer Support Specialist Process Evaluation Results T-1
Appendix U: HVAC Process Evaluation Results U-1
Appendix V: Special Study Panelist Instructions and Process Evaluation Form V-1

List of Tables

Table 1. Final Results for Mathematics 6

Table 2. Final Results for Reading 7

Table 3. Number of Initially Identified Programs by Occupation 34

Table 4. JSS Sessions and Workshops 39

Table 5. Summary of Panel Item Pools, Mathematics 47

Table 6. Summary of Panel Item Pools, Reading 47

Table 7. Pilot Study: College-Preparedness Postsecondary Institution Distributions 73

Table 8. Pilot Study: College-Preparedness Secondary-Level Institution Distributions 74

Table 9. Pilot Study: Panelist Distribution by Institution Type 76

Table 10. Pilot Study: Panelist Distribution by Demographic Characteristics 77

Table 11. Pilot Study: Geographic Distribution of Panelists 78

Table 12. Pilot Study: Student Populations Served by Job-Training Panelists 79

Table 13. Pilot Study: Abbreviations Used to Describe Panels 80

Table 14. Pilot Study: College-Preparedness Round 1 Results 81

Table 15. Pilot Study: College-Preparedness Round 2 Results 81

Table 16. Pilot Study: College-Preparedness Round 3 Results 82

Table 17. Pilot Study: College-Preparedness Round-to-Round Cut Score Changes by Panel 83

Table 18. Pilot Study: College-Preparedness Comparison of Cut Scores Based on Medians and Means 84

Table 19. Pilot Study: Automotive Master Technician Round 1 Results 85

Table 20. Pilot Study: Automotive Master Technician Round 2 Results 85

Table 21. Pilot Study: Automotive Master Technician Round 3 Results 86

Table 22. Pilot Study: Automotive Master Technician Round-to-Round Cut Score Changes by Panel 88

Table 23. Pilot Study: Automotive Master Technician Comparison of Cut Scores Based on Medians and Means 88

Table 24. Pilot Study: Exemplar Item Summary 89

Table 25. Pilot Study: Summary of Selected Evaluation Items by Panel 91

Table 26. Operational Session 1: Abbreviations Used to Describe Panels 106

Table 27. College-Preparedness Operational Workshop: Postsecondary Institution Distributions 108

Table 28. College-Preparedness Operational Workshop: Secondary-Level Institution Distributions 109

Table 29. College-Preparedness Operational Workshop: Panelist Distribution by Institution Type 110

Table 30. College-Preparedness Operational Workshop: Panelist Distribution by Demographic Characteristics 110

Table 31. College-Preparedness Operational Workshop: Geographic Distribution of Panelists 111

Table 32. College-Preparedness Operational Workshop: Round 1 Results 112

Table 33. College-Preparedness Operational Workshop: Round 2 Results 112

Table 34. College-Preparedness Operational Workshop: Round 3 Results 113

Table 35. College-Preparedness Operational Workshop: Round-to-Round Cut Score Changes by Panel 114

Table 36. College-Preparedness Operational Workshop: Comparison of Cut Scores Based on Medians and Means 115

Table 37. College-Preparedness Operational Workshop: Exemplar Item Summary 115

Table 38. College-Preparedness Operational Study: Summary of Selected Evaluation Items by Panel 117

Table 39. Automotive Master Technician Operational Workshop: Panelist Distribution by Institution Type 123

Table 40. Automotive Master Technician Operational Workshop: Panelist Distribution by Demographic Characteristics 124

Table 41. Automotive Master Technician Operational Workshop: Geographic Distribution of Panelists 124

Table 42. Automotive Master Technician Operational Workshop: Student Populations Served by Panelists 125

Table 43. Automotive Master Technician Operational Workshop: Round 1 Results 126

Table 44. Automotive Master Technician Operational Workshop: Round 2 Results 126

Table 45. Automotive Master Technician Operational Workshop: Round 3 Results 127

Table 46. Automotive Master Technician Operational Workshop: Round-to-Round Cut Score Changes by Panel 129

Table 47. Automotive Master Technician Operational Workshop: Comparison of Cut Scores Based on Medians and Means.....129

Table 48. Automotive Master Technician Operational Workshop: Exemplar Item Summary ...130

Table 49. Automotive Master Technician Operational Workshop: Summary of Selected Evaluation Items by Panel 131

Table 50. Operational Session 2: Abbreviations Used to Describe Panels.....138

Table 51. LPN Operational Workshop: Panelist Distribution by Institution Type.....140

Table 52. LPN Operational Workshop: Panelist Distribution by Demographic Characteristics.140

Table 53. LPN Operational Workshop: Geographic Distribution of Panelists.....141

Table 54. LPN Operational Workshop: Student Populations Served by Panelists.....141

Table 55. LPN Operational Workshop: Round 1 Results.....142

Table 56. LPN Operational Workshop: Round 2 Results.....143

Table 57. LPN Operational Workshop: Round 3 Results.....143

Table 58. LPN Operational Workshop: Round-to-Round Cut Score Changes by Panel.....145

Table 59. LPN Operational Workshop: Comparison of Cut Scores Based on Medians and Means 145

Table 60. LPN Operational Workshop: Exemplar Item Summary.....146

Table 61. LPN Operational Workshop: Summary of Selected Evaluation Items by Panel.....147

Table 62. Pharmacy Technician Operational Workshop: Panelist Distribution by Institution Type 154

Table 63. Pharmacy Technician Operational Workshop: Panelist Distribution by Demographic Characteristics 154

Table 64. Pharmacy Technician Operational Workshop: Geographic Distribution of Panelists 155

Table 65. Pharmacy Technician Operational Workshop: Student Populations Served by Panelists 156

Table 66. Pharmacy Technician Operational Workshop: Round 1 Results 157

Table 67. Pharmacy Technician Operational Workshop: Round 2 Results 157

Table 68. Pharmacy Technician Operational Workshop: Round 3 Results 158

Table 69. Pharmacy Technician Operational Workshop: Round-to-Round Cut Score Changes by Panel 160

Table 70. Pharmacy Technician Operational Workshop: Comparison of Cut Scores Based on Medians and Means 160

Table 71. Pharmacy Technician Operational Workshop: Exemplar Item Summary 161

Table 72. Pharmacy Technician Operational Workshop: Summary of Selected Evaluation Items by Panel 162

Table 73. Operational Session 3: Abbreviations Used to Describe Panels 168

Table 74. Computer Support Specialist Operational Workshop: Panelist Distribution by Institution Type 170

Table 75. Computer Support Specialist Operational Workshop: Panelist Distribution by Demographic Characteristics 170

Table 76. Computer Support Specialist Operational Workshop: Geographic Distribution of Panelists 171

Table 77. Computer Support Specialist Operational Workshop: Student Populations Served by Panelists 171

Table 78. Computer Support Specialist Operational Workshop: Round 1 Results 172

Table 79. Computer Support Specialist Operational Workshop: Round 2 Results 173

Table 80. Computer Support Specialist Operational Workshop: Round 3 Results 173

Table 81. Computer Support Specialist Operational Workshop: Round-to-Round Cut Score Changes by Panel 175

Table 82. Computer Support Specialist Operational Workshop: Comparison of Cut Scores Based on Medians and Means 175

Table 83. Computer Support Specialist Operational Workshop: Exemplar Item Summary 176

Table 84. Computer Support Specialist Operational Workshop: Summary of Selected Evaluation Items by Panel 177

Table 85. HVAC Operational Workshop: Panelist Distribution by Institution Type 184

Table 86. HVAC Operational Workshop: Panelist Distribution by Demographic Characteristics 185

Table 87. HVAC Operational Workshop: Geographic Distribution of Panelists 185

Table 88. HVAC Operational Workshop: Student Populations Served by Panelists 186

Table 89. HVAC Operational Workshop: Round 1 Results 187

Table 90. HVAC Operational Workshop: Round 2 Results 187

Table 91. HVAC Operational Workshop: Round 3 Results 188

Table 92. HVAC Operational Workshop: Round-to-Round Cut Score Changes by Panel 190

Table 93. HVAC Operational Workshop: Comparison of Cut Scores Based on Medians and Means 190

Table 94. HVAC Operational Workshop: Exemplar Item Summary 191

Table 95. HVAC Operational Workshop: Summary of Selected Evaluation Items by Panel 192

Table 96. Summary and Conclusions: Panel Identifications 197

Table 97. Summary of Cut Score Changes Across Rounds for Mathematics 203

Table 98. Summary of Cut Score Changes Across Rounds for Reading 203

Table 99. Standard Error Estimates and Confidence Intervals for Mean Cut Scores by Postsecondary Area, Mathematics 204

Table 100. Standard Error Estimates and Confidence Intervals for Mean Cut Scores by Postsecondary Area, Reading 205

Table 101. Special Study: Content Domain Cut Scores for Mathematics 221

Table 102. Special Study: Content Domain Cut Scores for Reading 221

Table 103. Special Study: Number of Items Deemed Irrelevant, Mathematics 222

Table 104. Special Study: Number of Items Deemed Irrelevant, Reading.....222

Table 105. Special Study: Process Evaluation Results.....223

List of Figures

Figure 1. Panel A Room Configuration 46
Figure 2. Sessions 50
Figure 3. Basic Structure of the Judgmental Standard-Setting Process 53
Figure 4. Sample CR Item as Presented in the CAB 57
Figure 5. Conceptual Representation of the OIB with CR Items 58
Figure 6. Sample Item Map 60
Figure 7. Sample Rater Location Chart from the CAB 63
Figure 8. Sample Booklet Score Chart 65
Figure 9. Sample Consequences Data Feedback 66
Figure 10. Pilot Study: College-Preparedness Mean Absolute Deviation (MAD) of Cut Scores by Panel 83
Figure 11. Pilot Study: Automotive Master Technician Mean Absolute Deviation (MAD) of Cut Scores by Panel 87
Figure 12. Pilot Study College-Preparedness: At the time I placed my bookmark, my understanding of the BPD was 92
Figure 13. Pilot Study College-Preparedness: The most accurate description of my level of confidence in my bookmark placement is 93
Figure 14. Pilot Study College-Preparedness: I believe my cut score is consistent with the BPD 94
Figure 15. Pilot Study Automotive Master Technician: At the time I placed my bookmark, my understanding of the BPD was 97
Figure 16. Pilot Study Automotive Master Technician: The most accurate description of my level of confidence in my bookmark placement is 98
Figure 17. Pilot Study Automotive Master Technician: I believe my cut score is consistent with the BPD 99
Figure 18. College-Preparedness Operational Workshop: Mean Absolute Deviation (MAD) of Cut Scores by Panel 113
Figure 19. College-Preparedness Operational Workshop: At the time I placed my bookmark, my understanding of the BPD was 118
Figure 20. College-Preparedness Operational Workshop: The most accurate description of my level of confidence in my bookmark placement is 119
Figure 21. College-Preparedness Operational Workshop: I believe my cut score is consistent with the BPD 120
Figure 22. Automotive Master Technician Operational Workshop: Mean Absolute Deviation (MAD) of Cut Scores by Panel 128

Figure 23. Automotive Master Technician Operational Workshop: At the time I placed my bookmark, my understanding of the BPD was 132

Figure 24. Automotive Master Technician Operational Workshop: The most accurate description of my level of confidence in my bookmark placement is 133

Figure 25. Automotive Master Technician Operational Workshop: I believe my cut score is consistent with the BPD.....134

Figure 26. LPN Operational Workshop: Mean Absolute Deviation (MAD) of Cut Scores by Panel.....144

Figure 27. LPN Operational Workshop: At the time I placed my bookmark, my understanding of the BPD was 148

Figure 28. LPN Operational Workshop: The most accurate description of my level of confidence in my bookmark placement is 149

Figure 29. LPN Operational Workshop: I believe my cut score is consistent with the BPD 150

Figure 30. Pharmacy Technician Operational Workshop: Mean Absolute Deviation (MAD) of Cut Scores by Panel.....159

Figure 31. Pharmacy Technician Operational Workshop: At the time I placed my bookmark, my understanding of the BPD was 163

Figure 32. Pharmacy Technician Operational Workshop: The most accurate description of my level of confidence in my bookmark placement is 164

Figure 33. Pharmacy Technician Operational Workshop: I believe my cut score is consistent with the BPD.....165

Figure 34. Computer Support Specialist Operational Workshop: Mean Absolute Deviation (MAD) of Cut Scores by Panel.....174

Figure 35. Computer Support Specialist Operational Workshop: At the time I placed my bookmark, my understanding of the BPD was 178

Figure 36. Computer Support Specialist Operational Workshop: The most accurate description of my level of confidence in my bookmark placement is 179

Figure 37. Computer Support Specialist Operational Workshop: I believe my cut score is consistent with the BPD.....180

Figure 38. HVAC Operational Workshop: Mean Absolute Deviation (MAD) of Cut Scores by Panel.....189

Figure 39. HVAC Operational Workshop: At the time I placed my bookmark, my understanding of the BPD was 193

Figure 40. HVAC Operational Workshop: The most accurate description of my level of confidence in my bookmark placement is 194

Figure 41. HVAC Operational Workshop: I believe my cut score is consistent with the BPD.....195

Figure 42. Cut Scores from Three Rounds of Ratings for Mathematics 199

Figure 43. Cut Scores from Three Rounds of Ratings for Reading.....200
Figure 44. Percents at or Above the Cut Scores from Round 3.....201
Figure 45. Mean Cut Scores and Confidence Intervals by Postsecondary Area, Mathematics...205
Figure 46. Mean Cut Scores and Confidence Intervals by Postsecondary Area, Reading.....206
Figure 47. My understanding of the tasks I was to accomplish during each round was 208
Figure 48. At the time I placed my bookmark, my understanding of the BPD was 209
Figure 49. I believe my cut score is consistent with the BPD 211
Figure 50. The most accurate description of my level of confidence in the cut score recommendations I provided was 213
Figure 51. I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark.....214
Figure 52. The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items.....215
Figure 53. I felt pressured by others in my group to make my cut score recommendation agree with theirs.....216
Figure 54. During the standard setting process, I found using CAB to be 217
Figure 55. Special Study: Sample Modified Item Map 220

Executive Summary

Overview

This report describes the process and outcomes of a series of judgmental standard-setting (JSS) studies conducted in both mathematics and reading to establish preparedness reference points—representing the academic performance required for placement in an entry-level, credit-bearing college course in the content area or for enrollment in a postsecondary job-training program—on the NAEP reporting scale for the two content areas. Four JSS sessions—a pilot study and three operational sessions—were held at the Westin St. Louis Hotel in St. Louis, Missouri, from April to July of 2011. The goals of each session were as follows:

- Finalize the descriptions of minimal knowledge and skills that describe what students need to know and be able to do to be prepared for placement in a college course and/or a job-training course; and
- Determine the score on the NAEP scale that corresponds to the level of performance at the borderline (the cut score) and the percentage of students performing above the cut score.

Additionally, test items illustrative of what students performing above the cut score know and can do were selected. In order to maximize standardization of the JSS process across the postsecondary and content areas involved, the Governing Board developed a Design Document (National Assessment Governing Board, 2010a) that guided all aspects of the project's implementation.

This report includes a description of the processes of and conclusions drawn from a pilot study (designed to evaluate the proposed methodology, logistics, training, and materials in advance of

the operational studies), a series of operational standard-setting workshops informed by findings from the pilot study, and a special study implemented at the end of the final operational session. All studies were conducted by WestEd, in partnership with Measured Progress and the Educational Policy Improvement Center (EPIC) and under contract with the National Assessment Governing Board (Governing Board). The descriptions of borderline performance, recommendations of cut scores, and recommendations of exemplar items that emerged from these studies will be submitted to the Governing Board for consideration.

Additional information about the technical decisions and computational procedures that were implemented in this series of JSS studies may be found in a separate Technical Report (Measured Progress & WestEd, 2011).

Standard-Setting Studies

The JSS pilot study session (comprising two workshops: a college-preparedness workshop and an occupation workshop) was held on April 26–29, 2011, while the three operational sessions (each session comprising two workshops: either a college-preparedness workshop and an occupation workshop or two occupation workshops) were held on May 24–27, June 7–10, and June 28–July 1, 2011. All sessions were held at the Westin St. Louis Hotel in St. Louis, Missouri. Agendas for the pilot study and the three operational JSS sessions are provided in Appendix A.

Postsecondary Areas. The focus of this project was on six postsecondary areas: preparedness for entry into college—more specifically, for placement into entry-level, credit-bearing college courses that meet general education requirements without the need for remedial coursework in mathematics or reading—and preparedness for entry into job-training programs in five exemplar

occupations: Automotive Master Technician; Computer Support Specialist; Heating, Ventilation, and Air Conditioning (HVAC); Licensed Practical and Licensed Vocational Nurse (LPN); and Pharmacy Technician.

Panelists. The Design Document developed by the Governing Board for this project guided the recruitment of panelists. WestEd proposed and the Governing Board approved additions to the sampling plans for recruiting college-preparedness and career-preparedness panelists. For each postsecondary area, recruitment followed a two-tiered process: (1) identification of eligible institutions and solicitation of nominees from qualified individuals within these institutions, and (2) recruitment of panelists from the pool of nominees. Given the distinct nature of each of the postsecondary areas, sampling plans and recruitment efforts were tailored to each in order to recruit the most qualified panelists, as described in relevant sections of this report.

Standard-Setting Process. A modified bookmark standard-setting method was used for the pilot study and the three operational JSS sessions. The specific process used was developed for NAEP achievement level setting by ACT, Inc. (ACT, Inc., 2007; ACT, Inc., 2010). Within this process, panelists reviewed assessment items that were ordered by difficulty based on item mapping using a response probability (RP) criterion of 0.67, starting with the easiest item and progressing to the most difficult. They evaluated each item against a description of borderline performance until they determined where a minimally prepared student could no longer respond correctly with a 0.67 probability. A bookmark was placed immediately preceding that item to locate the cut score. Within each panel, individuals' cut scores were used to compute the median, which served as the panel's cut score.

The bookmark method was modified in the following ways to be consistent with bookmark-method implementation in previous NAEP achievement-level-setting studies:

- Panelists were provided with actual test booklets to show examples of student performance on the assessment at and above the cut scores; and
- A spatially representative display of items on a student achievement scale was given to panelists to accompany the ordered item booklets used to place the cut scores.

The bookmark process was implemented using two independent replicate panels for each postsecondary area within each content area in an attempt to estimate the reliability of the replicate panels' cut scores. The item pools were divided into two comparable parts, with some overlap between the two parts, which were assigned to replicate panels. Panelists were selected and assigned to panels to create replicate panels that were as equivalent as possible.

In addition, Computer-Aided Bookmarking (CAB) software was developed and used for this project's standard setting. The computerization of the standard-setting process increased the efficiency of operations by reducing the time required for panelists to complete most steps in the process and for data analysis to be completed, allowing for greater efficiency in providing feedback to panelists between rounds of bookmarking.

The development of borderline performance descriptions (BPDs) was central to the standard-setting process, as BPDs describe the performance required for minimal preparedness in each postsecondary area and are, therefore, the statements of the performance standard to be represented on the NAEP scale by each cut score. For this study, the BPDs for each postsecondary area were developed by panelists recruited from within that area through an iterative three-step process: (1) panelists participated in online orientation webinars, which

introduced important background on the project, the NAEP framework, and the process by which the BPDs would be developed; (2) panelists completed online Content Objectives forms, in which they reviewed each objective for the NAEP framework and indicated whether the knowledge and skills reflected in the objectives were required for a student to be minimally prepared for entry into that postsecondary area's course or program, and provided sample course texts to content facilitators; and (3) content facilitators used the information collected from the panelists' review of the frameworks and examples of texts and tasks to draft preliminary BPDs, which were reviewed and refined by panelists throughout the in-person standard-setting process.

Standard-Setting Outcomes

Cut scores set by all replicate panels during the third and final round of ratings fell between 165 and 201 on the NAEP scale for mathematics, which has a maximum score of 300, and between 288 and 321 on the NAEP scale for reading, which has a maximum score of 500. For each content area, the difference between the lowest and highest cut score was near one standard deviation. The first tables present the final cut scores set by each panel, along with the percentages of high school students in 2009 who would have scored at or above each cut score. All cut scores are presented on the NAEP scale for the appropriate content. For each content area, approximately half of the cut scores are above the NAEP Proficient cut score. Additionally, an independent samples t-test was conducted for each postsecondary activity and content area mean cut score. Statistical significance between each postsecondary area's Panel A and B mean cut scores was investigated, and results of the significance test are presented in Tables 1 and 2. It is important to note that the final cut scores used by NAEP were based on the median (to limit the effects of outliers), and t-tests were conducted on means. Therefore, results should be interpreted with caution. Overall, the resulting cut scores reflect rather large variability across

replicate panels despite efforts to maintain equivalent standard-setting processes across postsecondary areas within the two content areas, although further analysis (detailed in the Technical Report) indicates that there are no detectable facilitator effects for the location of the final cut scores.

Table 1. Final Results for Mathematics

Postsecondary Activity	Panel	Number of Panelists	Median Cut Score	Percentage at or Above the Cut Score	Mean Cut Score	Significant Difference between Mean Cut Scores ($p \leq 0.05$)
College-Preparedness	A	10	201	8.0	199.5	Yes
	B	10	189	15.5	187.4	
Automotive Master Technician	A	7	167	36.1	172.4	No
	B	8	171	31.9	171.6	
Computer Support Specialist	A	10	165	38.1	163.9	Yes
	B	10	185	18.7	187.1	
Heating, Ventilation, and Air Conditioning	A	10	177	25.8	182.5	No
	B	9	172	31.0	175.7	
Licensed Practical Nurse	A	10	177	25.8	177.5	Yes
	B	10	193	12.7	191.7	
Pharmacy Technician	A	9	174	28.9	174.8	No
	B	9	176	26.9	173.0	

Note. The NAEP mathematics achievement level cut scores are set at 141 for Basic, 176 for Proficient, and 216 for Advanced.

Table 2. Final Results for Reading

Postsecondary Activity	Panel	Number of Panelists	Median Cut Score	Percentage at or Above the Cut Score	Mean Cut Score	Significant Difference between Mean Cut Scores ($p \leq 0.05$)
College-Preparedness	A	10	290	51.6	290.3	Yes
	B	9	304	36.4	305.1	
Automotive Master Technician	A	5	308	32.4	317.2	No
	B	6	294	47.3	296.5	
Computer Support Specialist	A	10	292	49.6	291.9	Yes
	B	10	307	33.5	304.4	
Heating, Ventilation, and Air Conditioning	A	10	289	52.9	290.4	No
	B	9	292	49.6	290.4	
Licensed Practical Nurse	A	10	307	33.5	309.9	Yes
	B	10	288	54.1	286.1	
Pharmacy Technician	A	10	321	20.2	319.5	Yes
	B	9	299	41.9	299.1	

Note. The NAEP reading achievement level cut scores are set at 265 for Basic, 302 for Proficient, and 346 for Advanced.

Procedural Validity

At the end of each bookmarking round and each day, panelists were provided with an evaluation form designed to assess their understanding of instructions, tasks, and materials. There were a total of five questionnaires administered over the course of each meeting. Most responses were collected on Likert scales, but several responses were narratives that addressed specific aspects of the process. These evaluations were reviewed at the end of each day, and any sources of confusion or misunderstanding were identified for clarification with individual panelists or the group as a whole. Selected results from these evaluations are presented in the body of this process report.

Overall, panelists indicated that they understood their tasks and the materials, felt comfortable and confident in making their decisions, felt free to make independent judgments, and found the

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

Recommendations

This set of standard-setting studies is an important component of the Governing Board's larger postsecondary preparedness initiative, and the focus on career-preparedness activities provides timely and useful information that will inform discussions surrounding the degree of overlap between preparedness for college and preparedness for the workplace. The methodology prescribed by the Governing Board, as described in the Design Document and implemented by Measured Progress, was thorough and comprehensive. Despite the rigor of the study design and its implementation, however, certain challenges arose, in particular within career-preparedness panels. In response to those challenges, the following lessons learned and recommendations are submitted for future standard-setting studies of this type.

Recruiting Procedures. The following section begins with a description of challenges related to the recruiting procedures implemented for college-preparedness and career-preparedness panelists and suggests approaches to improve this process in future studies.

College-Preparedness Panelists. Recruitment of postsecondary panelists for this project's college-preparedness panels was largely successful, with a substantial pool of qualified candidates from which to select optimal panelists. It is likely that offering to provide an honorarium to panelists assisted in recruitment; however, several potential candidates representing prestigious four-year postsecondary institutions declined to participate, stating that the honorarium was less than they would typically earn for consulting work. Recruitment of

secondary-level panelists proved somewhat more difficult, perhaps in part due to the timing of the pilot study and the first operational session (late April and late May, respectively—i.e., toward the end of the academic year), although sufficient numbers of qualified secondary-level panelists were recruited for the college-preparedness panels.

Career-Preparedness Panelists. Recruitment of job-training instructors to serve on certain career-preparedness panels was more difficult, for various possible reasons. Across the occupations, most job-training program heads and instructors were not familiar with NAEP and the type of activity entailed in standard setting studies; successfully explaining the importance, purpose, and approach of this type of study proved more difficult than when recruiting instructors from more traditional academic programs. The timing of the first and second operational workshops coincided with the end of the academic year for many, a difficult time to be away from classes; also, within at least some job-training programs, authorization to take a full week (four days for each workshop, plus a day for travel) away from classes appeared to be difficult to obtain. In addition, based on correspondence with several nominees, it is not uncommon for job-training instructors in some occupations (e.g., Automotive Master Technician, Computer Support Specialist, HVAC) to also work as practicing technicians, thus making it even more difficult to commit to the amount of time required for each workshop.

WestEd submits the following lessons that it learned through this study's recruitment process for consideration by the Governing Board and/or its contractors when recruiting for future studies.

- *Consider variability between occupations in responsiveness to recruitment efforts.* The Governing Board thoroughly and systematically reviewed a pool of potential occupations when selecting the five exemplar occupations to be included in this study, considering, among other factors, the availability of eligible programs and panelists. Even though the

number of formal job-training programs varied by occupation, all exemplar occupations seemed likely to produce the requisite numbers of panelists. Through its recruitment efforts, however, WestEd discovered that occupations varied dramatically in how they train their workforces, how their job-training programs are accredited or certified, and how communication flows to and among job-training programs; it also found that response rates differed considerably among job-training programs. For future studies involving occupations, advance planning and research is needed to estimate the amount of time required for successful recruitment to be completed.

- *Streamline nomination materials.* During initial recruitment efforts, WestEd used an introductory letter used for recruiting from typical academically focused audiences to request nominations from the heads of job-training programs. While these materials were effective in recruiting college-preparedness panelists, they were less appealing to job-training instructors. Therefore, during the recruitment process, WestEd transitioned to more graphical, streamlined materials. Response to the streamlined materials was greater.
- *Ensure that panelists' eligibility requirements are appropriate to the task.* It became apparent through the standard-setting process that some panelists in some of the occupational workshops lacked the content knowledge and skills to effectively interact with the NAEP content, particularly in mathematics. While panelists were required to be familiar with the content-specific knowledge, skills, and abilities required in their programs, they were not required to teach courses specifically addressing either content area. Across occupations, it is not reasonable to expect to find educators who teach reading-specific courses; however, in some occupations it is common for job-training programs to include mathematics-specific courses (such as Math for Pharmacy

Technicians). When recruiting from occupations that offer such courses in their job-training programs, it may be advisable to target recruitment to job-training instructors who teach those courses.

Advance Webinars in the Development of BPDs. The use of online webinars for training panelists and engaging them in the development of preliminary BPDs in advance of the standard-setting sessions was an innovation for NAEP standard setting. An evaluation of the effectiveness of this approach could inform its use in future standard-setting studies.

Facilitator Training. Given the scheduling of the JSS sessions, a total of eight process facilitators were required to participate in each session; these facilitators were selected from three organizations: WestEd, Measured Progress, and EPIC. While all facilitators had the requisite qualifications for conducting standard-setting on the NAEP assessments in mathematics and reading, they reflected the somewhat different styles of their organizations, and addressing these styles while the standard-setting sessions were in process posed some challenges. If future studies involve working with a large group of facilitators, it would be advisable to provide extensive training in advance of standard-setting and/or recruit all facilitators from the same organization.

Standard-Setting Procedures. The bookmark method stipulated by the Design Document worked well for the college-preparedness pilot study and operational workshop. Panelists for the college-preparedness workshops came from traditional college and secondary-level academic programs and were, as a whole, relatively familiar with NAEP and with the type of activities required of them. Recruitment of these panelists and implementation of the standard-setting process proceeded largely as planned. However, on the whole, job-training instructors recruited

for the career-preparedness workshops were less familiar with the objectives and structure of NAEP and with standard setting in general. As a group, they tended to struggle more than the college-preparedness panelists with the language of the NAEP frameworks and with the pools of NAEP items assigned to them. In addition, panelists from some occupations were not well versed in the academic prerequisites for their occupations. The diversity among occupations, panel groups, and panelists posed unique challenges to the implementation of the bookmark method as planned within the job-training workshops.

In response to the challenges represented by these lessons learned, modifications to the JSS process were made over the course of the pilot study and the three operational sessions, which yielded a refined implementation by the third operational session.

- The instructions and guidance provided to panelists when identifying knowledge, skills, and abilities (KSAs) for the items were refined through the pilot study and the first two operational sessions. It is recommended that item descriptions be provided to panelists—especially panelists not recruited from traditional academic programs—for use in the process of developing KSA annotations for future studies.

- The Design Document called for the sharing of content facilitators across pairs of replicate panels, and this design seemed appropriate for the college-preparedness panels. However, it is recommended that, for future standard-setting involving occupations, each job-training panel group be assigned its own content facilitator. Assigning a content facilitator full-time to a panel will allow more time and opportunities for panelists to seek guidance and consultation regarding content-related issues, such as KSAs.

- The decision to include secondary-level teachers in the final JSS session (for the Computer Support Specialist and HVAC workshops) was made to increase the content knowledge and skills represented among the panelists. However, in some workshops, the secondary-level instructors became too influential in establishing the content areas' BPDs. The inclusion of such instructors in the process should be carefully considered and their roles explicitly communicated in future studies.

- Across all workshops, the computerization of aspects of the standard-setting process proved successful for this project, and the continued use of computerized procedures in future studies is recommended.

Handling of Irrelevant Items. A number of panelists across the pilot study and the first two operational JSS studies reported NAEP items to be irrelevant to their job-training programs; therefore, the Governing Board requested and designed a special study to be implemented on the last day of the third operational JSS session to explore this issue. A systematic strategy for instructing panelists on how to rate seemingly irrelevant items, drawing upon information gleaned from this special study, is recommended for future standard-setting studies of this nature.

Introduction

Preparing students for postsecondary success—in college, in the workplace, and/or in the military—is a growing objective of the K–12 educational system and, in turn, of entities, such as the National Assessment Governing Board (Governing Board), that are tasked with evaluating the progress of student achievement. For over two decades, the 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 NAEP as a tool to predict students' academic preparedness for entry into postsecondary education or job-training programs, forming a Technical Panel on 12th Grade Preparedness Research (Technical Panel) that was tasked with assisting the Governing Board in planning relevant research and validity studies (National Assessment Governing Board, 2009). The Technical Panel contributed to the working definition of academic preparedness established by the Governing Board and recommended a multi-method approach to exploring the feasibility of reporting postsecondary preparedness on the 2009 Grade 12 NAEP scale for mathematics and reading. Four specific design methodologies were proposed, representing a balance between qualitative and quantitative studies:

- Content alignment studies between NAEP and assessments that are currently used as predictors of postsecondary preparedness;
- Statistical relationship studies with assessments that serve as measures of preparedness for college and job-training programs, as well as with postsecondary outcomes data (such as transcripts and employment outcomes);
- Criterion-based judgmental standard-setting (JSS) studies to identify reference points on the NAEP scale that indicate academic preparedness for admission into entry-level

general education college courses or job-training programs in occupations selected by the Governing Board; and

- National surveys of postsecondary institutions regarding the assessments used for course placement and the cut scores on these assessments used to identify the need for remediation.

The Governing Board began implementing these design methodologies in 2009. Content alignment studies—which compared the 2009 Grade 12 NAEP in mathematics and reading with ACCUPLACER, ACT, SAT, and WorkKeys—were completed in 2010, statistical relationship data were gathered in 2010 and 2011, and national surveys were developed in 2011. This report addresses the series of criterion-based JSS studies on the 2009 Grade 12 NAEP in mathematics and reading, which were conducted by WestEd under contract with the Governing Board. In implementing the studies, WestEd subcontracted with Measured Progress and the Educational Policy Improvement Center (EPIC) to oversee the standard-setting process and the development of borderline performance descriptions, respectively.

Governing Board's Approach to Preparedness

The Governing Board has focused its conceptualization of twelfth-grade preparedness on academic qualifications and does not propose to address a range of behavioral and attitudinal aspects of student performance in postsecondary activities that are not measured by NAEP (e.g., time-management skills, diligence). The Governing Board has further limited its definition of postsecondary preparedness to the academic skills required for placement in entry-level, credit-bearing college courses that count toward a four-year undergraduate degree, or for

placement into military or civilian job-training programs,[1] with no prediction of success in such college-level courses or job-training programs.

The working definitions of academic preparedness adopted by the Governing Board and used throughout the current studies follow:

- *Preparedness for college* refers to the mathematics and reading knowledge and skills necessary to qualify for placement into entry-level college credit courses that meet general education requirements without the need for remedial coursework in mathematics or reading.
- *Preparedness for workplace* refers to the reading and mathematics knowledge and skills needed to qualify for an occupation's job-training program; it does not necessarily mean that the qualifications to be hired for a job have been met. (National Assessment Governing Board, 2009)

Judgmental Standard-Setting (JSS) Studies

The objective of the JSS studies was to produce reference points, or cut scores, on the NAEP scale that represent academic preparedness for entry into credit-bearing college courses and for entry into job-training programs within five exemplar occupations selected by the Governing Board. For the purposes of this project, the Governing Board selected for inclusion the following occupations, as defined in the U.S. Department of Labor/Employment and Training Administration's Occupational Information Network (O*NET) database (http://www.onetonline.org):

[1] This conceptualization assumes that similar jobs in the military and civilian sectors require approximately similar academic skills and knowledge.

- Automotive Master Technician (O*NET Code 49-3023.01);[2]
- Computer Support Specialist (O*NET Code 15-1041.00);
- Heating, Ventilation, and Air Conditioning (HVAC) (O*NET Code 49-9021.01);[3]
- Licensed Practical and Licensed Vocational Nurse (LPN) (O*NET Code 29-2061.00); and
- Pharmacy Technician (O*NET Code 29-2052.00).

The Governing Board used the following criteria, as recommended in the Technical Panel's Final Report (National Assessment Governing Board, 2009), for selecting these occupations:

- Representation of O*NET Zones 2 & 3 (which require at least three months of training but less than a bachelor's degree);
- Civilian and military job-training counterparts;
- Broad coverage of industry sectors;
- Understanding within the public of the occupations' general functions, duties, and responsibilities;
- Positive employment level both currently and projected in the future;
- Coverage of reading and mathematics preparedness, representing a range of reading and mathematics skills along the NAEP scale; and

[2] O*NET refers to this occupation as "Automotive Master Mechanic"; however, this project's technical advisors suggested that, in the field, the term "Technician" is preferred. Therefore, throughout this report, the occupation is referred to as "Automotive Master Technician."

[3] The Governing Board initially requested the inclusion of the occupation of Plumber. When job-training programs were being identified for panelist recruitment, however, it became apparent that within the Plumber occupation the apprenticeship method of entering the profession was dominant, with not many formal job-training programs available from which to recruit. For this reason, in consultation with this project's technical advisors, WestEd recommended that the occupation of HVAC (O*NET refers to this occupation as "Heating and Air Conditioning Mechanics and Installers"; an alternative name is Heating, Ventilation, and Air Conditioning, or HVAC) replace the occupation of Plumber. The HVAC occupation meets the criteria laid out by the Governing Board and includes a network of formal job-training programs from which panelists were recruited.

- Representation of different training paths (e.g., vocational training or community college programs).

The Design Document stipulated the use of a modified bookmark methodology, for consistency with other recent NAEP standard-setting studies, and mandated the inclusion of a pilot study to evaluate study methodology, materials, and logistics; this pilot study was to include two postsecondary areas: college and one occupation. The Design Document also stipulated the following deliverables of the JSS studies:

- Borderline performance descriptions (BPDs)—descriptions of what a student at the borderline of academic preparedness should know and be able to do—to be used by panelists as standards when determining placement of bookmarks representing preparedness for the postsecondary areas; and
- Cut scores for all postsecondary areas within each content area.

In addition, NAEP exemplar items were selected by the panelists as illustrations of what academically prepared students know and can do.

As stipulated in the Design Document, the JSS studies used a replicate panel design, in which two replicate panels were convened for each postsecondary area within a content area in an attempt to estimate the reliability of the replicate panels' cut scores. In addition, the standard-setting process followed the recommendation of the Design Document in using computers whenever possible to increase the efficiency and effectiveness of the process, computerizing several key activities, including the capture of knowledge, skills, and abilities (KSA) annotations, the setting of cut scores, the provision of feedback, and the capture of process evaluation responses.

To capitalize on the potential for cost efficiencies and to accommodate a constrained timeline, within the pilot JSS study and each of the three operational JSS sessions, WestEd—with the support of Measured Progress and EPIC—conducted mathematics and reading JSS studies concurrently for two postsecondary areas, pairing occupations based on occupational similarities (e.g., pairing the two healthcare occupations of LPN and Pharmacy Technician) and panelist availability. The Design Document stipulated that the pilot study comprise college-preparedness and one of the five occupations; based on recruitment considerations,[4] college-preparedness was paired with the occupation of Automotive Master Technician for both the pilot study and the first of the operational sessions. The pilot study was conducted in April 2011, while the three operational studies were conducted from May through July 2011.

WestEd and Measured Progress—the subcontractor responsible for standard-setting activities—consulted with a JSS Technical Advisory Committee (JSS-TAC) convened for this study in all aspects of the project. The JSS-TAC was a five-member committee that collectively represents expertise in standard setting, psychometrics, transition to college and college academic requirements, and job-training requirements. The JSS-TAC met via conference call twice and in person four times over the course of the project and provided input on key components of the studies, including refinements to the methodology, panelist recruitment, implementation of the special study (described later in this report), data analysis procedures, and conclusions and recommendations to be submitted to the Governing Board.

[4] Of the five occupations, the Automotive Master Technician occupation had the largest pool of eligible job-training programs from which to recruit. Also, the prominence of one JSS-TAC member in the automotive industry—as Vice President, Test Development, of the National Institute for Automotive Service Excellence—positioned him to greatly support recruitment efforts in this occupation.

This report provides a detailed description of the method and outcomes of the pilot study and the three operational JSS sessions. It describes project activities preceding the pilot study that led to the development of draft BPDs. It also summarizes a special study that was implemented during the third operational JSS session to explore the utility of an alternative item map format within the context of these studies.

Project Overview

This section provides an overview of the project, including participants, recruitment, the standard-setting methodology, and data analysis. While certain details pertaining to panelist recruitment and the standard-setting methodology varied across the postsecondary areas and across the pilot and operational JSS sessions, underlying protocols were consistently implemented across all studies. These underlying protocols are described in this section; variations specific to particular postsecondary areas are described in later sections of this report.

Study Staff and Observers

JSS-TAC. At the recommendation of the Governing Board, WestEd established a JSS-TAC to provide advice and consultation throughout the course of this project. It comprised the following members:

- Dr. Bob Forsyth, Professor Emeritus, College of Education, University of Iowa;
- Dr. Ed Haertel, Jacks Family Professor of Education, Associate Dean for Faculty Affairs, Stanford University;
- Dr. Ron Hambleton, Distinguished University Professor, University of Massachusetts at Amherst;
- Dr. Chuck Kunce, Vice President, Test Development, National Institute for Automotive Service Excellence; and
- Dr. Lynn Webb, testing consultant.

In addition to providing key project consultation, Dr. Kunce attended the pilot study as an observer, and Dr. Forsyth attended both the pilot study and the second operational JSS session as an observer.

Project Management Staff. The following WestEd, Measured Progress, and EPIC employees made up the project management staff:

Dr. Stanley Rabinowitz (principal investigator and senior technical advisor, WestEd) and Dr. Dave Conley (senior technical advisor, EPIC) provided intellectual leadership, including spearheading advance planning of the study, overseeing the development of protocols and materials, consulting on recruitment and methodological issues related to career preparedness, and reviewing reports. Dr. Rabinowitz attended the first day of all JSS sessions.

Dr. Jennae Bulat (WestEd) and Dr. Luz Bay (Measured Progress) (project co-directors) shared project management responsibilities. Dr. Bulat executed day-to-day project management, including managing the schedule and budget, overseeing project staff, monitoring changes to the scope of work, recruiting participants, managing the JSS-TAC and content facilitators, and directing communication with the Governing Board. Dr. Bay served as the project's director of standard setting and, as such, oversaw all aspects of the standard-setting process and managed Measured Progress's team of psychometricians, data analysts, and hardware/software engineers. Both co-directors participated in all pre-standard-setting and standard-setting activities.

Dr. Tia Sukin (Measured Progress) served as the project's lead psychometrician, supporting the project onsite at all standard-setting sessions and overseeing all data analysis

Content Facilitators. The Design Document stipulated that two content facilitators—selected from among the members of the 2009 framework development panels for the Grade 12 NAEP in mathematics and reading—be recruited to support each standard-setting workshop, one for each content area (mathematics and reading). These content facilitators guided the development of BPDs through initial web-based development sessions and subsequent review

cycles within standard-setting workshops, ensuring that the performance descriptions were consistent with the knowledge and skills included in the NAEP frameworks. Content facilitators also helped lead discussions (e.g., during item review) within the standard-setting workshops. Because two workshops ran concurrently during each four-day session, a total of four content facilitators (two each for mathematics and reading) were required to support the pilot study and each of the operational sessions. The following individuals were engaged by WestEd as consultants to serve as content facilitators for this project:

- Dr. Carol Jago (reading), Director, California Reading and Literacy Project, University of California, Los Angeles;
- Dr. Michael Kamil (reading), Professor Emeritus, Psychological Studies in Education, School of Education, Stanford University;
- Dr. Jeremy Kilpatrick (mathematics), Regents Professor, Mathematics Education, University of Georgia; and
- Dr. Linda Wilson (mathematics), mathematics education consultant.

In addition, given the substantial time commitment required of each content facilitator and the aggressive schedule of this project, a backup facilitator was recruited for each of the content areas. These backup facilitators participated in all training sessions, participated in orientation webinars and the development of BPDs, and attended the pilot webinars and the pilot study, but they participated in operational workshops only when needed to replace a primary content facilitator (due to schedule conflicts, as described in relevant sections of this report).

The following backup content facilitators were engaged:

- Dr. Janice Dole (reading), Professor, College of Education, University of Utah; and

- Dr. Mary Lindquist (mathematics), mathematics education consultant and Fuller E. Callaway Professor of Mathematics Education, Emerita, Columbus State University.

Process Facilitators. The Design Document also called for four process facilitators to support each standard-setting workshop—two for each content area (mathematics and reading)—with one process facilitator responsible for each replicate panel within each content area in a workshop. These process facilitators were responsible for ensuring that the standard-setting process was executed faithfully. Because two workshops ran concurrently during a four-day session, eight process facilitators (four each for mathematics and reading) were required to support the pilot study and each of the operational sessions. The following eight process facilitators were recruited for this project:

- Ms. Kirsten Aspengren (reading), Director of the AP Course Audit, EPIC;
- Mr. Eric Crane (mathematics), Senior Research Associate, WestEd;
- Mr. Timothy Crockett (mathematics), Senior Vice President, Client Services, Measured Progress;
- Dr. Carole Gallagher (reading), Senior Research Associate, WestEd;
- Dr. Phil Robakiewicz (reading), Director, Client Services, Measured Progress;
- Dr. Joseph St. George (mathematics), Program Manager, Measured Progress;
- Ms. Marcia Tibbetts (reading), Program Director, Measured Progress; and
- Dr. Terri Ward (mathematics), Co-Director, the Center for Educational Policy Research, University of Oregon.

As with the content facilitators, given the aggressive schedule of this project, a backup process facilitator was recruited who could fill in as needed to replace a primary process facilitator. Given her background in facilitation and her familiarity with this project, Dr. Mary Seburn—the

Director of EPIC's Research, Design, and Analytics Division—was identified to fill this role. Dr. Seburn participated in all training sessions and attended the pilot webinars and the pilot study; she also attended the first operational session in order to replace a process facilitator who was unable to participate on the final day of that session.

Observers. The following external observers were invited to attend JSS sessions to provide feedback on the process:

- JSS-TAC members Dr. Bob Forsyth and Dr. Chuck Kunce attended the pilot study. Dr. Forsyth also attended the second operational session.
- Dr. Mildred Bazemore, member of the Governing Board's 12th Grade Preparedness Technical Advisory Group, attended the third operational session.
- Ms. Michelle Blair, Senior Research Associate at the Governing Board, attended the pilot study.
- Dr. Barbara Dodd, Professor of Educational Psychology at the University of Texas at Austin, current member of the NAEP Achievement Levels Committee on Standard Setting and contributor to several NAEP achievement-level-setting contracts, attended the first operational session.

In addition, Dr. Steve Viger, Manager of Measurement Research & Psychometrics at the Michigan Department of Education, received special permission from the Governing Board to observe the first operational session, although not in the same capacity as the other external observers.

Panelist Recruitment Plan

The objective of this study's panelist recruitment plan was to produce well-qualified panels, broadly representative of the following attributes:

- Gender;
- Race/ethnicity;
- Geographic location;
- Type of mathematics or reading experience; and
- Type of institutional affiliation.

This study design called for distinct groups of panelists to be recruited for six postsecondary areas: college-preparedness and job-training programs in five occupations. The postsecondary areas of college-preparedness and the Automotive Master Technician occupation were selected for inclusion in the pilot study; therefore, for these two areas, panelists were recruited for pilot and operational session panels. For each postsecondary area, recruitment followed a two-tiered process: (1) identification of eligible institutions using an approved sampling plan and solicitation of nominees from qualified individuals within these institutions, and (2) recruitment of panelists from the pool of nominees. To the extent possible given recruitment challenges, only one representative from a given institution was selected to participate in this series of studies. Requests for panelist nominations and panelist recruitment communications were primarily conducted via email; follow-up telephone calls were made as needed to ensure that the electronic communications had been received and to address questions or provide additional information. When email addresses were not available, potential nominators were contacted via U.S. mail.

Given the distinct nature of each of the postsecondary areas, sampling plans and recruitment efforts were tailored to each area in order to recruit the most qualified panelists. An overview of the recruitment process used across postsecondary areas is provided in this section; more detailed recruitment information specific to each panel is included in relevant sections of this report. For all pilot and operational panels, panelists were assigned to mathematics and reading content areas; within each content area, panelists were equally divided between two replicate panels (with the goal for each replicate panel to have six panelists for the pilot study and 10 panelists for the operational workshops).

Panelist Recruitment for College-Preparedness Panels. A total of 64 college-preparedness panelists was required for the pilot and operational college-preparedness JSS workshops

(24 for the pilot workshop and 40 for the first operational workshop), with panelists representing both postsecondary institutions (including two- and four-year institutions) and secondary-level institutions. The Governing Board recommended, and the JSS-TAC concurred with, a distribution of 80% postsecondary and 20% secondary instructors on each replicate operational panel (67% and 33% on each replicate pilot panel), thereby requiring a total of 48 postsecondary instructors (16 total for the pilot workshop and 32 total for the first operational workshop) and 16 secondary-level instructors (8 total for each of the pilot and operational workshops). In recognition that reading, per se, is not a typical college-level course, the Governing Board requested that half of the reading postsecondary instructors come from programs other than traditional English, composition, literature, and/or rhetoric programs. All postsecondary mathematics instructors were recruited from mathematics departments or programs.

Individuals with any of the following characteristics were considered to be eligible to serve as college-preparedness panelists:

- Instructors of two- and/or four-year higher education entry-level, credit-bearing English language arts or mathematics courses that fulfill general education requirements for a four-year degree program.
- Instructors of two- and/or four-year higher education entry-level, credit-bearing courses in other disciplines that fulfill general education requirements for a four-year degree program and that require students to engage in substantial amounts of reading, whether literary or informational texts.
- Instructors of remedial or developmental reading or mathematics courses in postsecondary institutions.
- Postsecondary instructors who specialize in mathematics or reading instruction or curriculum.
- Postsecondary instructors who have participated directly in the development of entry-level reading/English language arts or mathematics placement tests for a postsecondary institution.
- Postsecondary English language arts or mathematics instructors of entry-level courses who have participated directly in the development of high-school-to-college transition projects.
- Grade 12 high school English language arts or mathematics instructors who have worked with developers of college admission or placement tests or who have worked on high-school-to-college transition projects.

- Instructors with at least five years of grade 12 or postsecondary mathematics or reading teaching experience in courses appropriate to the targeted entry-level courses for student placement. For high school teachers, this may include teaching courses that count for college credit or teaching in dual-enrollment programs.
- Instructors judged to have very good professional performance by a supervisor or someone in the position to make that judgment.

The approved sampling plan for *postsecondary* colleges and universities used the following two-tiered sampling procedure for the recruitment of panelists:

Tier 1: At the first tier, postsecondary institutions were sampled from the U.S. Department of Education's Integrated Postsecondary Education Data System (IPEDS), maintained by the U.S. Department of Education's National Center for Education Statistics (National Center for Education Statistics, 2009). Data for the most recent available year (2009–10) were utilized, resulting in a dataset of 7,069 institutions. Institutions were sampled from this set according to the following characteristics based on IPEDS variable segments[5] (percentages represent targeted distributions stipulated in the sampling plan):

- Private/public status
  - 25% private institutions
  - 75% public institutions
- Selectivity of institution

[5] Categories were derived from the following IPEDS variables: Sector of Institution; Percent Admitted; Open Admission Policy (whether the institution accepts all or most entering first-time undergraduate-level students); and Total Enrollment (total men and women enrolled for credit in the fall of the academic year).

    - 63% open enrollment (81–100% applicants admitted)
    - 34% moderately selective (31–80% applicants admitted)
    - 3% highly selective (5–30% applicants admitted)
- Size of institution
    - 24% small (<5,000 enrollment)
    - 42% medium (5,000–19,999 enrollment)
    - 34% large (20,000+ enrollment)

Within these institutions, heads of relevant departments or programs were contacted with a request for panelist nominations (the nomination request email can be found in Appendix B); they were asked to nominate qualified instructors via an online form that solicited nominees' names, contact information, and basic qualifications (see Appendix C for the nomination form).

Tier 2: Once nominated by department or program heads via the online form, candidates were contacted and asked to submit résumés and complete an online informational form (see Appendix D). Once a pool of qualified panelist candidates was compiled, candidates were rated based on their background experiences and qualifications, with the most qualified panelist candidates selected and assigned to replicate panels and to table groups within panels in order to attain balance with respect to content area expertise, gender, demographics, NAEP geographic regions, and institutional characteristics.

The approved sampling plan for *secondary-level* institutions also used a two-tiered sampling procedure for the recruitment of panelists:

Tier 1: At the first tier, secondary institutions were sampled from the most recently available data (school year 2008–09) in the Common Core of Data (CCD) maintained by the U.S. Department of Education's National Center for Education Statistics (National Center for Education Statistics, 2008–09). The resulting dataset included a total of 20,381 schools. Institutions were sampled from this set according to the following characteristics based on CCD variables (percentages represent targeted distributions stipulated in the sampling plan):[6]

- Urbanicity[7]
    - 34% urban
    - 41% suburban
    - 25% town/rural
- Size of institution
    - 14% small (<500 enrollment)
    - 22% medium (501–1,000 enrollment)
    - 64% large (1,001+ enrollment)

Within these institutions, principals were contacted with a request for panelist nominations (the nomination request email can be found in Appendix E); they were asked to nominate qualified instructors via an online form that solicited nominees' names, contact information, and basic qualifications (see Appendix F for the nomination form).

[6] Variables used were Urban-centric Locale and Total Students.
[7] The Urban-centric Locale code is based on the school's physical address or mailing address and is a measure of a school's location relative to populous areas. For the sampling purposes of this study, only the first element of the 12 Urban-centric Locale categories was used and the data for the Rural and Town codes were combined. "Urban" corresponds to the CCD "City" category.

Tier 2: Once nominated by department or program heads via the online form, candidates were contacted and asked to submit résumés and complete an online informational form (see Appendix D). Once a pool of qualified panelist candidates was compiled, candidates were evaluated based on their background experiences and qualifications, with the most qualified panelist candidates selected and assigned to replicate panels and to table groups within panels in order to attain balance with respect to content area expertise, gender, demographics, NAEP geographic regions, and institutional characteristics.

Panelist Recruitment for Career-Preparedness Panels. A total of 24 Automotive Master Technician panelists was stipulated for the pilot study (12 in mathematics and 12 in reading, with each group divided into two replicate panels of six panelists each). For each occupation represented in an operational workshop, a total of 40 job-training panelists was stipulated
(20 in mathematics and 20 in reading, with each group divided into two replicate panels of 10 panelists each).

Individuals with the following characteristics were considered to be eligible to serve as career-preparedness panelists:

- Instructors knowledgeable of the content-area knowledge and skills required for entry into a job-training program in one of the occupations to be studied;
- Instructors with at least two years[8] of experience teaching entry-level courses in a job-training program in one of the occupations to be studied; and

[8] While the Design Document specified a minimum of five years of teaching experience for job-training instructors, JSS-TAC members with knowledge of the five occupations recommended reducing this requirement to two years.

- Instructors judged to have very good professional performance by a supervisor or someone in the position to make that judgment.

The approved sampling plan for job-training programs across the five occupations used a two-tiered sampling procedure for the recruitment of panelists:

Tier 1: A list of eligible job-training programs for each occupation was compiled. The process by which eligible programs were identified varied by occupation, as described later in this section. The heads of eligible job-training programs were contacted with a request for panelist nominations (see Appendix G for sample nomination materials); they were asked to nominate qualified instructors via an online form that solicited nominees' names, contact information, and basic qualifications (see Appendix H for a sample nomination form).

Tier 2: Nominated candidates were contacted and asked to submit résumés and complete an online informational form (see Appendix I for a sample form). Once a pool of qualified panelist candidates was compiled, candidates were evaluated based on their background experiences and qualifications, with the most qualified panelist candidates selected and assigned to replicate panels and to table groups within panels in order to attain balance with respect to content area expertise, gender, demographics, NAEP geographic regions, and institutional characteristics.

The process for determining program eligibility and recruiting panelists was largely unique to each program, although recruitment across all occupations began with the identification of a core pool of programs to target. Table 3 displays the numbers of programs that were initially identified as eligible for each of the five occupations, with the process for identifying additional programs and determining eligibility described for each occupation as follows.

Table 3. Number of Initially Identified Programs by Occupation

Occupation	Number of Programs
Automotive Master Technician	667
Computer Support Specialist	353
HVAC	43
LPN	160
Pharmacy Technician	230

For the occupation of Automotive Master Technician, it was determined, through consultation with the JSS-TAC, that the National Automotive Technicians Education Foundation (NATEF) is the primary accrediting body for this field. Therefore, the NATEF list of accredited programs (http://www.natef.org/certified.cfm) served as the basis for panelist recruitment. Also at the expert advice of a member of the JSS-TAC, programs that are housed in high schools and programs that offer fewer than the eight possible certification areas were excluded from recruitment, as these programs are not likely to offer training for Automotive Master Technicians and are, therefore, qualitatively different from those that offer all eight certification areas. To support panelist recruitment efforts, a letter of endorsement from the President and CEO of the National Institute for Automotive Service Excellence (ASE) was included with all recruitment communications. In addition, the president of NATEF and the president of Automotive Youth Educational Systems (AYES) provided membership lists from the North American Council of Automotive Teachers (NACAT) and AYES, respectively, and members representing eligible programs were asked to provide panelist nominations; the Executive Director of the Association for Career and Technical Education (ACTE) and the Executive Director of the National Association of State Directors of Career Technical Education Consortium (NASDCTEc) were also contacted to request nominations.

Identifying appropriate programs from which to recruit Computer Support Specialist panelists was complicated by the abundance of areas of focus within this occupation and the degree of overlap between many of these areas. In an attempt to ensure that potential panelists represented the types of programs reflected in the O*NET description of this occupation, eligible programs were primarily drawn from two primary databases that specifically reference the occupation of Computer Support Specialist: the College Board's Majors & Careers Central database (http://www.collegeboard.com/student/csearch/majors_careers/index.html), which lists colleges offering this major at the certificate and associate's degree levels, and CareerOneStop's database (http://www.careerinfonet.org/edutraining/Default.aspx?searchMode=occupation), which is sponsored by the U.S. Department of Labor's Employment and Training Administration and uses data collected in the U.S. Department of Education's IPEDS. To supplement these efforts, computer support instructors who had been identified by EPIC for a prior study were contacted, emails were sent to a database of computer support instructors acquired from Market Data Retrieval (MDR), and recruitment information was disseminated through the ACTE network.

For the occupation of HVAC, upon consultation with the Director of Education for the Air Conditioning, Heating, and Refrigeration Institute (AHRI), it was determined that the Partnership for Air-Conditioning, Heating, Refrigeration Accreditation (PAHRA) is the primary accreditation body for this field. Therefore, the PAHRA list of accredited programs (http://pahrahvacr.org/Content/Schools_38.aspx) served as the initial list of eligible programs. In addition, a request for nominations, including an endorsement from the Director of Education for AHRI, was distributed throughout the Council for Air Conditioning and Refrigeration Educators (CARE) network by the Director of Education for AHRI; recruitment information was

disseminated through the ACTE network; and emails were sent to a database of HVAC instructors acquired from MDR.

The National League for Nursing Accrediting Commission (NLNAC) is responsible for the specialized accreditation of nursing education programs, both postsecondary and higher-degree, that offer either a certificate, a diploma, or a recognized professional degree (clinical doctorate, master's/post-master's, baccalaureate, associate's, diploma, or practical nursing). NLNAC makes available a searchable database of accredited nursing programs (http://www.nlnac.org/Forms/directory_search.htm), and this database was used to identify accredited institutions offering LPN programs.

It was determined that the American Society of Health-System Pharmacists (ASHP) is the leading national accrediting body specifically for Pharmacy Technician training programs. Therefore, the ASHP's Pharmacy Technician Training Program Directory (http://accred.ashp.org/aps/pages/directory/technicianProgramDirectory.aspx) was used to identify institutions that offer training in Pharmacy Technician skills.

In addition to the aforementioned occupation-specific strategies, certain strategies supported recruitment activities across all five occupations:

- Networks of proprietary schools offering eligible programs were solicited for nominations;
- The community college networks from 44 states were asked to disseminate project information to their member community colleges; and
- Qualified panelists were asked to nominate equally qualified colleagues as part of the recruitment process.

To ensure that panelists who were recruited through these means came from eligible programs, such panelists were asked to confirm that their programs met the O*NET definitions of the occupations.

For the occupations included in the pilot study and the first two operational sessions—Automotive Master Technician, LPN, and Pharmacy Technician—panelists were recruited only from within job-training programs. While all panelists had to be familiar with the mathematics or reading knowledge and skills required by their programs, there was no requirement to recruit only job-training instructors who directly taught mathematics or reading courses within these programs; indeed, in many occupations and job-training programs, such courses—especially for reading—do not exist. However, for the final operational session, which included the Computer Support Specialist and HVAC occupations, it was decided to add secondary-level teachers of English language arts and mathematics courses to the replicate panels in order to provide a perspective on what mathematics and reading skills are taught at the high school level. These secondary-level teachers were recruited from networks of qualified instructors within WestEd, EPIC, and Measured Progress.

Despite occupation-specific recruitment efforts, recruitment of job-training instructors—particularly in the Automotive Master Technician, HVAC, and LPN occupations—proved more difficult than was anticipated and more difficult than the recruitment of college-preparedness workshop panelists. Across the occupations, most job-training program heads and instructors were not familiar with either NAEP or standard-setting activities; successfully explaining the importance, purpose, and approach of this type of study proved more difficult than when recruiting instructors from more traditionally academically focused secondary schools and colleges. The timing of the first and second operational workshops coincided with the end of the

academic year for many, a difficult time to be away from classes; also, within at least some job-training programs, authorization to take a full week (four days for each workshop, plus a day for travel) away from classes appeared to be difficult to obtain, perhaps in part because budgetary constraints within some institutions made it difficult to find replacement instructors to cover classes. In addition, based on correspondence with nominees, it is not uncommon for job-training instructors in some occupations (e.g., Automotive Master Technician, Computer Support Specialist, HVAC) to also work as practicing technicians, thus making it even more difficult to commit to the amount of time required for each workshop.

In order to meet targeted numbers of panelists, the Governing Board allowed the recruitment of multiple representatives from the same institution, especially when such representatives were assigned to different workshops (e.g., a pilot study workshop and an operational workshop) or to different content areas within a workshop. Where panelists representing the same institution were recruited for the same content area within a workshop, they were assigned to different panels or, in a few instances, to different table groups within a panel (see each workshop's section in this report for the numbers of panelists representing multiple institutions). In addition, due to the unanticipated challenges of recruiting sufficient panelists for the Computer Support Specialist, HVAC, LPN, and Pharmacy Technician panels, a number of nominators of panelists from these occupations were paid a recruitment incentive of $100 for each nominee who was selected to participate in the study.[9]

[9] A total of 34 incentive payments were made.

Judgmental Standard-Setting Process

The JSS process refers to all activities through which components of the standard—the borderline performance descriptions (BPDs), the cut scores and percentages of students at or above the cut scores, and the items illustrative of what students who score above the cut scores know and can do—are obtained.

Four JSS sessions—the pilot study and three operational sessions—were held at the Westin St. Louis Hotel in St. Louis, Missouri, from April to July of 2011. Each session comprised two workshops that ran concurrently, as displayed in Table 4.

Table 4. JSS Sessions and Workshops

JSS Session	Dates	Workshops	Panels
Pilot Study	April 26–29, 2011	College-Preparedness	Mathematics Panels A & B Reading Panels A & B
		Automotive Master Technician	Mathematics Panels A & B Reading Panels A & B
Operational Session 1	May 24–27, 2011	College-Preparedness	Mathematics Panels A & B Reading Panels A & B
		Automotive Master Technician	Mathematics Panels A & B Reading Panels A & B
Operational Session 2	June 7–10, 2011	LPN	Mathematics Panels A & B Reading Panels A & B
		Pharmacy Technician	Mathematics Panels A & B Reading Panels A & B
Operational Session 3	June 28–July 1, 2011	Computer Support Specialist	Mathematics Panels A & B Reading Panels A & B
		HVAC	Mathematics Panels A & B Reading Panels A & B

Following is a description of the process implemented in this series of JSS studies, including the development of BPDs, the provision of advance materials, the use of a computer-aided design, the use of replicate panels, the division of panels and item pools, the use of NAEP-like scales, facilitator and panelist training, and the standard-setting implementation itself.

Development of Borderline Performance Descriptions. The development of BPDs was a critical component of the JSS process, because BPDs describe the performance required for minimal preparedness in each postsecondary area and are, therefore, the statements of the performance standard to be represented on the NAEP scale by each cut score. This study was designed to have the BPD for each postsecondary area developed by panelists recruited from within that area. Instructors and job trainers recruited to serve on the JSS panels were engaged in an iterative three-step process to develop the BPDs.

Step 1: Orientation Webinars. Each panelist was expected to participate in a two-hour web-based orientation led by the project management team and content facilitators. Webinars were scheduled approximately two to three weeks prior to each JSS session. Panelists who were recruited after the webinar for their workshop was held and panelists who could not accommodate the timing of their scheduled webinar were sent a link to an online recording of a webinar and asked to watch the webinar. The objectives of the webinars were as follows:

- Introduce the study purpose, context, methodology, and panelists' roles;
- Introduce the Governing Board's definition of minimal academic preparedness;
- Present and review the NAEP framework;
- Describe the process for developing the BPDs; and
- Instruct panelists on how to complete the next step in the development of the BPDs.

Step 2: Content Framework Objectives Review. Following the webinar, each panelist was asked to review the objectives within the NAEP framework and indicate whether the knowledge and skills reflected in each objective were required for a student to be minimally prepared for entry into that postsecondary area's course or program. At the content facilitators' request, reading

panelists in all of the five occupations and mathematics panelists in the Computer Support Specialist and HVAC occupations were also asked to submit to WestEd examples of texts or tasks that a student would encounter in an introductory course, and to indicate what sections of those texts or tasks posed particular challenges to incoming students; the content facilitators reviewed these texts to better understand the types of materials that students in the job-training programs had to read in an entry-level course.

Step 3: Development and Refinement of BPDs. Content facilitators used the information collected from the panelists' review of the frameworks and examples of texts and tasks to draft preliminary BPDs. The content facilitators developed preliminary pilot study BPDs during the April 14–15, 2011, Facilitators' Training. Refinement of these pilot study BPDs and development of all operational BPDs occurred either remotely or when content facilitators were convened for JSS workshops. NAEP objectives that had been rated as being required for minimal preparedness by a majority of panelists in their content objective reviews were included in the preliminary descriptions, as were objectives not selected by panelists but recommended for inclusion by the content facilitators. BPDs for operational workshops were informed by the pilot study BPDs, with modifications made as needed to capture the unique input of each workshop's panelists. Draft BPDs for each JSS workshop were submitted to the project co-directors and the Governing Board for review the week prior to each workshop. Approved preliminary BPDs were then shared with panelists just prior to each JSS workshop.

On Day 1 of each JSS workshop, content facilitators reviewed the NAEP frameworks and then led panelists in a review of the preliminary BPDs, with particular attention to those NAEP objectives not identified as important by the majority of panelists and to objectives recommended for inclusion by the content facilitators. The development of BPDs was a joint

activity of the replicate panels in each content area and postsecondary area. Facilitators engaged panelists in discussions of the objectives and helped them to understand the knowledge and skills involved. Panelists came to agreement on edits to the preliminary BPDs. Following the Design Document, pilot study panelists were given the opportunity to continue modifying the BPDs throughout Days 2 and 3, with final versions submitted by each pair of replicate panels just prior to the last round of standard setting. However, based upon recommendations by content facilitators and the JSS-TAC, this process was revised for the operational sessions. For the three operational sessions, panelists were instructed to develop the BPDs as fully as possible during the Day 1 session so that the first round of standard setting would be based upon fully developed BPDs; panelists were then allowed the opportunity to refine the BPDs as needed during Day 2 and Day 3 discussions, with the expectation that such refinements would require relatively minimal changes to the BPDs.

Advance Materials. Upon confirming their availability to participate in this study, panelists were sent nondisclosure agreements, consultant engagement forms, and information pertaining to the meeting site and travel logistics; this information included instructions for contacting the designated travel agent, information about expense reimbursement policies, and information and maps pertaining to ground transportation and meal options in and around the hotel.

Prior to commencing JSS-related activities, panelists were sent both electronic and paper copies of the following documents for their respective content areas (mathematics or reading):

- *Mathematics Framework for the 2009 National Assessment of Educational Progress* (National Assessment Governing Board, 2008);

- *Reading Framework for the 2011 National Assessment of Educational Progress* (National Assessment Governing Board, 2010b);
- Governing Board's definition of 12th grade student preparedness (National Assessment Governing Board, 2010a);
- Governing Board's policy definitions of the three NAEP achievement levels for reading and mathematics: Basic, Proficient, and Advanced (National Center for Education Statistics, 2010);
- *The Nation's Report Card: Grade 12 Reading and Mathematics 2009 National and Pilot State Results* (National Center for Education Statistics, 2010); and
- Additional Grade 12 released items that were not reported in *The Nation's Report Card.*

Panelists were instructed to review these documents prior to participating in the online orientation webinars, to use them to inform their review of the NAEP objectives, and to bring them to the JSS sessions for use in the refinement of the BPDs and in the standard-setting process.

The week prior to each JSS session, panelists were sent electronic copies of the agenda for the session and the JSS Briefing Booklet that was prepared for the session, which provided a highly detailed and technical description of every step in the study methodology (Briefing Booklets for all sessions are provided in Appendix J). Panelists were instructed to review both documents and to bring the Briefing Booklet to the JSS session as a reference tool that could be used throughout the standard-setting process. They were also sent electronic copies of the draft preliminary BPDs that had been developed by content facilitators and were instructed to review their respective BPDs and to be prepared to discuss them during Day 1 of the JSS workshop.

Computer-Aided Bookmarking (CAB). For the JSS implementation of this project's bookmark standard-setting methodology, significant elements of the process were computerized using Computer-Aided Bookmarking software (CAB) to improve efficiencies of the process and enhance the quality of the panelists' experience. The computerized elements of the process included the following:

- Provision of electronic OIB (Ordered Item Booklet) and CROIB (Constructed-Response Ordered Item Booklet) (although paper versions were also available);
- Collection of KSA annotations;
- Collection of panelists' ratings;
- Collection of panelists' evaluations;
- Presentation of consequences data feedback;
- Presentation of Rater Location Chart; and
- Selection of exemplar items.

Each panelist used a netbook computer to perform his or her tasks. All panelists entered their ratings and responses to evaluation questionnaires into the CAB. Panelists also used netbook computers to record the KSAs identified during the item review of the electronic OIBs and CROIBs. The CAB was also used to present consequences data feedback, or impact data. Within the CAB, the consequences data feedback feature is an interactive mechanism that displays the consequences data resulting from moving cut scores up or down the NAEP scale. The CAB also calculates the percentage of students at or above each scale score, as well as indicating on which side of the cut score each item falls as a panelist moves the placement of the cut score.

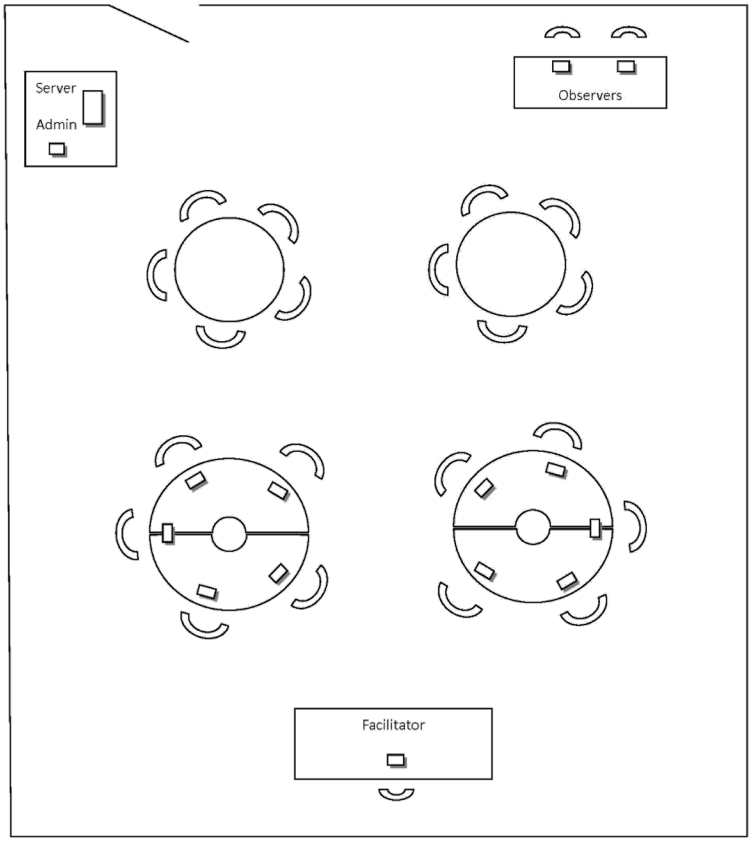
Replicate Panels. The Governing Board's design called for the use of replicate panels throughout the standard-setting process. Such replication of panels was included in the study design as a way of estimating how consistently the cut score was set across panels, to assess the reliability of the judgments.

Each operational workshop was designed to include 40 panelists. Twenty panelists were mandated for each operational standard-setting workshop in each content area (replicate panels of 10 panelists each), and 12 panelists were mandated for each pilot study workshop (replicate panels of six panelists each). For each content area (mathematics or reading), panelists were assigned to one of two replicate panels: Panel A or Panel B. 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 of panelists were considered when assigning members to replicate panels and tables, and the panelists were selected to maximize equivalence of the replicate panels. Similarly, table group assignments were made to maximize equivalence across the groups. In all other aspects, the assignments were random. The goal was to have replicate panels and tables as similar as possible with respect to panelist type (i.e., educator role), gender, geographic region, and race/ethnicity.

With the exception of general sessions, large-group training, and sessions pertaining to the development and refinement of the BPDs, the two replicate panels for each content area worked independently in separate rooms, with each replicate panel led by a designated process facilitator. The process facilitator assigned to a replicate panel provided instructions (e.g., for activities such as the rounds of ratings), provided feedback information between rounds, and led discussions about different feedback information.

For sessions in which the two replicate panels were combined (e.g., to discuss or edit BPDs), Panel B panelists joined Panel A panelists in the Panel A room. Both Panel A and Panel B room configurations had two round tables for panelists, where netbook computers were set up. However, Panel A rooms had two additional tables at the back of the room to accommodate Panel B panelists. The Panel A room configuration is presented in Figure 1.

Figure 1. Panel A Room Configuration



Item Pool Division. The NAEP item pools were divided into two corresponding sets, A and B, for use by replicate panels in order to limit the number of items reviewed by each panelist and, therefore, minimize possible fatigue. The division also created a design that allowed the

reliability of the process to be evaluated. The item pools used by replicate panels did not include items that were released to the public. The resulting item rating pools contained 49% to 53% of the items in the assessment. Items included in both pools were referred to as common items. Equivalence was monitored with regard to (a) content area subscale representation, (b) item type representation, and (c) item difficulty. Tables 5 and 6 present summaries of the item pools by panel and overall for mathematics and reading, respectively. Item difficulty was calculated using each item's scale value (or score point, for polytomous items) for which a correct response probability (probability of receiving that score point or higher) of 0.67 was expected.

Table 5. Summary of Panel Item Pools, Mathematics

Panel	Number of Items	Percent by Subscale[a]				Percent by Item Type[b]			Item Difficulty				
		NPO	M&G	DAP	ALG	MC	DCR	PCR	Total Possible Points	Mean	SD[c]	Min. Item Difficulty	Max. Item Difficulty
A	81	11	32	24	33	65	10	25	113	195.8	38.7	88.0	300.0
B	82	12	29	27	32	67	9	24	111	194.3	40.3	71.0	300.0
Pool[d]	163	12	31	24	34	66	9	25	225	196.0	39.2	71.0	300.0

[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 that were excluded from the panel Ordered Item Booklets.

Table 6. 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. Item Difficulty	Max. Item Difficulty
A	69	28	72	59	10	30	96	300.5	52.5	172.0	444.0
B	71	28	72	61	8	31	100	291.9	51.0	158.0	444.0
Pool[d]	131	31	69	58	9	33	186	297.0	50.8	158.0	444.0

[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 that were excluded from the panel Ordered Item Booklets.

NAEP-like Scales. For this project, different NAEP-like scales were used to avoid contamination of the standard-setting process (e.g., influence by replicate Panel A on the cut score decision of replicate Panel B) and the impact of having the NAEP achievement-level cut scores for grade 12 released. Each scale was a linear transformation of the NAEP reporting scale. Different NAEP-like scales were used for each panel for the two postsecondary areas in a session, and different NAEP-like scales were used for the replicate panels within a content area.

Facilitator Training. Efforts were made to ensure that the facilitators were properly trained to implement the process uniformly across the eight panels within each JSS session. The four content facilitators and eight process facilitators attended a two-day training, held at the Measured Progress facilities in Dover, New Hampshire, prior to the pilot study. The project co-director overseeing the judgmental standard-setting process (director of standard setting) led the training. The Governing Board Contracting Officer's Representative (COR) provided the overview of the study within the context of the Governing Board's Preparedness Research Program. The WestEd project co-director presented the panelist-recruitment process and the BPD-development process, and the psychometrician for the project described the technical underpinnings for mapping items on the NAEP scale for the purpose of rank-ordering the items for the implementation of the bookmark standard-setting method. A walkthrough of the standard-setting process was attempted, during which the facilitators were introduced to the CAB and how it would be used by panelists. The attempt was not particularly successful, due to software performance issues when multiple users were accessing the CAB at the same time. While the session was helpful in uncovering such performance issues, it did not provide the anticipated level of process training. Additionally, standard-setting process materials that were intended to be distributed to the facilitators were not available at the time of the training.

Recognizing that facilitators may introduce individual differences that can result in slightly different instructions, Facilitator Handbooks were prepared for the content and process facilitators. The handbooks served as the "script" for providing instructions, describing the activities, and explaining the feedback and how it could be used in subsequent rounds. Additionally, the process facilitators were provided PowerPoint presentation materials for all workshops to further standardize panelist instructions. The handbooks that were available during the training were heavily modified for the pilot study and subsequently updated for each operational session. The Facilitator Handbooks are provided in Appendix K.

In addition to the two-day facilitator training, a facilitator meeting was held the day before each JSS session. In this meeting, the director of standard setting walked the facilitators through the four-day JSS process, discussing each element of the process, for the purpose of having a solid common understanding of the session's objectives, deliverables, and timing. Also discussed were improvements made to the JSS process as a result of lessons learned during the prior JSS session.

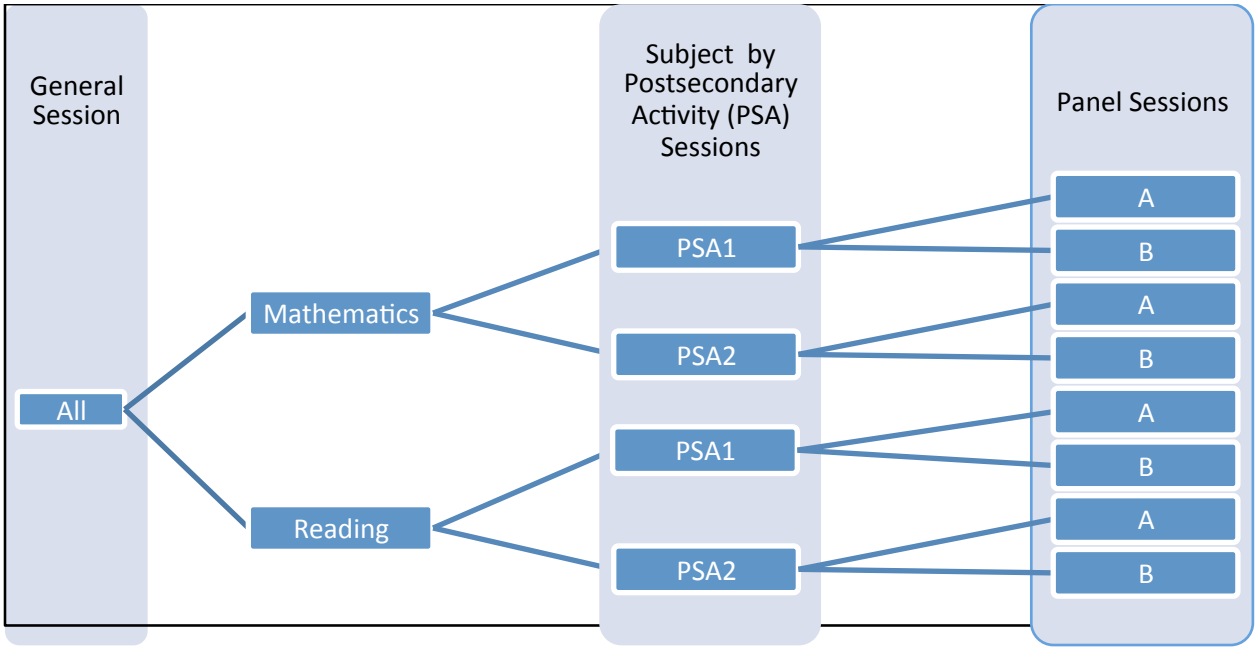
During each JSS session, the director of standard setting convened regular meetings with the facilitators. As needed, the director of standard setting scheduled meetings at the beginning of each day to discuss the prior day's activities and process evaluation results. Following each day's standard-setting activities, the director of standard setting convened a facilitator debriefing to discuss issues encountered that day. A debriefing meeting was also held at the end of the last day of the first and second operational JSS sessions to discuss the overall success of the session. While the majority of facilitators attended these meetings, in most cases not all facilitators were available to participate.

Panelist Training. Instructions were provided to the panelists at three different levels of grouping:

- During general sessions;
- In content area by postsecondary activity sessions; and
- Within each replicate panel.

These levels are represented in Figure 2.

Figure 2. Sessions



In general sessions, instructions were given to all panelists. The purpose of general sessions was to provide the same information and instructions to panelists across all panels. The first general session occurred at the beginning of the first day. Other general sessions were held throughout the four-day sessions in order to introduce major parts of the process. All instructions in the general sessions were provided by the director of standard setting.

The instructions provided for the different groupings of panelists were color-coded in the Agendas and the Briefing Booklets to indicate instructional sessions that were for all panelists in the workshop, for a pair of replicate panels within a content area/postsecondary area, or for a single replicate panel. In addition to the electronic copies of these materials that were sent to the panelists in advance of the JSS sessions, at the JSS sessions, panelists were given hard copies of these materials; the materials were also made available electronically within the CAB.

Bookmark Standard-Setting Design. As prescribed in the Design Document, a modified bookmark standard-setting methodology (ACT, Inc., 2007; ACT, Inc., 2010) was used for all JSS sessions. The bookmark methodology was introduced in 1996 (Lewis, Mitzel, & Green, 1996) and has since become the most widely used standard-setting methodology in state assessments (Council of Chief State School Officers, 2001).

The goals of each JSS session were as follows:

- Finalize the descriptions of knowledge and skills that describe what students need to know and be able to do to be minimally prepared for placement in a college credit-bearing course and/or a job-training course; and
- Determine the score on the NAEP scale that corresponds to the level of performance at the borderline (the cut score) of minimal preparedness and the percentage of students performing above the cut score.

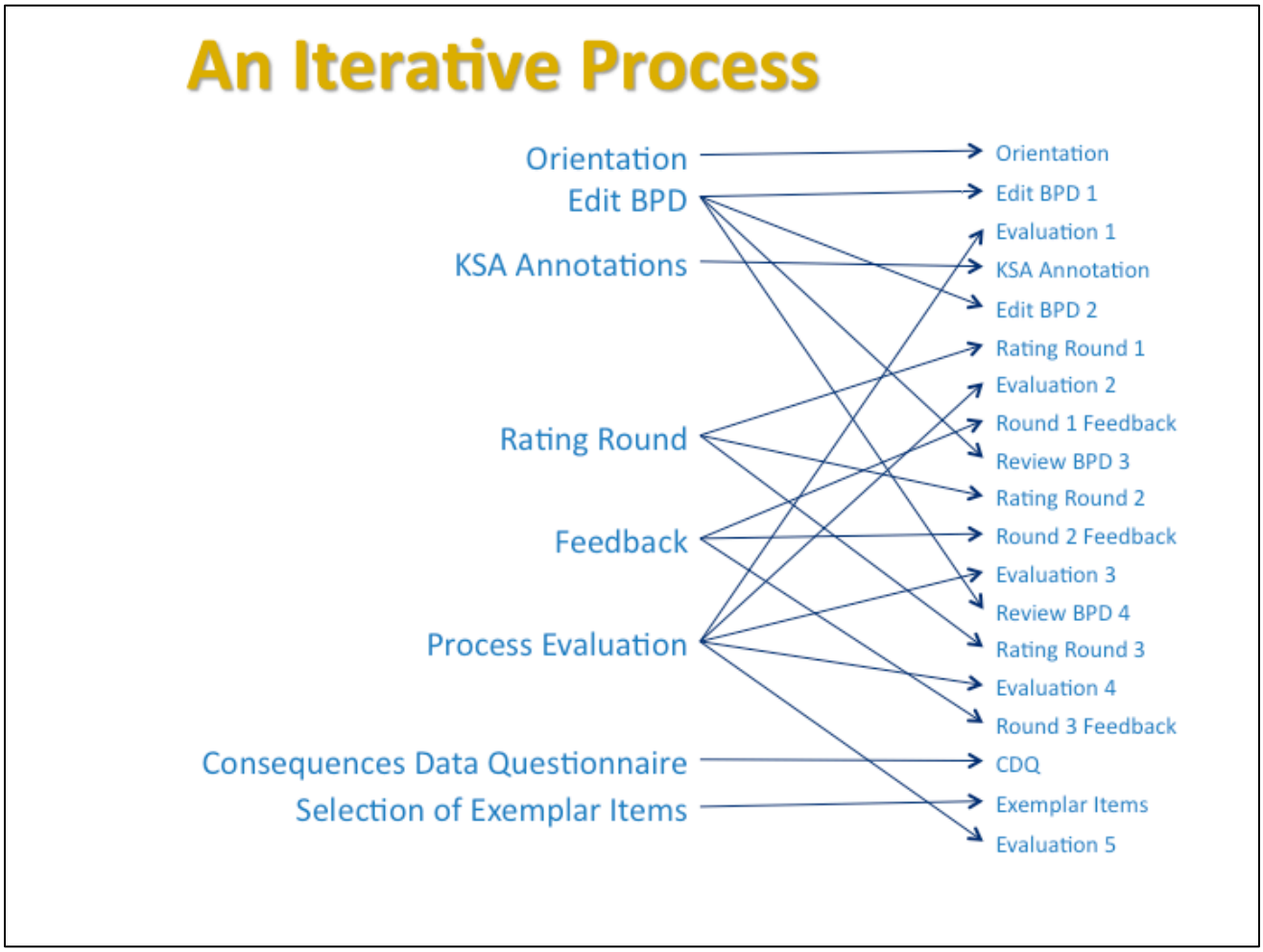
Additionally, test items illustrative of what students performing above the cut score know and can do were selected.

The standard-setting process consisted of the following types of activities:

- Training, whereby panelists were provided with information and instructions to enable them to provide informed judgments;
- Data collection, whereby panelists provided their judgments in three rating rounds; and
- Feedback, whereby panelists were provided with additional information based on the judgments they provided in the prior standard-setting rounds.

These three types of activities were executed through an iterative process, as presented in Figure 3 and described in the agendas used for the JSS sessions (provided in Appendix A). Each component of the process was described to the panelists in a Briefing Booklet, which was sent to them in advance of the JSS session. Copies of the Briefing Booklets for the JSS sessions are provided in Appendix J.

Figure 3. Basic Structure of the Judgmental Standard-Setting Process



This section outlines the standard-setting steps that were implemented in each of the pilot study and operational JSS workshops. Modifications to these steps that were implemented within particular JSS sessions are summarized at the end of this section and are discussed in greater detail in the sections of this report related to those JSS sessions.

Orientation. Each JSS session began with introductions of the project staff, an overview of the JSS process, background information on the Governing Board's research program on academic preparedness, information about 12th-grade student performance on the mathematics and reading NAEP, and a description of how panelists were recruited for participation.

In these sessions, the director of standard setting introduced the panelists to the bookmark standard-setting method that would be used to establish preparedness cut scores on NAEP. She described to the panelists the judgments that they would be making in the process of recommending cut scores and how they would be prepared; she also explained to the panelists that they would be using netbook computers and CAB software to perform most of their tasks during the four-day process. Panelists were also told that they would be trained to make informed judgments, as this was a necessary condition for the results to be deemed reasonable and valid. The presentation slides used in the orientation sessions are provided in Appendix L.

Taking and Scoring of a NAEP Test. In these sessions, panelists took a form of the 2009 reading or mathematics NAEP exam for grade 12 under conditions similar to those of an actual student administration. The purpose of this activity was not to test panelists' abilities but to give them an opportunity to experience the exam as the student experiences it, as test conditions—such as timing and instructions—are important factors to consider in standard-setting work. Also, by taking the exam and reviewing their own responses, panelists gained familiarity with NAEP test items and scoring guides.

NAEP Frameworks and BPD Discussion. Following the NAEP administrations, panelists separated into their content area groups within their respective postsecondary areas, where the JSS content facilitators provided the panelists with an orientation to the appropriate framework (i.e., mathematics or reading). This orientation provided a critical opportunity for the panelists to understand the frameworks, an important step in their reaching a useful understanding of what students should know and be able to do in reading or mathematics to be considered prepared for entry into college-level or occupational-training program courses. For most panelists, this orientation was a repetition of the framework overview provided during the online orientation

webinars, but the repetition was considered useful. During these discussions, panelists reviewed and discussed the topics and aspects of the assessment determined by the framework. Approximately ninety minutes were allotted for these sessions, and most panels consumed the full amount of time; the sessions were helpful for panelists to start focusing on how to relate the NAEP frameworks to the development of the BPDs.

Since the main focus of the standard-setting process was to identify the lower boundary of performance that is considered to represent minimal preparedness, it was important that panelists be familiar with and clear about the concept of borderline performance. Within the content area groups, content facilitators led discussions about borderline performance, introducing draft BPDs that had been developed using information collected from the panelists following the online orientation webinars. Panelists worked within their content area groups for these activities to further refine the BPDs and reach agreement regarding the knowledge and skills that students should have in order to be prepared for placement in a credit-bearing college-level course or in a job-training program. Panelists had further opportunities to review draft BPDs before each round of bookmarking (setting the cut score) for preparedness.

KSA Review of CROIBs. Following the orientation trainings and discussions surrounding NAEP frameworks and BPDs, panelists began to familiarize themselves with the items in their item pools. Panelists were also introduced to the CAB software. For each item, they were instructed to think about the KSAs needed to correctly answer each item or (for CR items) to score at a specific score point on the rubric.

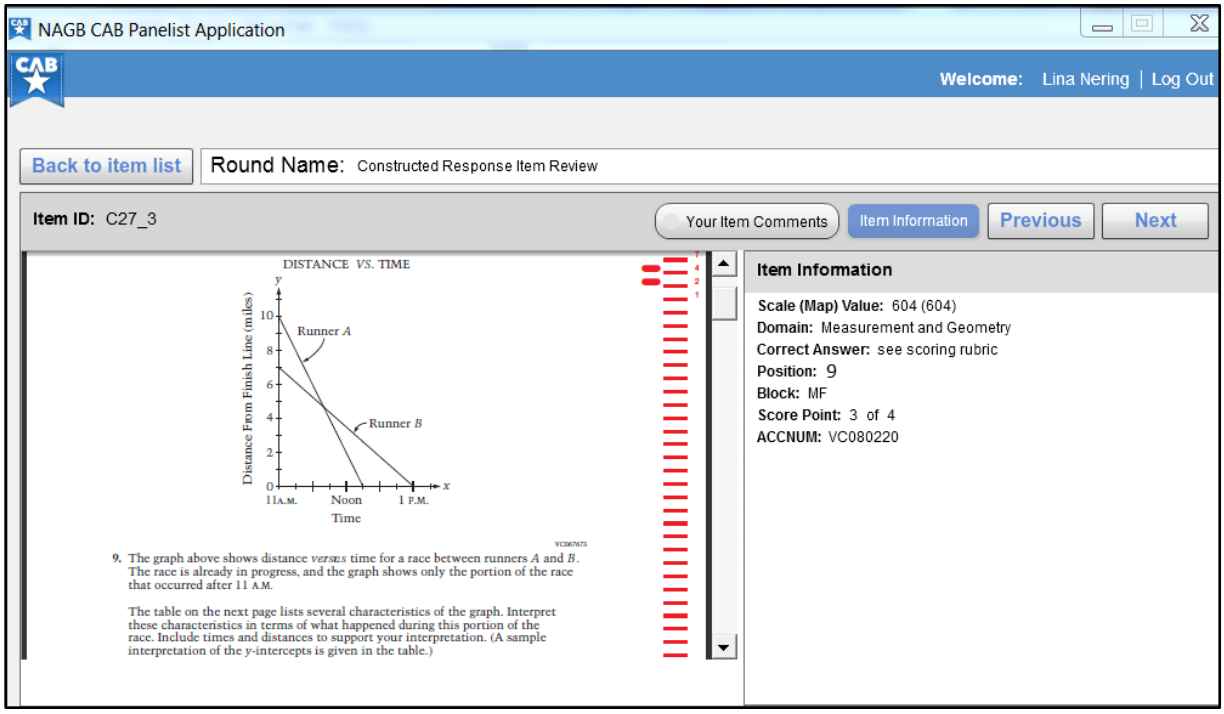
For this activity, panelists were given CROIBs (Constructed-Response Ordered Item Booklets), both within the CAB and in a paper book, which contained all CR items and score points. The

CR items were ordered by item difficulty of the full-credit response, from easiest to most difficult. In the CROIB, each item was included only once. Having panelists review the KSAs of CR items using the CROIB provided them the opportunity to interact with each CR item as a whole and not with one score level at a time. The following information was included in the CROIB:

- The page with the item, which included an information box with the item ID and the page numbers where the item's highest score point can be found in the OIB;
- The scoring rubric (note: KSAs were written for credited responses only); and
- An example of a student response at each score level, including incorrect (noncredited) responses.

In the CAB, score points for CR items were listed as individual items, indicated by an underscore and a number following the item ID. The number after the underscore corresponded to the rubric score. An example of a CR item as presented in the CAB is shown in Figure 4.

Figure 4. Sample CR Item as Presented in the CAB



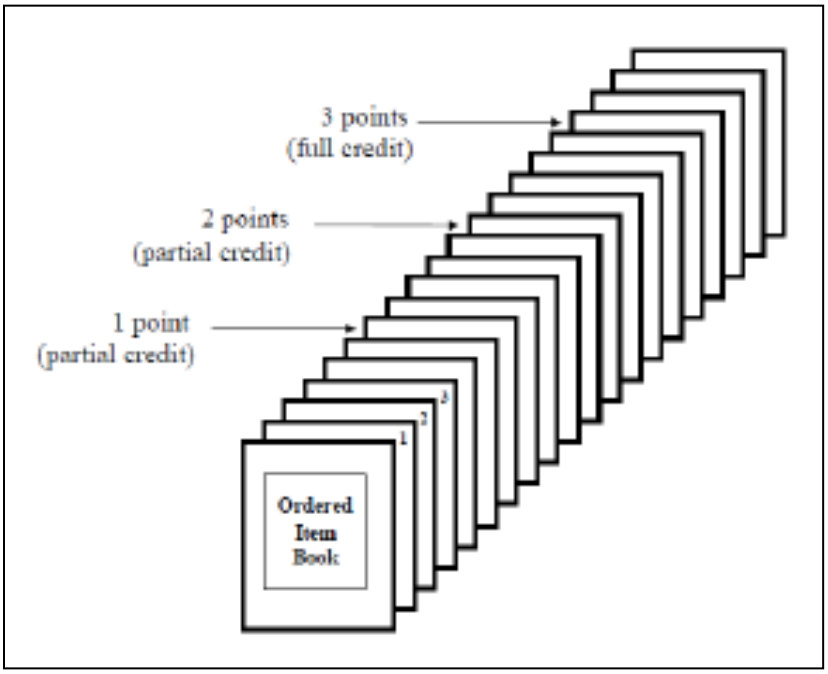
The KSA review process involved multiple stages, including review of constructed-response (CR) items alone and with multiple-choice (MC) items, review with the whole group, review within the table group, and independent review. Each stage was designed to help panelists gain a clearer understanding of what the assessment was measuring and the performance required of students. Panelists used the CAB software for the first time during each group's review of CR items, during which panelists viewed and noted the KSAs required to receive a specific score point or higher. Exact instructions (with screenshots) given to the panelists are provided in the Facilitator Handbooks in Appendix K. Panelists were also introduced to item maps for their use during the OIB KSA review.

KSA Review of OIBs. Following their review of CR items in the CROIBs to determine KSAs, panelists were given OIBs (Ordered Item Booklets), both within the CAB and in paper form, which contained all of the items in panelists' item rating pools, and were instructed to continue

the rating process. The OIBs contained all of the items in the item pool to which their panel would be exposed during rating rounds. The items were ordered by difficulty (starting with the easiest items) based on their scale location, using a response probability criterion of 0.67. Item difficulty was based on grade 12 student performance on the 2009 NAEP. Scoring rubrics for CR items were included. CR items appeared multiple times, once for each credited score level. Each OIB included one page for each MC item and at least two pages for each score point of a CR item (i.e., one page for the item and one page for the scoring rubric).

Facilitators reviewed a few items with the panelists so that panel groups could discuss the KSAs before continuing the KSA review task independently. Figure 5 is a conceptual presentation of the OIB showing a CR item scored at different levels.

Figure 5. Conceptual Representation of the OIB with CR Items



Because there was insufficient time for each panelist to review all items individually, panelists were assigned specific items to review. For mathematics, panelists on each panel (i.e., A and B)

were assigned between 59 and 61 score points to review. For reading, panelists on each panel were assigned between 65 and 67 score points to review. Score points are the total of the number of MC items (i.e., one score point per item) and the number of partial- or full-credit score levels for CR items. Each item was included in the list of at least two panelists in each table group; thus, each panelist had an opportunity to interact with each item during table group discussions.

Provision of Item Maps. Panelists were provided item maps, in paper form only (a sample item map is displayed in Figure 6). An item map is a spatially representative display of items by difficulty. Items were ordered on the map from easiest at the bottom to hardest at the top. 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 vertical difference 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. Panelists were further instructed to note the following characteristics of item maps:

- Items were ordered from easiest (at the bottom of the page) to hardest (at the top of the page);
- Items were mapped to a scale value according to the difficulty of the items, based on performance of students in the 2009 NAEP assessment; and
- CR items were mapped once for each credited response.

Figure 6. Sample Item Map

Redacted

Refining of BPDs. Following the KSA review process, Panels A and B reconvened to further refine the BPDs. This was the last BPD review prior to the first round of ratings. Panelists had opportunities to review the BPDs prior to each round of ratings: for the pilot study, the intent was to finalize the BPDs prior to the last round of ratings, with panelists instructed to review and modify them prior to each round of ratings; this process was refined in the operational sessions, however, during which the intent was to finalize the BPDs prior to the first round of ratings.

Bookmarking. Working independently, panelists translated the BPDs onto the score scale by placing a bookmark to divide the items in the OIB into two groups:

- Items easy enough for two-thirds of students who match the BPD to answer correctly; and
- Items too difficult for that expectation.

Within this process, panelists reviewed NAEP assessment items that were presented in an OIB. As indicated, within the OIB, each MC item is located at the ability level (scale score) that students would need in order to have a 0.67 probability of answering the item correctly. Each CR score point has a unique location on the scale, and the location of a given CR score point is defined as the position on the ability scale for which students have a 0.67 probability of achieving at least that score point—that is, that score point or higher. Once panelists reviewed these items and developed KSAs for all items assigned to them, they then evaluated each item against a description of borderline performance—the minimal level of performance required in the content area to represent preparedness for eligibility for acceptance into a job-training program or for placement in a credit-bearing college-level course—until they came to an item they judged to be too difficult for minimally prepared students. They then placed a bookmark immediately preceding that item to locate the cut score.

It is important to note that panelists were instructed to not place their bookmarks immediately upon finding an item that seemed too difficult; 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 the CAB automatically stored each panelist's selected cut score in the database. As a precaution against data loss, panelists were also asked to document the item ID where they placed their bookmarks and the corresponding scaled score.

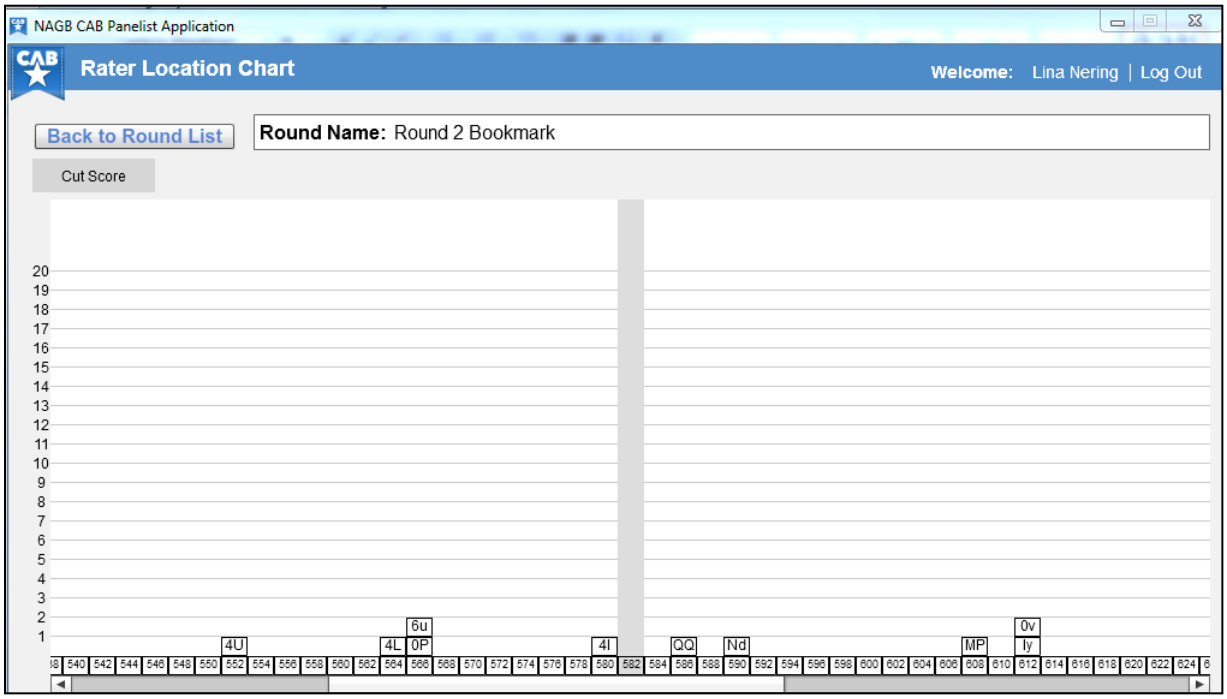
Additionally, panelists were asked to place their bookmarks in their paper OIBs corresponding to their placement of their bookmarks in the CAB.

In the pilot study, some job-training panelists stated that certain NAEP items were not relevant to their occupations; therefore, in the operational sessions, panelists were given specific instructions for placing their bookmarks when there were items that they deemed irrelevant to their training programs.

Provision of Feedback. Once bookmarks were set, panelists were provided with various forms of feedback to use in reconsidering and possibly refining their cut scores, including actual test booklets as examples of student performance on the assessment at the cut score and at the middle of each achievement level and spatially representative displays of items on the NAEP achievement scale (a linear transformation was used in the studies). Feedback was based on the cut score computed for the panel, which was the median of the cut scores for each panelist on the panel (A or B).

Rater Location Charts. Panelists were given a rater location chart after each round of bookmarking. The rater location chart displayed the distribution of cut scores for all panelists on Panel A or Panel B for a given round of bookmarking, thus providing information on the interrater consistency of the panelists' judgments. In the CAB, panelists were identified using codes to protect confidentiality. The rater location chart also displayed the median cut score for the panel group. An example of the rater location chart from the CAB is shown in Figure 7.

Figure 7. Sample Rater Location Chart from the CAB



Panelists were instructed to evaluate their individual cut scores relative to the other cut scores and to the median cut score, to help determine whether their conceptualizations and understandings of the BPD differed from those of others in the group.

Item Maps. After each round of bookmarking, panelists were also given item maps. Panelists were instructed to mark their personal cut scores and their panel's cut score on their item maps. In comparing their BPD to the KSAs of items around their individual cut scores and the KSAs of items around the panel cut score, panelists were to determine whether their understanding of the BPD was more consistent with the KSAs around their cut scores or around the panel's cut score. They were also to determine if there was another location for their bookmark that would better represent their understanding of the BPD than their current cut score.

Whole-Booklet Feedback. Panelists were given whole-booklet feedback after the first round of bookmarking in order to provide a holistic view of student performance. Twelve examples of student booklets from two forms were presented to panelists. Each whole booklet had the responses of a student to the two blocks of items in the form. These student responses had been scored, and a scaled score had been assigned to each booklet. These booklets were selected such that they were distributed with respect to the round 1 median cut score. Six booklets were selected from one form of the NAEP assessment and six from a second form. Two booklets from each form (for a total of four out of the 12) 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).

Panelists were instructed to compare the knowledge, skills, and abilities exhibited in each booklet to their understanding of the BPD to gain an insight of whether their cut scores were appropriately located. Panelists were instructed to not score the booklets. Instead, they were to try to understand how the performance of students in each text booklet compared to their understanding of minimal performance required for preparedness.

Booklet Score Chart. Panelists were also provided booklet score charts after the first round of bookmarking: charts that showed the expected percentage of points earned by the student for each of the 12 booklets included in the whole-booklet feedback. The chart also indicated the location of each booklet on the NAEP-like scale, as well as its location relative to the panel cut score. A sample booklet score chart is displayed in Figure 8.

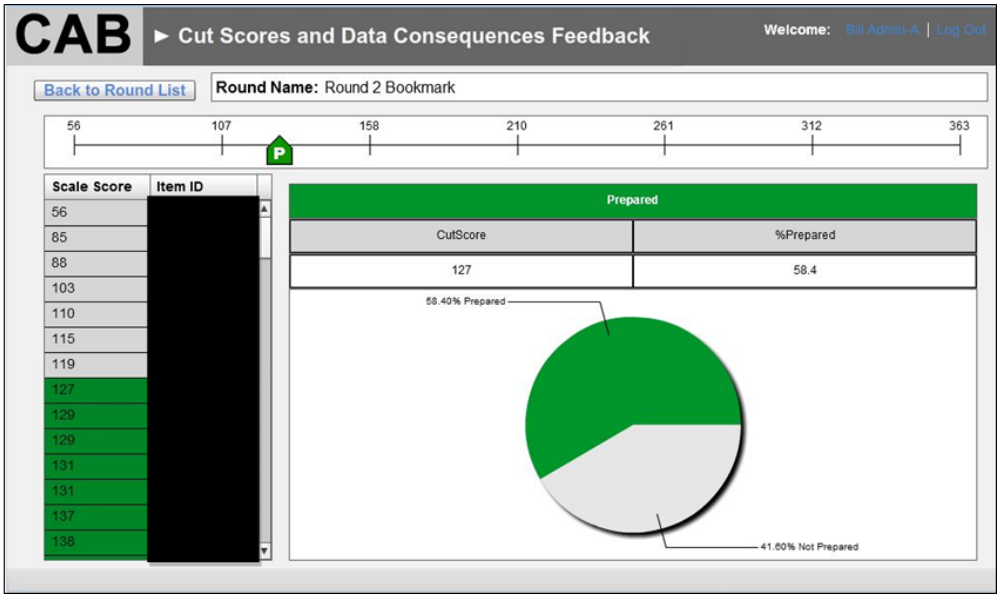
Figure 8. Sample Booklet Score Chart

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					

Consequences Data Feedback. After the second round of ratings, panelists were presented with feedback that displayed the percentage of 12th-grade students in 2009 who scored at or above the group cut score. This consequences data feedback was presented in the CAB, allowing the percentage of students above the cut score and the list of items mapped below and above the cut score to change as panelists adjusted the placement of the cut score. An image of the consequences data feedback screen in the CAB is displayed in Figure 9.

Figure 9. Sample Consequences Data Feedback

Partially Redacted (Item ID Column only)



Within each panel, panelists discussed whether the consequences data seemed reasonable in light of the BPD (what students *should know and be able to do*) and in light of what panelists knew about student performance in this content area (what students *know and can do*). Panelists were asked to consider the following questions and instructions:

- *Having seen these data, do you want to adjust your cut score?*

- *Did students generally perform better or worse than you expected?*
- *Whatever your reaction to the consequences data, you should keep in mind that it is the BPD that is to serve as the criterion for placing your bookmark.*

Consequences Data Questionnaire. After panelists set their preparedness cut scores for the third and final time, they were once again provided with consequences data based on their ratings as well as consequences data based on the ratings provided by panelists from the other replicate panel. They were asked to share their opinions regarding the percentage of students at or above the cut score for their group. A questionnaire was developed to collect their opinions regarding the consequences of their group's cut scores.

Selection of Exemplar Items. After the bookmarking rounds were completed, panelists made recommendations for exemplar items—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 were identified from items in the booklets that were used when panelists took the NAEP assessment on Day 1. An item or score level (on a CR item) from this pool was identified as a potential exemplar if the probability of a correct response was at least 0.67 at the cut score.

In this task, panelists rated the potential exemplar items as to whether they should be used to illustrate preparedness. Panelists were asked to indicate whether they felt the items should definitely be used, were OK 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 the CAB independently.

Process Evaluations. Panelists were asked to complete evaluation forms after each major activity or phase of the standard-setting process. The evaluation forms included many statements that panelists responded to using a rating scale (such as “strongly agree” to “strongly disagree”). In addition, panelists were asked to provide written responses to more general, open-ended questions and were given space to comment on any aspect they felt would be helpful to the Governing Board for evaluating the process.

The staff reviewed the evaluations at the end of each day to see if panelists were experiencing any difficulties with performing the tasks. Evaluation data helped to improve the standard-setting process and were an important source of validity evidence for the cut-score recommendations. Evaluation data are included in each JSS workshop section of this report. As indicated in the agenda provided for each JSS session, five evaluation questionnaires were administered; sample process evaluations (for the pilot study) can be found in Appendix M.

Differences in Implementation Across JSS Sessions. A pilot study of the process was implemented using the procedures planned for the operational standard-setting sessions. The purpose of the pilot study was to determine whether modifications for training, instructions, materials, timing, and logistics were needed, as well as to provide an opportunity for the facilitators to practice the process before moving to the operational setting. One major change resulting from the pilot study and implemented in the operational sessions was the addition of a general session prior to every major standard-setting activity. This change was intended to ensure that panelists on all panels heard the same instructions from the director of standard setting. Another change for the purpose of ensuring standardized instructions across panels was to provide PowerPoint slides to each facilitator with instructions for panelists to follow for each task. Changes to the process for finalizing the BPDs were discussed previously.

One aspect of implementation that was modified over operational sessions was the amount of training, support, and information provided to the panelists for the KSA review. The ability of the CAB to collect panelists' KSAs provided rich information that led to improvements in the instructions provided. It also provided information on how well panelists were attuned to the language of the framework and the assessment. Information from the CAB regarding KSAs corroborated observations regarding some facilitators' understanding of the panelists' tasks and panelists' understanding of the NAEP content. Collectively, this information prompted changes in the process of developing KSAs that culminated in a decision to provide item descriptions to panelists in the final operational session for use in writing KSAs.

Another 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 passage made the KSA review of CR 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.

Data Analysis

Within each JSS workshop, each panel completed three rounds of bookmarking. The results for each of the rounds are presented in the respective workshop's section of this report, with results presented for each content area and replicate panel by postsecondary area (i.e., college-preparedness and Automotive Master Technician). The results reported include the following: (1) median cut score; (2) standard error of the median, including an empirical-based method (Maritz & Jarrett, 1978) and a bootstrap method (Efron & Gong, 1983); (3) mean absolute deviation (MAD); (4) percent of students considered prepared based on the median cut score; (5) a

summary of the changes in cut scores made by individual panelists between rounds; and (6) a comparison of cut scores based on medians and means. The first four sets of results in each section are presented in the same table for each round. MAD results are further summarized graphically so that comparisons can be made between rounds. The fifth and sixth sets of results are presented in tables at the end of each results presentation as well.

The median panel cut score was used to calculate round cut scores, as it has been used in past NAEP bookmark-based standard-setting studies, with the rationale that the median is less sensitive to outliers than the mean (ACT, Inc., 2005a, 2005b). Outliers can easily occur due to inexperience as a panelist or due to the panelist intending to influence the mean. Use of the median helps prevent these influences. For informational purposes, mean cut scores are presented at the end of the round results.

The standard error of a cut score is an estimate of the uncertainty in the reported cut score due to various sources of error. Due to the difficulty surrounding calculating the standard error of a median, two nonparametric standard error procedures were used: an empirical-based method (EmpSE) and a bootstrap method (BootSE). The Technical Report for these studies (Measured Progress & WestEd, 2011) provides a more in-depth treatment of how these standard errors were calculated. Reported standard errors for rounds 2 and 3 should be interpreted with caution, as panelist judgments for these rounds are no longer independent due to the standard-setting process implemented; after round 1, panelist cut scores are influenced by the group cut scores and cut score distributions. Panelists are generally more comfortable when their cut scores are close to one another, so there is a regression to the group cut score (ACT, Inc., 2005a). Estimates of the standard error of the final cut score do not account for this regression to the median. For this reason, estimates of the standard error at the final round tend to be smaller and are more likely to

overestimate differences between replications of a method using the same item pools but different groups of panelists. In addition, cut scores established in rounds 2 and 3 are based on the baseline established in the first round and do not tend to vary substantially from the previous round.[10]

Another indication of the variability of bookmark placements within a panel is the mean absolute deviation (MAD), which is the average difference between each panelist's cut score and the median cut score and is further explained in the Technical Report. As panelists review results and feedback together, outliers and variability tend to decrease as panelists gain a shared understanding of the borderline preparedness description and once they have learned the round 1 panel cut score.

Finally, a study of the change in cut scores by round provides additional information about how panelists were responding to the feedback provided (i.e., panel cut scores and distributions after rounds 1 and 2, whole-booklet feedback after round 1, and impact data after round 2). A table near the end of each results section reports the number of panelists whose cut scores increased, decreased, or had no change from the previous round.

[10] Note that even if there are substantial variations among individual panelists' cut scores, there still might not be a variation in the median.

Pilot Study

The pilot JSS study, conducted on April 26–29, 2011, was intended to evaluate the methodology, facilities, materials, software, training, and study logistics in order to identify needed refinements for the operational sessions. The postsecondary areas included in the pilot study were college-preparedness and the Automotive Master Technician occupation.

Pilot Study Panelists

Panelists for the pilot study college-preparedness workshop were recruited from 17 different postsecondary institutions and seven different secondary institutions. Panelists for the pilot study Automotive Master Technician workshop were recruited from 19 different Automotive Master Technician job-training programs.

The sampling plan for the pilot study college-preparedness workshop stipulated representation of both public (75%) and private (25%) postsecondary institutions. As displayed in Table 7, all postsecondary college-preparedness panelists came from public institutions. The selected institutions closely reflected the targeted distribution for institutional selectivity. Looking at institutional size, however, somewhat more small institutions were selected than planned (35% selected, compared to the target of 24%), with fewer medium (35% versus the target of 42%) and large (29% versus the target of 34%) institutions represented.

Table 7. Pilot Study: College-Preparedness Postsecondary Institution Distributions[11]

Post-secondary Area	Content Area	Panel	Private/Public Status		Selectivity			Size			Total Number of Institutions
			Public	Private	Open	Mod.	High	Small	Med.	Large	
College-Preparedness PILOT	Math	A	4 (100%)	0 (0%)	3 (75%)	1 (25%)	0 (0%)	0 (0%)	3 (75%)	1 (25%)	4 (100%)
		B	5 (100%)	0 (0%)	2 (40%)	2 (40%)	1 (20%)	3 (60%)	1 (20%)	1 (20%)	5 (100%)
	Reading	A	4 (100%)	0 (0%)	2 (50%)	2 (50%)	0 (0%)	2 (50%)	2 (50%)	0 (0%)	4 (100%)
		B	4 (100%)	0 (0%)	3 (75%)	1 (25%)	0 (0%)	1 (25%)	0 (0%)	3 (75%)	4 (100%)
College-Preparedness Institution Totals (N = 17)			17 (100%)	0 (0%)	10 (59%)	6 (35%)	1 (6%)	6 (35%)	6 (35%)	5 (29%)	17 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of institution distribution across replicate panels.

The sampling plan for college-preparedness secondary-level institutions called for a distribution of institutions across the categories of urban (34%), suburban (41%), and town/rural (25%). As shown in Table 8, while the representation of suburban secondary institutions (43%) met the target, urban institutions were overrepresented (at 43%) and town/rural institutions were underrepresented (at 14%). The sampling target for institutional size was not met, with no medium-sized secondary institutions sampled and the sample including 29% small institutions (compared to the target of 14%) and 71% large institutions (compared to the target of 64%).

[11] Throughout this report, the sums of table column and/or row cells may not equal the reported total for a given column or row due to rounding.

Table 8. Pilot Study: College-Preparedness Secondary-Level Institution Distributions

Post-secondary Area	Content Area	Panel	Urbanicity			Size			Total Number of Institutions
			Urban	Sub-urban	Rural	Small	Med.	Large	
College-Preparedness PILOT	Math	A	1 (50%)	1 (50%)	0 (0%)	0 (0%)	0 (0%)	2 (100%)	2 (100%)
		B	0 (0%)	0 (0%)	1 (100%)	1 (100%)	0 (0%)	0 (0%)	1 (100%)
	Reading	A	1 (50%)	1 (50%)	0 (0%)	1 (50%)	0 (0%)	1 (50%)	2 (100%)
		B	1 (50%)	1 (50%)	0 (0%)	0 (0%)	0 (0%)	2 (100%)	2 (100%)
College-Preparedness Institution Totals (N = 7)			3 (43%)	3 (43%)	1 (14%)	2 (29%)	0 (0%)	5 (71%)	7 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of institution distribution across replicate panels.

The process of acquiring the sample of nominees and of recruiting both college-preparedness and Automotive Master Technician panelists for participation is described in the Panelist Recruitment Plan subsection of this report (pp. 26–38). The Design Document called for a total of 48 panelists for the pilot study: 24 for the college-preparedness workshop (12 in mathematics and 12 in reading, with each group divided into two replicate panels of six panelists) and 24 for the Automotive Master Technician preparedness workshop (12 in mathematics and 12 in reading, with each group divided into two replicate panels of six panelists). This target was achieved for panelists in the college-preparedness workshop. However, the recruitment challenges previously described resulted in the recruitment of only 19 Automotive Master Technician panelists: 10 for mathematics and nine for reading. Panelists were assigned to replicate panels and table groups based upon background and demographic characteristics in order to render the replicate panels as similar as possible; for the Automotive Master Technician replicate panels, this meant table sizes of as few as two panelists.

The target was to include two secondary-level instructors and four postsecondary instructors on each pilot college-preparedness replicate panel. Each pilot college-preparedness replicate panel did include two secondary-level instructors, with one exception—in mathematics Panel B, a secondary-level instructor left the project just prior to the pilot study and was replaced by a panelist from a postsecondary four-year institution. As shown in Table 9, among postsecondary panelists, there was comparable representation of two-year postsecondary institutions (33% of college-preparedness panelists) and four-year postsecondary institutions (38% of college-preparedness panelists). Five of the eight college-preparedness postsecondary reading instructors were recruited from traditional English- or composition-type programs (e.g., literature, composition, English, developmental reading), while the remaining three instructors represented four-year biological science and philosophy programs and a two-year communications program. As shown in Table 10, the college-preparedness mathematics replicate panels were evenly distributed by gender, although women outnumbered men on the two reading replicate panels (83% versus 17% on Panel A; 67% versus 33% on Panel B); panelists who reported their race/ethnicity were predominantly White/Caucasian (79% across all panels), with minimal ethnic diversity on three of the college-preparedness panels and none on the fourth (mathematics Panel A).

All Automotive Master Technician panelists were recruited from postsecondary institutions, as illustrated in Table 9. While the majority (84%) of Automotive Master Technician panelists represented public two-year community or technical programs, one mathematics Panel A panelist taught in a technical program embedded within a public four-year institution, and two reading Panel A panelists taught at proprietary technical schools. As shown in Table 10, no gender balance was achieved on any of the Automotive Master Technician panels, as all panelists were

male, and only one panelist (5%) reported his ethnicity as Non-White/Caucasian. The remaining 18 panelists (95%) reported themselves to be White/Caucasian.

Table 9. Pilot Study: Panelist Distribution by Institution Type

Post-secondary Area	Content Area	Panel	Type of Institution					Total Panelists
			4-Year Public	4-Year Private	2-Year Public (Community/ Technical)	2-Year Private	Secondary	
College-Preparedness PILOT	Math	A	2 (33%)	0 (0%)	2 (33%)	0 (0%)	2 (33%)	6 (100%)
		B	3 (50%)	0 (0%)	2 (33%)	0 (0%)	1 (17%)	6 (100%)
	Reading	A	1 reading (17%) 1 "other" (17%)	0 (0%)	1 reading (17%) 1 "other" (17%)	0 (0%)	2 (33%)	6 (100%)
		B	1 reading (17%) 1 "other" (17%)	0 (0%)	2 reading (33%)	0 (0%)	2 (33%)	6 (100%)
College-Preparedness Totals (N = 24)			9 (38%)	0 (0%)	8 (33%)	0 (0%)	7 (29%)	24 (100%)
Automotive Master Technician PILOT	Math	A	1 (20%)	0 (0%)	4 (80%)	0 (0%)	N/A	5 (100%)
		B	0 (0%)	0 (0%)	5 (100%)	0 (0%)	N/A	5 (100%)
	Reading	A	0 (0%)	0 (0%)	2 (50%)	2 (50%)	N/A	4 (100%)
		B	0 (0%)	0 (0%)	5 (100%)	0 (0%)	N/A	5 (100%)
Automotive Master Technician Totals (N = 19)			1 (5%)	0 (0%)	16 (84%)	2 (11%)	N/A	19 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

Table 10. Pilot Study: Panelist Distribution by Demographic Characteristics

Postsecondary Area	Content Area	Panel	Gender		Race/Ethnicity			Total Panelists
			Female	Male	White/ Caucasian	Non-White/ Caucasian	Not Specified	
College-Preparedness PILOT	Math	A	3 (50%)	3 (50%)	6 (100%)	0 (0%)	0 (0%)	6 (100%)
		B	3 (50%)	3 (50%)	4 (67%)	2 (33%)	0 (0%)	6 (100%)
	Reading	A	5 (83%)	1 (17%)	5 (83%)	1 (17%)	0 (0%)	6 (100%)
		B	4 (67%)	2 (33%)	4 (67%)	1 (17%)	1 (17%)	6 (100%)
College-Preparedness Totals (N = 24)			15 (63%)	9 (38%)	19 (79%)	4 (17%)	1 (4%)	24 (100%)
Automotive Master Technician PILOT	Math	A	0 (0%)	5 (100%)	5 (100%)	0 (0%)	0 (0%)	5 (100%)
		B	0 (0%)	5 (100%)	4 (80%)	1 (20%)	0 (0%)	5 (100%)
	Reading	A	0 (0%)	4 (100%)	4 (100%)	0 (0%)	0 (0%)	4 (100%)
		B	0 (0%)	5 (100%)	5 (100%)	0 (0%)	0 (0%)	5 (100%)
Automotive Master Technician Totals (N = 19)			0 (0%)	19 (100%)	18 (95%)	1 (5%)	0 (0%)	19 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

While geographic distribution was not specified as part of the sampling plan, an attempt was made to balance geographic location as much as possible across replicate panels and table groups within panels while maintaining balance across institution and panelist characteristics. The geographic distribution of the pilot study panelists is shown in Table 11. A map indicating the U.S. Census Bureau census regions can be found in Appendix N.

Table 11. Pilot Study: Geographic Distribution of Panelists

Postsecondary Area	Content Area	Panel	Geographic Region*				Total Panelists
			Northeast	South	Midwest	West	
College-Preparedness PILOT	Math	A	1 (17%)	3 (50%)	0 (0%)	2 (33%)	6 (100%)
		B	2 (33%)	2 (33%)	1 (17%)	1 (17%)	6 (100%)
	Reading	A	0 (0%)	3 (50%)	0 (0%)	3 (50%)	6 (100%)
		B	0 (0%)	0 (0%)	2 (33%)	4 (67%)	6 (100%)
College-Preparedness Totals (N = 24)			3 (13%)	8 (33%)	3 (13%)	10 (42%)	24 (100%)
Automotive Master Technician PILOT	Math	A	1 (20%)	1 (20%)	2 (40%)	1 (20%)	5 (100%)
		B	1 (20%)	1 (20%)	2 (40%)	1 (20%)	5 (100%)
	Reading	A	2 (50%)	0 (0%)	2 (50%)	0 (0%)	4 (100%)
		B	1 (20%)	0 (0%)	3 (60%)	1 (20%)	5 (100%)
Auto. Master Technician Totals (N = 19)			5 (26%)	2 (11%)	9 (47%)	3 (16%)	19 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.
*Based upon U.S. Census Bureau census regions.

As shown in Table 10, all four census regions were represented in the pool of college-preparedness pilot study panelists, with the lowest percentages of panelists (13% each) representing the Northeast and the Midwest and the highest percentage (42%) representing the West. However, there was not full geographic representation within all of the replicate panels. Only mathematics Panel B had representatives from all four regions.

All four census regions were also represented in the pool of Automotive Master Technician pilot study panelists, with the lowest percentage of panelists (11%) representing the South and the highest percentage (47%) representing the Midwest. However, there was not full geographic representation within all of the replicate panels. While both mathematics panels had

representatives from all four regions, panelists on reading Panel B represented only three regions, and panelists on reading Panel A represented only two regions.

During the recruitment process, job-training panelists were asked to indicate the predominant student population served by their programs (i.e., students coming directly from high school or students returning to school after a year or more of absence). This information was not used to assign panelists to panels or table groups; however, it was helpful in understanding the perspectives brought to the standard-setting process by the job-training panelists. Information about student populations served by the programs represented by pilot study job-training panelists is shown in Table 12.

Table 12. Pilot Study: Student Populations Served by Job-Training Panelists

Postsecondary Area	Content Area	Panel	Predominant Student Population Served by Program		Total Panelists
			Students Coming Directly from High School	Students Returning to School after Absence	
Automotive Master Technician PILOT	Math	A	4 (80%)	1 (20%)	5 (100%)
		B	3 (60%)	2 (40%)	5 (100%)
	Reading	A	4 (100%)	0 (0%)	4 (100%)
		B	3 (60%)	2 (40%)	5 (100%)
Auto. Master Technician Totals (N = 19)			14 (74%)	5 (26%)	19 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

As reported by panelists, the majority of Automotive Master Technician panelists (74%) represented job-training programs that predominantly served students coming directly from high school. Across the four panels, only five panelists (26%) taught in job-training programs that predominantly served students returning to school after an absence.

Pilot Study JSS Process

The pilot study implemented the JSS process as described in the Judgmental Standard-Setting Process subsection of this report (pp. 39–69). The pilot study agenda can be found in Appendix A.

Pilot Study BPDs and Numerical Results

The process by which the BPDs were developed and refined for use in the pilot study bookmarking is described in the Development of Borderline Performance Descriptions subsection of this report (pp. 40–42). The preliminary mathematics and reading pilot study BPDs that were developed by content facilitators, using panelists' responses to the online Content Objectives form, prior to the JSS pilot study session and the final BPD versions agreed upon by the mathematics and reading pairs of replicate panels for each postsecondary area are provided in Appendix O.

Table 13 displays how to interpret panel abbreviations used in pilot study numerical results tables and graphical displays.

Table 13. Pilot Study: Abbreviations Used to Describe Panels

Abbreviation	Content Area	Postsecondary Area	Panel
MCA	Mathematics	College-Preparedness	A
MCB			B
RCA	Reading	College-Preparedness	A
RCB			B
MAA	Mathematics	Automotive Master Technician	A
MAB			B
RAA	Reading	Automotive Master Technician	A
RAB			B

College-Preparedness Workshop Results. The following tables display standard-setting results for the pilot study college-preparedness workshop.

When round 1 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The round 1 results are presented in Table 14.

Table 14. Pilot Study: College-Preparedness Round 1 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	214	1.6	1.4	2.7	3.4
	B	181	8.0	7.2	9.8	22.1
Reading	A	309	10.8	8.6	16.5	31.5
	B	337	9.9	9.3	14.8	9.3

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 2 bookmarks were placed, the CAB recalculated the median cut scores for each panel and the associated impact data. The results of the panels' round 2 judgments are presented in Table 15.

Table 15. Pilot Study: College-Preparedness Round 2 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	210	4.2	4.7	6.0	4.4
	B	203	7.1	6.9	10.0	7.1
Reading	A	313	5.1	3.2	6.8	27.5
	B	342	4.3	4.5	5.0	7.0

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 3 bookmarks were placed, the CAB once again calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 3 decisions are presented in Table 16.

Table 16. Pilot Study: College-Preparedness Round 3 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	205	4.2	4.7	8.5	6.3
	B	196	7.1	6.9	18.1	10.8
Reading	A	313	5.1	3.2	4.3	27.5
	B	331	4.3	4.5	7.0	12.9

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

Figure 10 summarizes the MAD results across rounds for the four college-preparedness panels. One would typically expect the variability of panel cut scores to decrease as (1) panelists gain a shared understanding of the borderline performance description and (2) panelists learn the round 1 panel cut score. However, within these panels, MAD increased across rounds in most cases, except for college-preparedness reading Panels A (across all rounds) and B (between rounds 1 and 2). The mathematics panels experienced greater increases in variability overall. The increase in variability among panelist cut scores in later rounds is not easily explainable. However, one hypothesis may be that panelists were reacting strongly and differently to the post-round feedback.

Figure 10. Pilot Study: College-Preparedness Mean Absolute Deviation (MAD) of Cut Scores by Panel

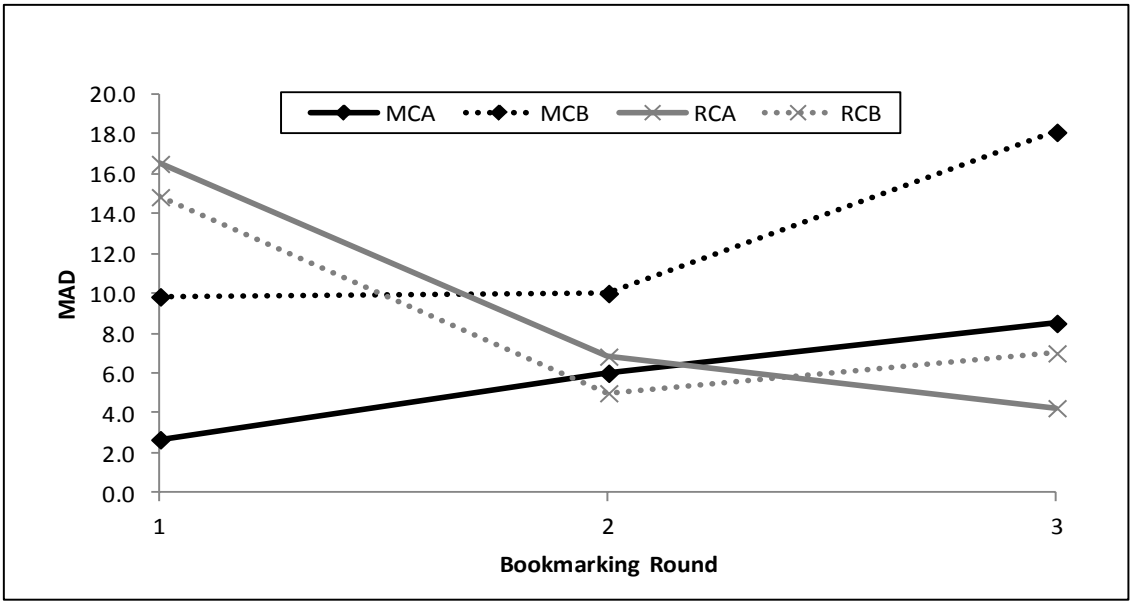


Table 17 summarizes the changes in cut scores made by individual panelists between rounds, while Table 18 presents a comparison of median and mean cut scores for each panel. One would typically expect to see either the same number of panelists or fewer panelists change their cut scores between rounds 2 and 3 than between rounds 1 and 2. This pattern is observed in Table 17.

Table 17. Pilot Study: College-Preparedness Round-to-Round Cut Score Changes by Panel

Panel	Round	Increased N (%)	No Change N (%)	Decreased N (%)
MCA	R1–R2	0 (0.0%)	1 (16.7%)	5 (83.3%)
	R2–R3	0 (0.0%)	2 (33.3%)	4 (66.7%)
MCB	R1–R2	5 (83.3%)	0 (0.0%)	1 (16.7%)
	R2–R3	0 (0.0%)	4 (66.7%)	2 (33.3%)
RCA	R1–R2	3 (50.0%)	2 (33.3%)	1 (16.7%)
	R2–R3	0 (0.0%)	4 (66.7%)	2 (33.3%)
RCB	R1–R2	2 (33.3%)	3 (50.0%)	1 (16.7%)
	R2–R3	0 (0.0%)	3 (50.0%)	3 (50.0%)

Differences between mean and median cut scores, as shown in Table 18, are due to the presence of outliers (i.e., panelists who set cut scores significantly higher or lower than their peers). The largest absolute difference between the mean and median cut scores in Table 18 occurs for college-preparedness reading Panel A in round 2 (6.2), where the direction of the difference would result in a higher cut score being set if the mean instead of the median were used. Overall, in the third round, the effect of outliers on mean cut scores is minimal.

Table 18. Pilot Study: College-Preparedness Comparison of Cut Scores Based on Medians and Means

Panel	Round 1			Round 2			Round 3		
	Median	Mean	Median–Mean	Median	Mean	Median–Mean	Median	Mean	Median–Mean
MCA	414.5	415.3	**-0.8**	410.5	406.3	**4.2**	405.5	402.2	**3.3**
MCB	403.0	407.2	**-4.2**	424.5	424.7	**-0.2**	418.0	416.7	**1.3**
RCA	510.0	514.5	**-4.5**	514.0	520.2	**-6.2**	514.0	514.3	**-0.3**
RCB	558.5	558.5	**0.0**	564.0	561.7	**2.3**	553.0	554.7	**-1.7**

Automotive Master Technician Workshop Results. The following tables display standard-setting results for the pilot study Automotive Master Technician workshop.

When round 1 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The round 1 results are presented in Table 19.

Table 19. Pilot Study: Automotive Master Technician Round 1 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	197	18.6	18.5	19.2	10.3
	B	172	16.1	16.0	18.2	31.0
Reading	A	346	20.4	20.2	27.3	5.5
	B	317	14.5	15.8	17.0	23.7

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 2 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 2 judgments are presented in Table 20.

Table 20. Pilot Study: Automotive Master Technician Round 2 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	194	12.9	12.8	13.8	12.1
	B	183	7.3	5.6	8.2	20.4
Reading	A	346	17.6	17.3	23.0	5.5
	B	321	5.6	5.5	6.0	20.2

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 3 bookmarks were placed, the CAB once again calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 3 decisions are presented in Table 21.

Table 21. Pilot Study: Automotive Master Technician Round 3 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	189	5.9	5.0	6.8	15.5
	B	183	3.6	3.4	4.6	20.4
Reading	A	327	6.9	7.5	7.3	15.5
	B	321	5.9	6.2	6.8	20.2

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.

[b] Empirical-based Standard Error.

[c] Bootstrap Standard Error.

[d] Mean Absolute Deviation.

[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

Figure 11 summarizes the MAD results across rounds for the Automotive Master Technician panels. Unlike the college panels, the automotive MAD panel results more closely follow the expectation that variability of panel cut scores decrease as panelists gain a shared understanding of the borderline preparedness description and have learned the round 1 panel cut score. Figure 11 shows a slight increase in variability for reading Panel B between rounds 2 and 3, which suggests that this panel reacted more strongly and differently to the impact data that was presented.

Figure 11. Pilot Study: Automotive Master Technician Mean Absolute Deviation (MAD) of Cut Scores by Panel

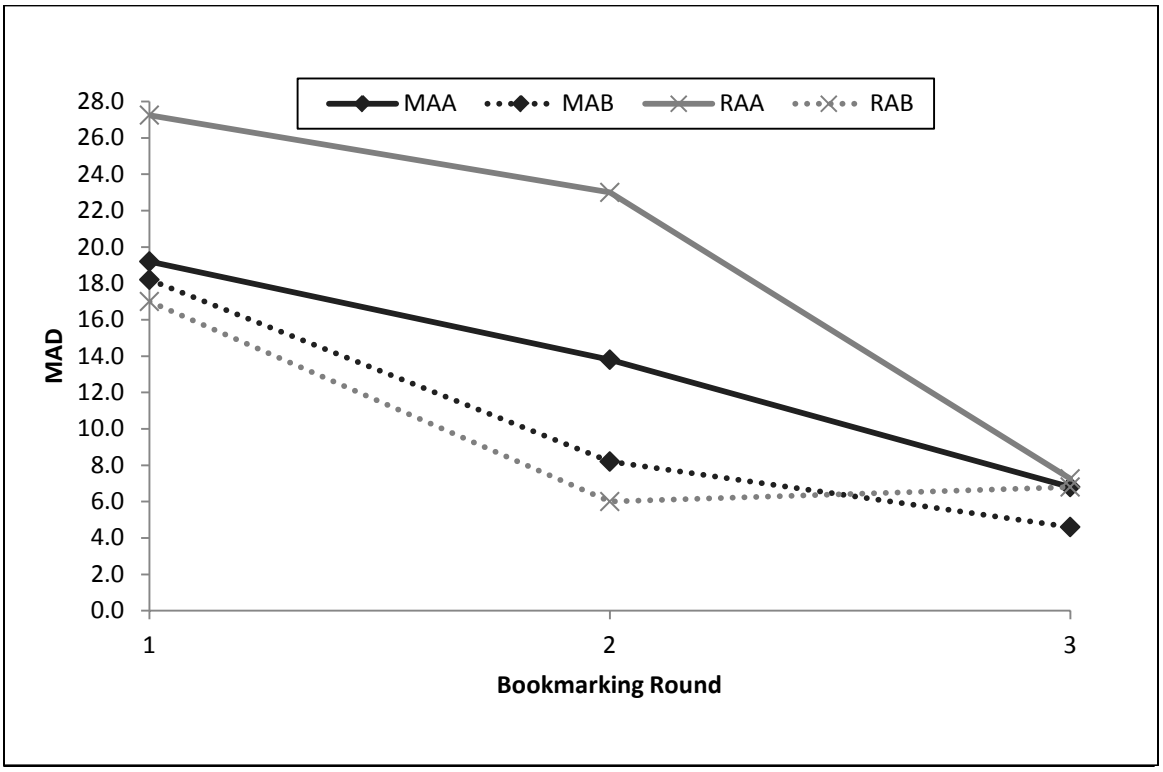


Table 22 summarizes the changes in cut scores made by individual panelists between rounds, while Table 23 presents a comparison of median and mean cut scores for each panel. The expectation was that the same number of panelists or fewer panelists changed their cut scores between rounds 2 and 3 than between rounds 1 and 2. This was observed for three of the four panels in Table 22. One more panelist on reading Panel A changed her/his cut score between rounds 2 and 3 than between rounds 1 and 2, indicating that this panel may have reacted more strongly to the impact data presented than other panels. As was true for the college panels, the effect of outliers on mean cut scores in the third round was minimal, as observed in Table 23.

Table 22. Pilot Study: Automotive Master Technician Round-to-Round Cut Score Changes by Panel

Panel	Round	Increased N (%)	No Change N (%)	Decreased N (%)
MAA	R1–R2	2 (40.0%)	1 (20.0%)	2 (40.0%)
	R2–R3	2 (40.0%)	1 (20.0%)	2 (40.0%)
MAB	R1–R2	3 (60.0%)	0 (0.0%)	2 (40.0%)
	R2–R3	1 (20.0%)	2 (40.0%)	2 (40.0%)
RAA	R1–R2	1 (25.0%)	2 (50.0%)	1 (25.0%)
	R2–R3	0 (0.0%)	1 (25.0%)	3 (75.0%)
RAB	R1–R2	4 (80.0%)	0 (0.0%)	1 (20.0%)
	R2–R3	0 (0.0%)	3 (60.0%)	2 (40.0%)

Table 23. Pilot Study: Automotive Master Technician Comparison of Cut Scores Based on Medians and Means

Panel	Round 1			Round 2			Round 3		
	Median	Mean	Median–Mean	Median	Mean	Median–Mean	Median	Mean	Median–Mean
MAA	398.0	392.4	**5.6**	395.0	393.2	**1.8**	390.0	392.8	**-2.8**
MAB	394.0	395.4	**-1.4**	405.0	410.4	**-5.4**	405.0	406.4	**-1.4**
RAA	546.5	544.3	**2.3**	547.0	544.5	**2.5**	528.0	524.3	**3.8**
RAB	539.0	532.4	**6.6**	543.0	545.0	**-2.0**	543.0	543.4	**-0.4**

Pilot Study Exemplar Item Ratings. Following the rounds of bookmarking, panelists were asked to identify items that, if responded to correctly (i.e., MC items) or if responses earned a specified number of points (i.e., CR items), would exemplify preparedness for entry-level courses in the program of study (i.e., college or Automotive Master Technician). Exemplar item ratings were collected through the CAB.

Potential exemplar items were drawn from blocks of the assessment that were released to the public. Panelists rated items and score points for which the probability was 0.67 or lower that a student who was prepared (as determined by the round 3 panel cut score) could correctly answer the item or attain the score point. For each item, panelists were asked to indicate if they felt the

item was very good, OK, or should not be used to illustrate preparedness. Table 24 presents a summary of the number of items presented to each panel and the number of items for which panelists expressed 100% agreement and at least 75% agreement that the item/score point would be at least OK to demonstrate preparedness. The average scale value for items selected as at least OK appears parenthetically within the table. Panelists did not reach a high level of agreement on most items.

Table 24. Pilot Study: Exemplar Item Summary

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)
MCA	6	15	205	1 (238)	3 (225)
MCB	6	19	196	5 (231)	8 (223)
RCA	6	12	313	5 (324)	7 (323)
RCB	6	6	331	2 (366)	6 (368)
MAA	5	21	189	1 (216)	7 (209)
MAB	5	23	183	0 (N/A)	2 (211)
RAA	4	6	327	5 (368)	5 (368)
RAB	5	8	321	5 (365)	6 (357)

Pilot Study Process Evaluation Results

The validity of standard-setting outcomes depends in part on what is called *procedural validity*. Procedural validity is provided in the form of evidence that the procedures were carried out as intended and were understood by the panelists. At the end of each round and each day, panelists were provided with an evaluation form designed to assess their understanding of instructions, tasks, and materials. There were a total of five questionnaires administered over the course of each meeting. Most responses were collected using Likert scales, but several responses were narratives that addressed specific aspects of the process. These evaluations were reviewed at the end of each day, and any sources of confusion were identified for clarification with individual

panelists or the group as a whole. The process evaluation questionnaires are presented in their entirety in Appendix M. Along with the questions, the appendix shows the frequency of responses per Likert-scale category, and the average response.

Selected results are presented in the body of this section and include summaries of the following topics:

- Understanding of Tasks;
- Understanding of the Borderline Performance Description;
- Comfort and Confidence;
- Independence of Judgment; and
- Helpfulness of Software.

Results are presented independently for each postsecondary activity. Following the presentation of the process evaluation results, reactions to the consequences data are summarized for each postsecondary activity.

College-Preparedness Workshop Process Evaluation Results.

Understanding of Tasks. Across all panels, when panelists were asked to respond to the following statement: *"My understanding of the tasks I was to accomplish during each round was . . . ,"* the average response was 4.0 or higher, which corresponds to a verbal description of "Adequate" or "Totally Adequate." Average results by panel can be found in Table 25.

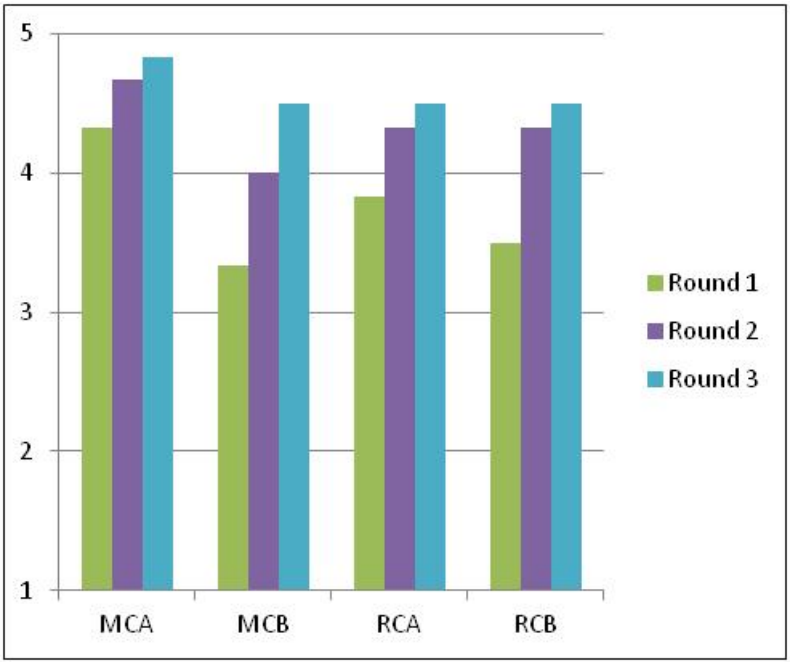
Table 25. Pilot Study: Summary of Selected Evaluation Items by Panel

Evaluation Item	Average Response by Panel							
	College-Preparedness				Automotive Master Technician			
	Mathematics		*Reading*		*Mathematics*		*Reading*	
	MCA	**MCB**	**RCA**	**RCB**	**MAA**	**MAB**	**RAA**	**RAB**
My understanding of the tasks I was to accomplish during each round was . . . (1 = Totally Inadequate; 5 = Totally Adequate)	4.7	4.0	4.0	4.5	4.3	4.2	4.3	3.8
The most accurate description of my level of confidence in the cut score recommendations I provided was . . . (1 = Not at All Confident; 5 = Very Confident)	3.7	4.2	4.8	4.3	4.4	3.8	4.5	4.0
I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark. (1 = Totally Disagree; 5 = Totally Agree)	4.5	3.3	3.5	3.3	4.0	3.8	4.3	3.6
The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items. (1 = Totally Disagree; 5 = Totally Agree)	2.8	2.7	3.7	3.2	2.8	3.6	3.5	2.8
I felt pressured by others in my group to make my cut score recommendation agree with theirs. (1 = Totally Disagree; 5 = Totally Agree)	1.2	1.3	1.2	1.0	1.0	1.6	1.0	1.4
During the standard setting process, I found using CAB to be . . . (1 = Not at All Helpful; 5 = Very Helpful)	4.7	4.0	4.3	4.8	4.6	4.4	4.3	4.0

Understanding of the Borderline Performance Description. After each bookmarking round, panelists were asked to respond to the following statement: *"At the time I placed my bookmark, my understanding of the BPD was . . ."* For round 1, the average response was 3.3 or higher, which corresponds to a verbal description of at least "Somewhat Adequate." Across all panels, the average reported understanding of the BPD steadily increased between rounds, with the lowest average being 4.5 after the third round of bookmarking. Average results by panel are shown in Figure 12.

Figure 12. Pilot Study College-Preparedness: At the time I placed my bookmark, my understanding of the BPD was . . .

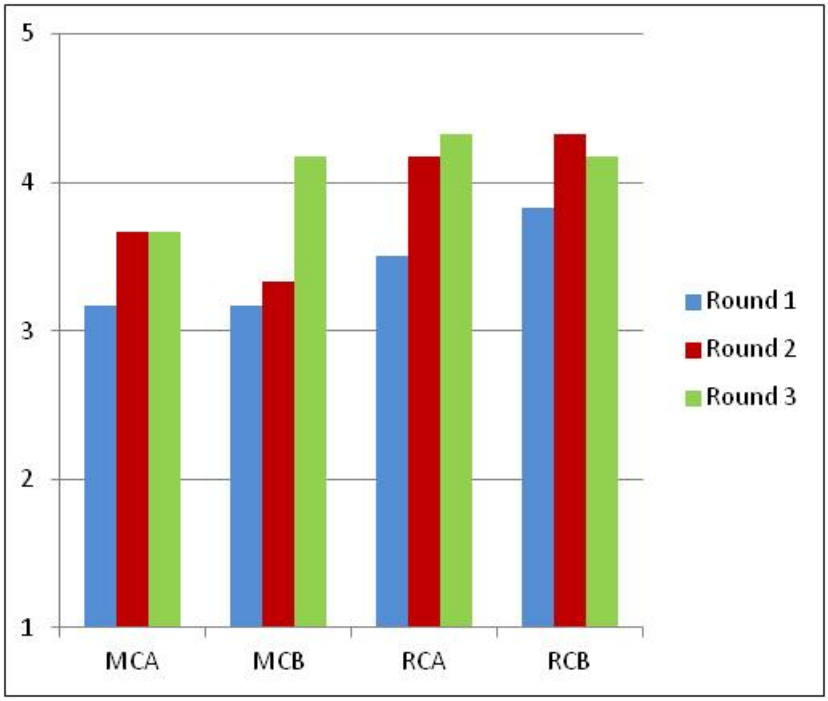
(1 = Totally Inadequate; 5 = Totally Adequate)



Comfort and Confidence. Several statements were developed to solicit how comfortable and/or confident panelists were in setting their cut scores or with components of the standard-setting process. After each bookmarking round, panelists were asked to respond to the following statement: "*The most accurate description of my level of confidence in my bookmark placement is . . .*" For round 1, the average response was 3.2 or higher, which corresponds to a verbal description of at least "Somewhat Confident." Across three of the four panels, the average reported comfort level steadily increased or was maintained between rounds. Reading Panel B did not follow this same pattern, but confidence after the third round of bookmarking was still evident. Average results by panel are shown in Figure 13.

Figure 13. Pilot Study College-Preparedness: The most accurate description of my level of confidence in my bookmark placement is . . .

(1 = Not at All Confident; 5 = Very Confident)



As a follow-up to this statement, panelists were asked to respond to the following statement: *"The most accurate description of my level of confidence in the cut score recommendations I provided was . . ."* Across all panels, the average response was at least 3.7, which corresponds to a verbal description of at least "Somewhat Confident." Average results by panel can be found in Table 25 (p. 91).

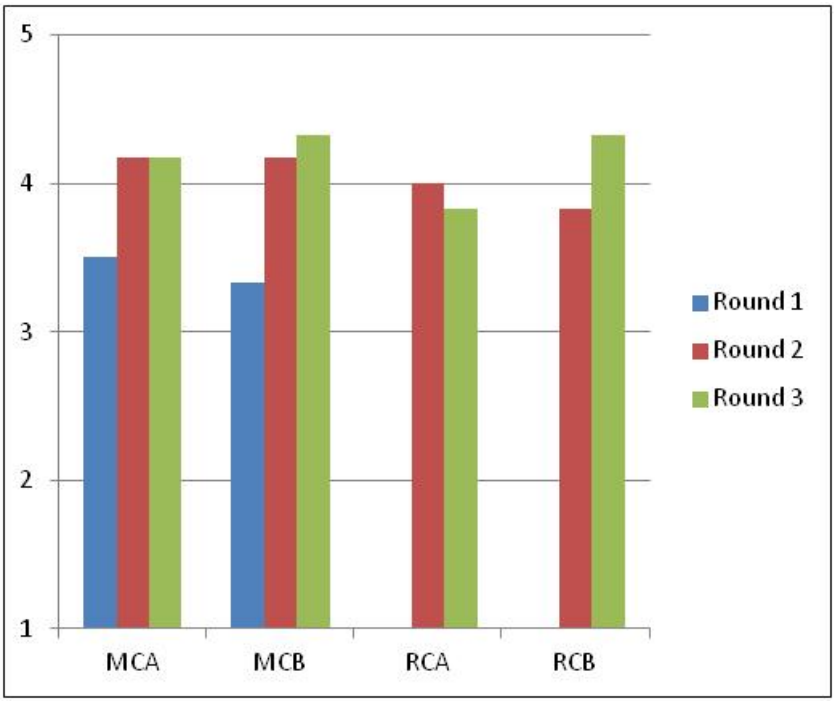
Additionally, after each round of bookmarking, panelists were asked to respond to the following statement: *"I believe my cut score is consistent with the BPD."* It is important to note that round 1 data were not analyzed for all panels.[12] After round 3, all panels had an average response of at least 3.8, which corresponds to a verbal description of at least "Somewhat Agree." Further,

[12] Although data were collected from all of the panels, a technical issue within the CAB did not allow for a clean output of all evaluation data for analysis. This issue was resolved, and all subsequent data outputs were complete.

agreement increased or remained the same across rounds of bookmarking for three of the four panels (MCA, MCB, RCB). Average results by panel are shown in Figure 14.

Figure 14. Pilot Study College-Preparedness: I believe my cut score is consistent with the BPD

(1 = Totally Disagree; 5 = Totally Agree)



Specific to the standard-setting method are the reliance on an understanding of how to use a response probability in placing a bookmark and a general acceptance that items are ordered by relative difficulty. Thus, each of these assumptions was evaluated. Panelists were asked to respond to the following statement:*"I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark."* Across all panels, the average response was at least 3.3, which corresponds to a verbal description of at least "Somewhat Agree." Average results by panel can be found in Table 25 (p. 91).

Panelists were also asked to respond to the following statement: *"The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items."* Average responses

across panels varied from 2.7 to 3.7, which corresponds to verbal descriptions between "Disagree" and "Agree." Average results by panel can be found in Table 25 (p. 91).

Independence of Judgment. Across all panels, when panelists were asked to respond to the following statement: *"I felt pressured by others in my group to make my cut score recommendation agree with theirs,"* the average response was no higher than 1.3, which corresponds to a verbal description of "Disagree" or "Totally Disagree." Average results by panel can be found in Table 25 (p. 91).

Helpfulness of Software. As the CAB software system was implemented to aid the standard-setting process, panelists were asked to respond to the following statement: *"During the standard setting process, I found using CAB to be . . ."* Across all panels, the lowest average response was 4.0, which corresponds to a verbal description of "Helpful" or higher. Average results by panel can be found in Table 25 (p. 91).

Panelists were asked to review both their final round 3 cut scores and the associated impact data. They were asked to respond to the following questions with a "Yes" or "No" response:

- *Does the [impact data] percentage reflect your expectations about the proportion of students whose NAEP score would indicate at least minimal preparedness for placement?*
- *Would you change the cut score recommended by your panel to the Governing Board if you could?*

Results across panels varied significantly, with 17% to 83% of panelists (depending on which panel they belonged to) responding "Yes" to the first question and 50% to 80% of panelists responding "No" to the second question. The full set of results can be found in Appendix M.

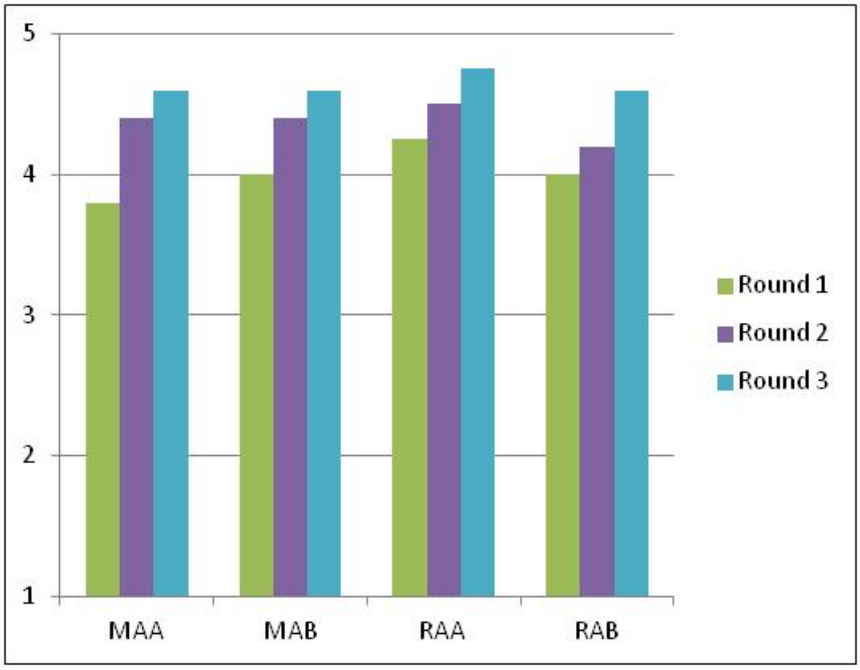
Automotive Master Technician Workshop Process Evaluation Results.

Understanding of Tasks. Across all panels, when panelists were asked to respond to the following statement: *"My understanding of the tasks I was to accomplish during each round was . . ."* the average response was 3.8 or higher, which corresponds to a verbal description of at least "Somewhat Adequate." Average results by panel can be found in Table 25 (p. 91).

Understanding of the Borderline Performance Description. After each bookmarking round, panelists were asked to respond to the following statement: *"At the time I placed my bookmark, my understanding of the BPD was . . ."* For round 1, the average response was 3.8 or higher, which corresponds to a verbal description of at least "Somewhat Adequate." Across all panels, the average reported understanding of the BPD steadily increased between rounds, with the lowest average being 4.6 after the third round of bookmarking. Average results by panel are shown in Figure 15.

Figure 15. Pilot Study Automotive Master Technician: At the time I placed my bookmark, my understanding of the BPD was . . .

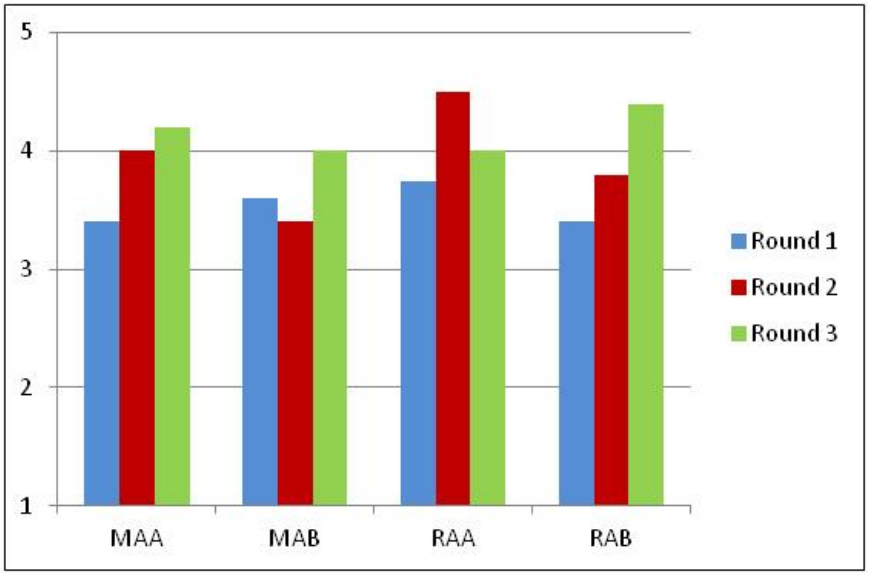
(1 = Totally Inadequate; 5 = Totally Adequate)



Comfort and Confidence. Several statements were developed to solicit how comfortable and/or confident panelists were in setting their cut scores or with components of the standard-setting process. After each bookmarking round, panelists were asked to respond to the following statement: "*The most accurate description of my level of confidence in my bookmark placement is . . .*" For round 1, the average response was 3.4 or higher, which corresponds to a verbal description of at least "Somewhat Confident." Across most panels (MAA, MAB, RAB), the average reported comfort level was higher after the third round of bookmarking, although no clear patterns of increasing confidence were evident between each round and across panels. Average results by panel are shown in Figure 16.

Figure 16. Pilot Study Automotive Master Technician: The most accurate description of my level of confidence in my bookmark placement is . . .

(1 = Not at All Confident; 5 = Very Confident)



As a follow-up to this statement, panelists were asked to respond to the following statement: *"The most accurate description of my level of confidence in the cut score recommendations I provided was . . ."* Across all panels, the average response was at least 3.8, which corresponds to a verbal description of at least "Somewhat Confident." Average results by panel can be found in Table 25 (p. 91).

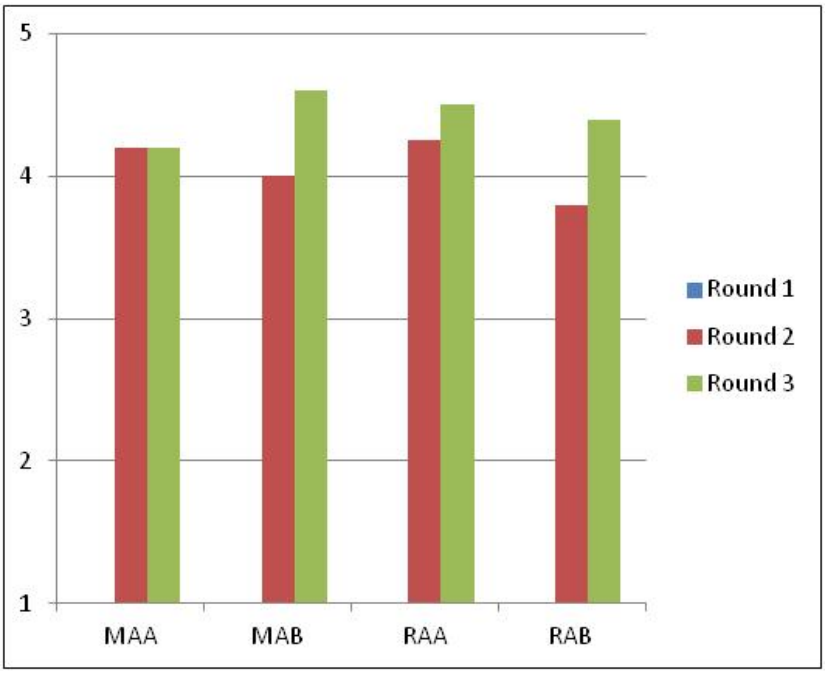
Additionally, after each round of bookmarking, panelists were asked to respond to the following statement: *"I believe my cut score is consistent with the BPD."* It is important to note that round 1 data could not be analyzed for all panels.[13] After round 2, all panels had an average response of at least 3.8, which corresponds to a verbal description of at least "Somewhat Agree."

[13] Although data were collected from all of the panels, a technical issue within the CAB did not allow for a clean output of all evaluation data for analysis. This issue was resolved, and all subsequent data outputs were complete.

Furthermore, agreement increased or remained the same between rounds 2 and 3 of bookmarking for all panels. Average results by panel are shown in Figure 17.

Figure 17. Pilot Study Automotive Master Technician: I believe my cut score is consistent with the BPD

(1 = Totally Disagree; 5 = Totally Agree)



Specific to the standard-setting method are the reliance on an understanding of how to use a response probability in placing a bookmark and a general acceptance that items are ordered by relative difficulty. Thus, each of these assumptions was evaluated. Panelists were asked to respond to the following statement: *"I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark."* Across all panels, the average response was at least 3.6, which corresponds to a verbal description of at least "Somewhat Agree." Average results by panel can be found in Table 25 (p. 91).

Panelists were also asked to respond to the following statement: *"The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items."* Average responses across panels varied from 2.8 to 3.6, which corresponds to verbal descriptions between "Disagree" and "Agree." Average results by panel can be found in Table 25 (p. 91).

Independence of Judgment. Across all panels, when panelists were asked to respond to the following statement: *"I felt pressured by others in my group to make my cut score recommendation agree with theirs,"* the average response was no higher than 1.6, which corresponds to a verbal description of "Disagree" or "Totally Disagree." Average results by panel can be found in Table 25 (p. 91).

Helpfulness of Software. As the CAB software system was implemented to aid the standard-setting process, panelists were asked to respond to the following statement: *"During the standard setting process, I found using CAB to be . . ."* Across all panels, the lowest average response was 4.0, which corresponds to a verbal description of "Helpful" or higher. Average results by panel can be found in Table 25 (p. 91).

Panelists were asked to review both their final round 3 cut scores and the associated impact data. They were asked to respond to the following questions with a "Yes" or "No" response:

- *Does the [impact data] percentage reflect your expectations about the proportion of students whose NAEP score would indicate at least minimal preparedness for placement?*
- *Would you change the cut score recommended by your panel to the Governing Board if you could?*

Results across panels varied significantly, with 60% to 100% of panelists (depending on which panel they belonged to) responding “Yes” to the first question and 0% to 100% of panelists responding “No” to the second question.

Overall, results from the process evaluations provide evidence that panelists understood their tasks and were comfortable with the process, although they were not necessarily satisfied with the results. The full set of pilot study process-evaluation results can be found in Appendix M.

Operational Session 1

The first operational JSS session was conducted on May 24–27, 2011, and included college-preparedness and Automotive Master Technician workshops.

For this first operational session, the JSS process was largely implemented as described in the Judgmental Standard-Setting Process subsection of this report (pp. 39–69). Based on feedback from pilot study panelists, facilitators, and observers, however, the following enhancements were implemented in this session.

A general session was held at the beginning of each day and prior to every major stage of the process, including introductions to new functionalities of the CAB. These sessions provided the opportunity to recap the previous day's activities, address issues identified from panelists' responses to process evaluation questionnaires, and give uniform instructions to panelists for the current day's tasks. Instructions for specific tasks were repeated by the process facilitators in the panel rooms. PowerPoint slides with instructions for different tasks were provided to each process facilitator.

Given that college-preparedness and Automotive Master Technician workshops were included in the pilot study, the pilot study college-preparedness and Automotive Master Technician BPDs were used as the starting point for developing the operational BPDs. The content facilitators consolidated the pilot study BPDs and the Operational Session 1 panelists' responses to the online Content Objectives form to develop preliminary BPD drafts to share with the Operational Session 1 panelists, highlighting for discussion KSAs that were in the pilot study BPDs but not selected by the operational panelists as required for job-training program preparedness as well as those KSAs selected as required by the operational panelists but not present in the pilot study

BPDs. The BPD session on the first day of the standard-setting workshops addressed these highlighted KSAs. Based on feedback from the pilot study, knowledge of how achievement standard-setting procedures are typically implemented, and JSS-TAC recommendations, within the operational workshops BPDs were finalized earlier in the JSS process than had occurred in the pilot study, in order to have a stable standard throughout the bookmarking process. Panelists were instructed to finalize the BPDs prior to the first round of item rating and to review and discuss the BPDs prior to subsequent rounds; if they identified an error or desired modification, such edits were made with the approval of the content facilitators. The preliminary mathematics and reading BPDs that were developed by content facilitators prior to the JSS session and the final BPD versions agreed upon by the mathematics and reading pairs of replicate panels for each postsecondary area are provided in Appendix O.

The KSA review of items was enhanced by having the content facilitators introduce the topic to content area groups from each postsecondary activity. Using items from the released blocks of NAEP items, the content facilitators led the panelists in discussions of what students should know and be able to do in order to receive credit for their responses to CR items, modeling the thought process of describing the KSA required to correctly respond to each item. The content facilitators had been given examples of KSA descriptions, and they encouraged panelists to suggest ways to describe the KSAs needed for each credited response to CR items. This activity preceded the part of the process in which panelists reviewed KSAs for MC items, so only CR items were used in this discussion. Following this training, panelists developed and discussed KSAs for four assigned CR items in their panel groups. During this activity, process facilitators had access to descriptions for all credited responses for each item and were instructed to use these descriptions to help guide panelists in meaningful discussions.

Based on review of KSAs developed by pilot study panelists, panelists were given instructions on what *not* to do when identifying the KSAs required to respond to an MC item correctly or to receive a certain score level for CR items. Instructions to panelists included the following:

- Do NOT write your opinion of the item or rubric.
- Do NOT write about how the KSA relates to the BPD.
- Do NOT rate the items as “hard” or “easy.”
- Do NOT rate the items as “needed/relevant” or “not needed/irrelevant” for your particular postsecondary area.

Furthermore, for items that measure KSAs that are not relevant to the requirements of panelists’ programs or courses, panelists were instructed to make note of such items but to not rate all items in terms of relevance.

For the bookmarking rounds, panelists were given the following instruction on how to place bookmarks relative to items deemed irrelevant to their programs or courses: if the first item that a panelist finds too difficult for students performing at the borderline of academic preparedness is preceded by an item or items that the panelist deems irrelevant for his/her program, then the panelist should place her/his bookmark immediately after the last item relevant to the program that is not too difficult for a student performing at the borderline. This instruction was provided in conjunction with other considerations, such as the range of uncertainty, when deciding where to place the bookmark.

Another enhancement between the pilot study and this session was a reordering of items in the reading CROIBs to reflect the sequence in which they were presented to students within blocks. Within the NAEP reading assessment, reading items are passage-based, with one primary

passage accompanying a block of items. Prior to this modification, CR items were ordered in the CROIB based on the scale value of the highest score level, with all of the score points for each CR item presented together. However, adjacent CR items might be associated with two different passages. Within the modified CROIB, all items associated with a passage were presented with that passage. Blocks were ordered by difficulty, with the easiest blocks appearing first; block difficulty was based on the average difficulty of the CR items within the block. This ordering of items resulted in each passage appearing only once in printed materials, which made the printed OIBs and CROIBs less cumbersome. This modification also meant that, during the KSA review of CR items, panelists read each passage once to review all CR items for that block. The ordering of items remained consistent between printed OIBs and CROIBs and those presented in the CAB.

Finally, during this workshop, scoring guides accompanying whole booklets were provided to panelists. Providing scoring guides for whole-booklet review has not been customary in NAEP standard setting. The whole booklet is meant to be a holistic feedback; thus, scoring individual responses was not encouraged. However, given that some panelists lacked familiarity with NAEP assessment content, they had a difficult time reviewing performance holistically; therefore, giving them correct responses to the MC items and scoring rubrics for the CR items in the whole-booklet forms eliminated their need to refer to the OIBs for this information, an additional step that pilot study panelists found cumbersome.

The abbreviations used to describe panels in the first operational session are displayed in Table 26.

Table 26. Operational Session 1: Abbreviations Used to Describe Panels

Abbreviation	Content Area	Postsecondary Area	Panel
MCA	Mathematics	College-Preparedness	A
MCB			B
RCA	Reading	College-Preparedness	A
RCB			B
MAA	Mathematics	Automotive Master Technician	A
MAB			B
RAA	Reading	Automotive Master Technician	A
RAB			B

College-Preparedness Operational Workshop

The following subsections describe the college-preparedness operational workshop panelists, numerical results, and process evaluation results.

College-Preparedness Operational Workshop Panelists

College-preparedness operational panelists were recruited from 30 postsecondary institutions and eight secondary institutions. The college-preparedness sampling plan stipulated representation from both public (75%) and private (25%) postsecondary institutions. As displayed in Table 27, the majority of postsecondary college-preparedness operational panelists (97%) were from public institutions, with only one mathematics panelist representing a private institution; this corresponds with the distribution of pilot study college-preparedness panelists. The selected institutions closely reflected the targeted representation for highly selective institutions (3%), although open-enrollment institutions were underrepresented on the panels (43% versus the target of 63%), and moderately selective institutions were overrepresented on the panels (53% versus the target of 34%). The percentage of medium-sized institutions (47%) came close to matching the target of 42%; however, fewer small institutions were selected than planned (13% versus the target of 24%), and more large institutions were selected than planned (40% versus the target of 34%). Relative to the pilot study panelists, somewhat more moderately selective institutions were represented in the operational workshop, while fewer small institutions were represented in the operational workshop.

Table 27. College-Preparedness Operational Workshop: Postsecondary Institution Distributions

Post-secondary Area	Content Area	Panel	Private/Public Status		Selectivity			Size			Total Different Institutions
			Public	Private	Open	Mod.	High	Small	Med.	Large	
College-Prepared-ness	Math	A	8 (100%)	0 (0%)	5 (63%)	3 (38%)	0 (0%)	0 (0%)	5 (63%)	3 (37%)	8 (100%)
		B	7 (88%)	1 (13%)	4 (50%)	4 (50%)	0 (0%)	2 (25%)	3 (38%)	3 (38%)	8 (100%)
	Reading †	A	8 (100%)	0 (0%)	1 (13%)	6 (75%)	1 (13%)	2 (25%)	3 (38%)	3 (38%)	8 (100%)
		B*	6 (100%)	0 (0%)	3 (50%)	3 (50%)	0 (0%)	0 (0%)	3 (50%)	3 (50%)	6 (100%)
College-Preparedness Institution Totals (N = 30)			29 (97%)	1 (3%)	13 (43%)	16 (53%)	1 (3%)	4 (13%)	14 (47%)	12 (40%)	30 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of institution distribution across replicate panels.

†Two reading panelists came from the same university system but from different campuses with different demographics; therefore, these campuses are treated as different institutions.

*One reading Panel B panelist represented the same institution as a reading Panel A panelist; that institution is included in counts for Panel A but not in counts for Panel B.

The college-preparedness sampling target for secondary-level institutions called for a distribution across the categories of urban (34%), suburban (41%), and town/rural (25%). The institutions represented on these panels, however, were evenly distributed across urban (50%) and suburban (50%) locations, with no town/rural institutions represented, as shown in Table 28. The target for institutional size for the panel was not fully met; while approximately the targeted percentage of small institutions was recruited (13% versus the target of 14%), medium-sized institutions were underrepresented (13% versus the target of 22%) and large institutions were overrepresented (75% versus the target of 64%).This distribution of institutions is similar to what was represented in the pilot study.

Table 28. College-Preparedness Operational Workshop: Secondary-Level Institution Distributions

Postsecondary Area	Content Area	Panel	Urbanicity			Size			
			Urban	Sub-urban	Town/ Rural	Small	Med.	Large	Total Different Institutions
College-Preparedness	Math	A	0 (0%)	2 (100%)	0 (0%)	0 (0%)	0 (0%)	2 (100%)	2 (100%)
		B	2 (100%)	0 (0%)	0 (0%)	1 (50%)	1 (50%)	0 (0%)	2 (100%)
	Reading	A	2 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (100%)	2 (100%)
		B	0 (0%)	2 (100%)	0 (0%)	0 (0%)	0 (0%)	2 (100%)	2 (100%)
College-Preparedness Institution Totals (N = 8)			4 (50%)	4 (50%)	0 (0%)	1 (13%)	1 (13%)	6 (75%)	8 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of institution distribution across replicate panels.

Thirty-nine panelists were recruited to participate in the college-preparedness operational session. Table 29 displays the distribution of panelists by type of institution, and Table 30 displays the distribution of panelists by demographic characteristics. The Design Document called for a total of 40 panelists for this workshop: 20 in mathematics and 20 in reading, with each group divided into two replicate panels of 10 panelists. This target was nearly achieved, but one reading panelist dropped from the study too late to be replaced.

Eight of the 15 college-preparedness postsecondary reading instructors were recruited from traditional English- or composition-type programs (e.g., literature, composition, English, developmental reading), while the remaining seven panelists were instructors in other departments for which the curriculum involves a heavy reading load for students: communications, education, history, sociology, and business courses in two- and four-year colleges. Across all panels, over half of postsecondary panelists came from four-year institutions, a higher proportion than was seen in the pilot study. In addition, all college-preparedness replicate panels included two secondary-level teachers of 12th graders. As shown in Table 30,

some gender diversity was achieved on each college-preparedness panel, although women consistently outnumbered men. Panelists were predominantly White/Caucasian, with limited ethnic diversity on all four panels. This distribution is comparable to what was obtained in the pilot study.

Table 29. College-Preparedness Operational Workshop: Panelist Distribution by Institution Type

Post-secondary Area	Content Area	Panel	Type of Institution					Total Panelists
			4-Year Public	4-Year Private	2-Year Public (Community/ Technical)	2-Year Private	Secondary	
College-Preparedness	Math	A	5 (50%)	0 (0%)	3 (30%)	0 (0%)	2 (20%)	10 (100%)
		B	5 (50%)	1 (10%)	2 (20%)	0 (0%)	2 (20%)	10 (100%)
	Reading	A	3 reading (30%) 4 “other” (40%)	0 (0%)	1 reading (10%)	0 (0%)	2 (20%)	10 (100%)
		B	2 reading (22%) 2 “other” (22%)	0 (0%)	2 reading (22%) 1 “other” (11%)	0 (0%)	2 (22%)	9 (100%)
College-Preparedness Totals (N = 39)			21 (54%)	1 (3%)	9 (23%)	0 (0%)	8 (21%)	39 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

Table 30. College-Preparedness Operational Workshop: Panelist Distribution by Demographic Characteristics

Postsecondary Area	Content Area	Panel	Gender		Race/Ethnicity			Total Panelists
			Female	Male	White/ Caucasian	Non-White/ Caucasian	Not Specified	
College-Preparedness	Math	A	6 (60%)	4 (40%)	9 (90%)	1 (10%)	0 (0%)	10 (100%)
		B	6 (60%)	4 (40%)	8 (80%)	2 (20%)	0 (0%)	10 (100%)
	Reading	A	6 (60%)	4 (40%)	8 (80%)	2 (20%)	0 (0%)	10 (100%)
		B	5 (56%)	4 (44%)	7 (78%)	2 (22%)	0 (0%)	9 (100%)
College-Preparedness Totals (N = 39)			23 (59%)	16 (41%)	32 (82%)	7 (18%)	0 (0%)	39 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

As shown in Table 31, all four census regions were represented on the college-preparedness operational panels, with the lowest percentage of panelists (13%) representing the Northeast and the highest percentage (38%) representing the South. In three of the four college-preparedness replicate panels, panelists represented all four geographic regions. In reading Panel B, however, there were no panelists from the Northeast.

Table 31. College-Preparedness Operational Workshop: Geographic Distribution of Panelists

Postsecondary Area	Content Area	Panel	Geographic Region*				Total Panelists
			Northeast	South	Midwest	West	
College-Preparedness	Math	A	1 (10%)	5 (50%)	2 (20%)	2 (20%)	10 (100%)
		B	2 (20%)	5 (50%)	2 (20%)	1 (10%)	10 (100%)
	Reading	A	2 (20%)	2 (20%)	3 (30%)	3 (30%)	10 (100%)
		B	0 (0%)	3 (33%)	2 (22%)	4 (44%)	9 (100%)
College-Preparedness Totals (N = 39)			5 (13%)	15 (38%)	9 (23%)	10 (26%)	39 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.
*Based upon U.S. Census Bureau census regions.

College-Preparedness Operational Workshop Numerical Results

The following tables display standard-setting results for the college-preparedness operational workshop.

When round 1 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The round 1 results are presented in Table 32.

Table 32. College-Preparedness Operational Workshop: Round 1 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	196	8.9	9.3	14.9	10.8
	B	184	10.1	9.8	16.8	19.6
Reading	A	273	10.3	9.7	23.9	68.7
	B	303	12.9	13.0	23.7	37.6

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 2 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 2 judgments are presented in Table 33.

Table 33. College-Preparedness Operational Workshop: Round 2 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	201	1.8	2.4	4.6	7.9
	B	191	2.8	3.1	4.9	14.0
Reading	A	288	5.3	5.4	10.2	36.4
	B	304	1.5	1.6	3.8	36.4

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 3 bookmarks were placed, the CAB once again calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 3 decisions are presented in Table 34.

Table 34. College-Preparedness Operational Workshop: Round 3 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	201	1.4	1.6	3.1	7.9
	B	189	2.5	2.6	4.8	15.5
Reading	A	290	2.2	2.1	5.9	51.6
	B	304	1.3	1.4	3.6	36.4

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.

[b] Empirical-based Standard Error.

[c] Bootstrap Standard Error.

[d] Mean Absolute Deviation.

[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

Figure 18 summarizes the MAD results across rounds for the college-preparedness operational panels. In contrast to results from the pilot study, variation among panelists' cut scores decreased between rounds for all panels.

Figure 18. College-Preparedness Operational Workshop: Mean Absolute Deviation (MAD) of Cut Scores by Panel

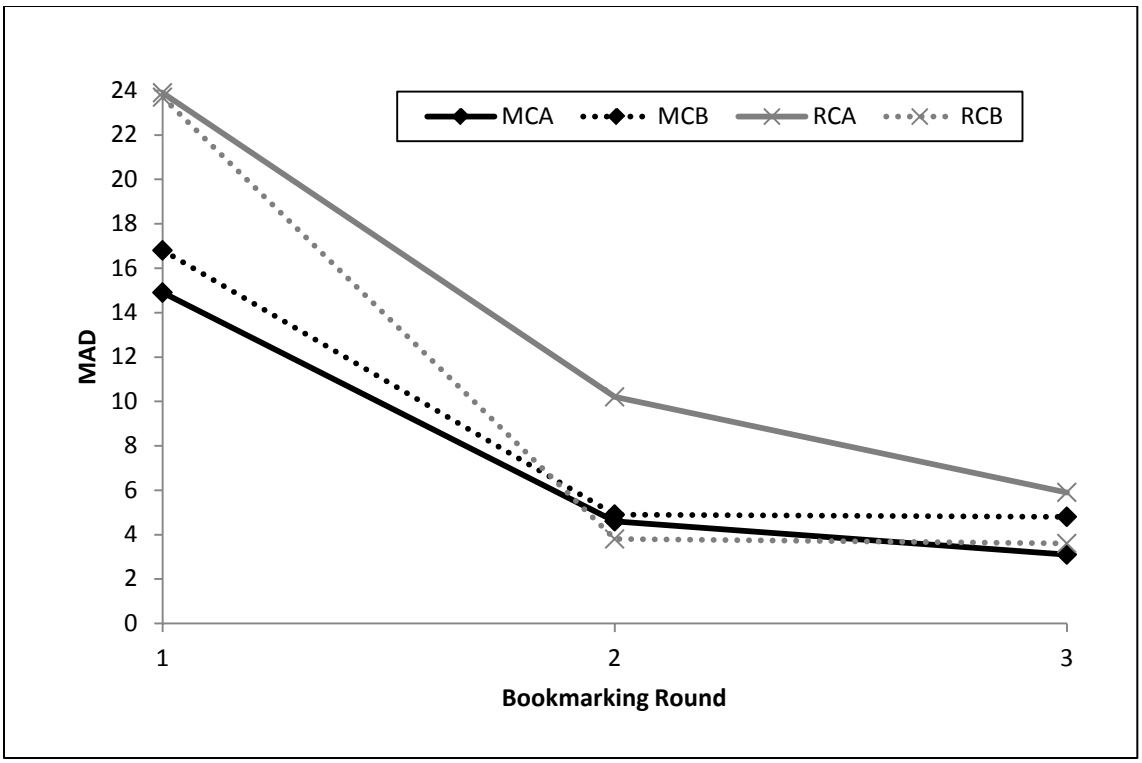


Table 35 summarizes the changes in cut scores made by individual panelists between rounds, while Table 36 presents a comparison of median and mean cut scores for each panel. The expectation was that the same number of panelists or fewer panelists changed their cut scores between rounds 2 and 3 than between rounds 1 and 2. This pattern is observed in Table 35. Differences between mean and median cut scores are due to the presence of outliers (i.e., panelists who set cut scores significantly higher or lower than their peers). The largest absolute difference between the mean and median cut scores in Table 36 occurs for reading Panel A in round 1 (8.6), where the direction of the difference would result in a higher cut score being set if the mean instead of the median were used. Overall, in the third round, the effect of outliers on mean cut scores is negligible.

Table 35. College-Preparedness Operational Workshop: Round-to-Round Cut Score Changes by Panel

Panel	Round	Increased N (%)	No Change N (%)	Decreased N (%)
MCA	R1–R2	5 (50.0%)	1 (10.0%)	4 (40.0%)
	R2–R3	6 (60.0%)	2 (20.0%)	2 (20.0%)
MCB	R1–R2	5 (50.0%)	1 (10.0%)	4 (40.0%)
	R2–R3	2 (20.0%)	5 (50.0%)	3 (30.0%)
RCA	R1–R2	8 (80.0%)	0 (0.0%)	2 (20.0%)
	R2–R3	5 (50.0%)	4 (40.0%)	1 (10.0%)
RCB	R1–R2	4 (44.4%)	0 (0.0%)	5 (55.6%)
	R2–R3	0 (0.0%)	8 (88.9%)	1 (11.1%)

Table 36. College-Preparedness Operational Workshop: Comparison of Cut Scores Based on Medians and Means

Panel	Round 1			Round 2			Round 3		
	Median	Mean	Median–Mean	Median	Mean	Median–Mean	Median	Mean	Median–Mean
MCA	196.0	190.7	**5.3**	201.0	197.2	**3.8**	201.0	199.5	**1.5**
MCB	184.0	186.6	**-2.6**	191.0	187.5	**3.5**	188.5	187.4	**1.1**
RCA	272.5	281.1	**-8.6**	288.0	286.8	**1.2**	289.5	290.3	**-0.8**
RCB	303.0	303.1	**-0.1**	304.0	305.3	**-1.3**	304.0	305.1	**-1.1**

College-Preparedness Operational Workshop Exemplar Item Ratings. Following the rounds of bookmarking, panelists were asked to identify items that, if responded to correctly (i.e., MC items) or if responses earned a specified number of points (i.e., CR items), would exemplify preparedness for entry-level college courses.

Table 37 presents a summary of the number of items presented to each panel and the number of items for which panelists expressed 100% agreement and at least 75% agreement that the item/score point would be at least OK to demonstrate preparedness. The average scale value for items selected as at least OK appears parenthetically within the table as well.

Table 37. College-Preparedness Operational Workshop: Exemplar Item Summary

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)
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)

College-Preparedness Operational Workshop Process Evaluation Results

Following this presentation of the process evaluation results, reactions to the consequences data are summarized. The process evaluation questionnaires for this workshop are presented in their entirety in Appendix P; along with the questions, the appendix shows the frequency of responses per Likert-scale category, and the average response.

Understanding of Tasks. Across all panels, when panelists were asked to respond to the following statement: *"My understanding of the tasks I was to accomplish during each round was . . ."* the average response was 4.1 or higher, which corresponds to a verbal description of "Adequate" or "Totally Adequate." Average results by panel can be found in Table 38.

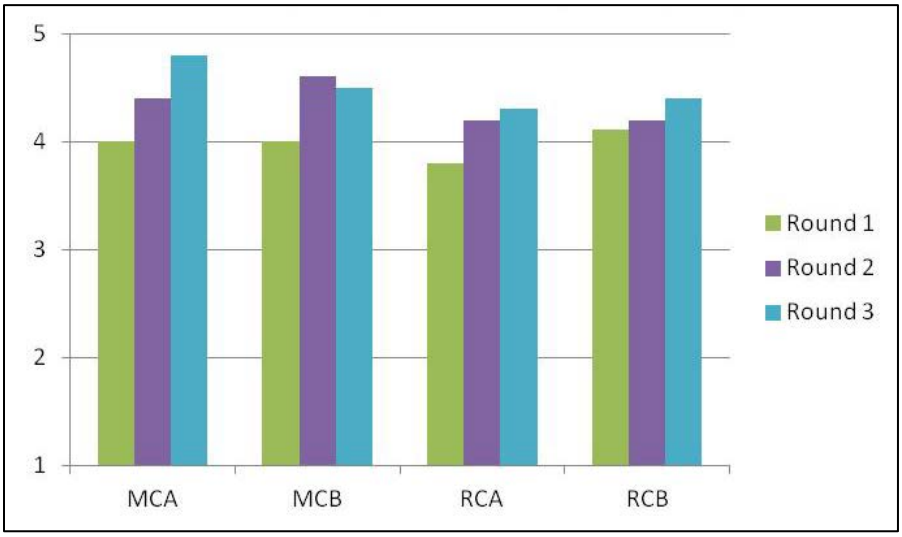
Table 38. College-Preparedness Operational Study: Summary of Selected Evaluation Items by Panel

Evaluation Item	Average Response by Panel			
	College-Preparedness			
	Mathematics		*Reading*	
	MCA	**MCB**	**RCA**	**RCB**
My understanding of the tasks I was to accomplish during each round was . . . (1 = Totally Inadequate; 5 = Totally Adequate)	4.2	4.4	4.1	4.3
The most accurate description of my level of confidence in the cut score recommendations I provided was . . . (1 = Not at All Confident; 5 = Very Confident)	4.4	3.9	4.5	4.7
I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark. (1 = Totally Disagree; 5 = Totally Agree)	3.9	3.8	3.8	3.1
The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items. (1 = Totally Disagree; 5 = Totally Agree)	4.4	3.6	3.0	3.2
I felt pressured by others in my group to make my cut score recommendation agree with theirs. (1 = Totally Disagree; 5 = Totally Agree)	1.5	1.7	1.8	1.8
During the standard setting process, I found using CAB to be . . . (1 = Not at All Helpful; 5 = Very Helpful)	4.7	4.3	4.4	4.9

Understanding of the Borderline Performance Description. After each bookmarking round, panelists were asked to respond to the following statement: *"At the time I placed my bookmark, my understanding of the BPD was . . ."* For round 1, the average response was 3.8 or higher, which corresponds to a verbal description of at least "Somewhat Adequate." Across all panels, the average reported understanding of the BPD increased between rounds 1 and 3, with the lowest average being 4.3 after the third round of bookmarking. Average results by panel are shown in Figure 19.

Figure 19. College-Preparedness Operational Workshop: At the time I placed my bookmark, my understanding of the BPD was . . .

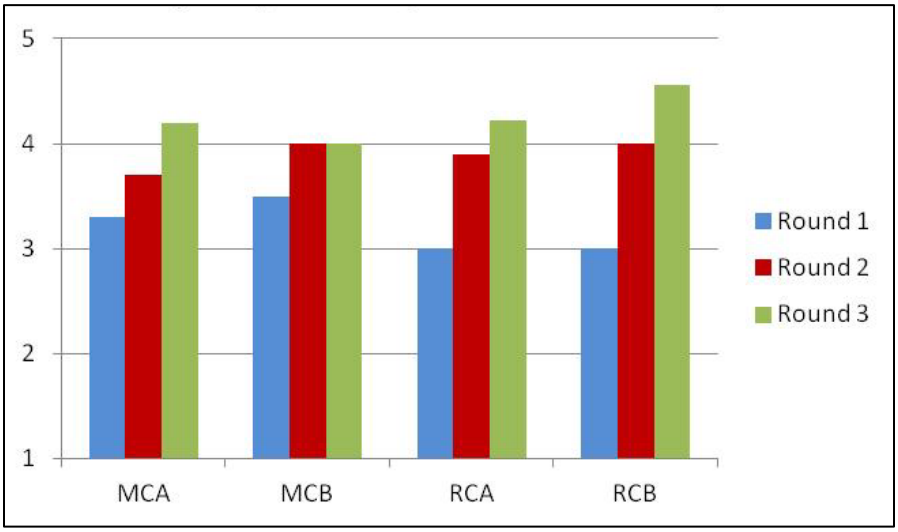
(1 = Totally Inadequate; 5 = Totally Adequate)



Comfort and Confidence. Several statements were developed to solicit how comfortable and/or confident panelists were in setting their cut scores or with components of the standard-setting process. After each bookmarking round, panelists were asked to respond to the following statement: *"The most accurate description of my level of confidence in my bookmark placement is . . ."* For round 1, the average response was 3.0 or higher, which corresponds to a verbal description of at least "Somewhat Confident." Across all panels, the average reported comfort level steadily increased or was maintained between rounds, with the lowest average being 4.0 after the third round of bookmarking. Average results by panel are shown in Figure 20.

Figure 20. College-Preparedness Operational Workshop: The most accurate description of my level of confidence in my bookmark placement is . . .

(1 = Not at All Confident; 5 = Very Confident)

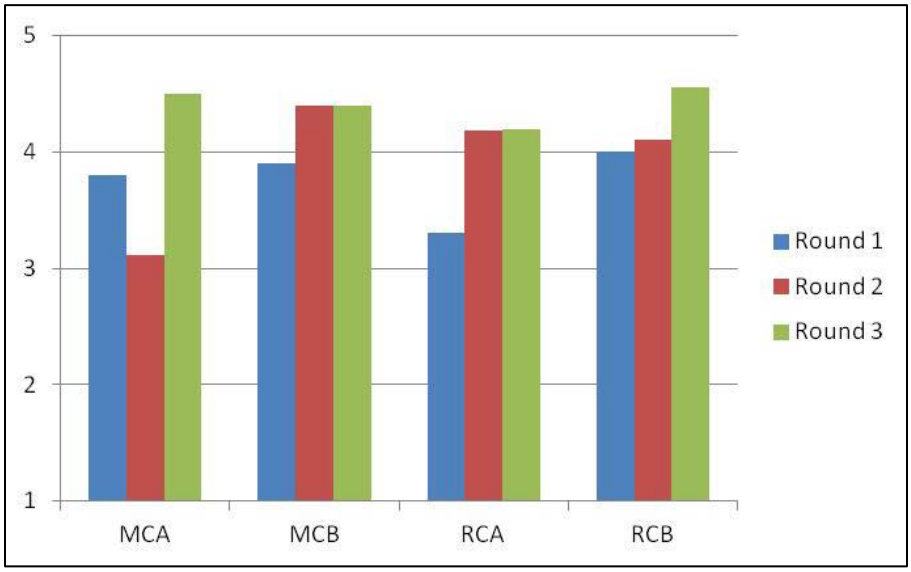


As a follow-up to this statement, panelists were asked to respond to the following statement: *"The most accurate description of my level of confidence in the cut score recommendations I provided was . . ."* Across all panels, the average response was at least 3.9, which corresponds to a verbal description of at least "Somewhat Confident." Average results by panel can be found in Table 38 (p. 117).

Additionally, after each round of bookmarking, panelists were asked to respond to the following statement: *"I believe my cut score is consistent with the BPD."* After round 1, all panels had an average response of at least 3.3, which corresponds to a verbal description of at least "Somewhat Agree." Further, agreement increased or remained the same for all panels by the end of the third bookmarking round. Average results by panel are shown in Figure 21.

Figure 21. College-Preparedness Operational Workshop: I believe my cut score is consistent with the BPD

(1 = Totally Disagree; 5 = Totally Agree)



Specific to the standard-setting method are the reliance on an understanding of how to use a response probability in placing a bookmark and a general acceptance that items are ordered by relative difficulty. Thus, each of these assumptions was evaluated. Panelists were asked to respond to the following statement: *"I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark."* Across all panels, the average response was at least 3.1, which corresponds to a verbal description of at least "Somewhat Agree." Average results by panel can be found in Table 38 (p. 117).

Panelists were also asked to respond to the following statement: *"The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items."* Average responses across panels varied from 3.0 to 4.4, which corresponds to verbal descriptions between "Somewhat Agree" and "Totally Agree." Average results by panel can be found in Table 38 (p. 117).

Independence of Judgment. Across all panels, when panelists were asked to respond to the following statement: *"I felt pressured by others in my group to make my cut score recommendation agree with theirs,"* the average response was no higher than 1.8, which corresponds to a verbal description of "Disagree" or "Totally Disagree." Average results by panel can be found in Table 38 (p. 117).

Helpfulness of Software. As the CAB software system was implemented to aid the standard-setting process, panelists were asked to respond to the following statement: *"During the standard setting process, I found using CAB to be . . ."* Across all panels, the lowest average response was 4.3, which corresponds to a verbal description of "Helpful" or higher. Average results by panel can be found in Table 38 (p. 117).

Panelists were asked to review both their final round 3 cut scores and the associated impact data. They were asked to respond to the following questions with a "Yes" or "No" response:

- *Does the [impact data] percentage reflect your expectations about the proportion of students whose NAEP score would indicate at least minimal preparedness for placement?*
- *Would you change the cut score recommended by your panel to the Governing Board if you could?*

Sixty percent to 83% of panelists (depending on which panel they belonged to) responded "Yes" to the first question, and 55% to 80% of panelists responded "No" to the second question, indicating that most panelists were satisfied with the final cut scores. The full set of results can be found in Appendix P.

Automotive Master Technician Operational Workshop

The following subsections describe the Automotive Master Technician operational workshop panelists, numerical results, and process evaluation results.

Automotive Master Technician Operational Workshop Panelists

Automotive Master Technician operational panelists were recruited from 23 Automotive Master Technician job-training programs. From these programs, 26 Automotive Master Technician panelists were recruited to participate in the first operational session.[14] Table 39 displays the distribution of panelists by type of institution, and Table 40 displays the distribution of panelists by demographic characteristics.

The Design Document called for a total of 40 panelists for this workshop: 20 in mathematics and 20 in reading, with each group divided into two replicate panels of 10 panelists. However, the recruitment challenges previously described resulted in the recruitment of only 26 Automotive Master Technician panelists: 15 for mathematics and 11 for reading. Panelists were assigned to replicate panels and table groups within panels with one exception: all five panelists on reading Panel A were assigned to one table group in order to avoid having a table with fewer than three panelists.

All panelists were recruited from postsecondary Automotive Master Technician institutions, as illustrated in Table 39. While the majority (92%) represented public two-year community or technical programs, one panelist on each of the mathematics replicate panels taught at a

[14]Three panelists came from one program, and two panelists came from another program. Panelists from the same program were assigned to separate replicate panels when possible or, if not possible, to separate table groups within a panel.

proprietary technical school. This distribution is comparable to that seen in the pilot study Automotive Master Technician workshop, although, unlike the pilot study, no four-year institutions were represented on the operational panels. While no female Automotive Master Technician panelists were recruited for the pilot study, as shown in Table 40, two female instructors were recruited for the operational workshop: one for mathematics and one for reading. The overwhelming majority (92%) of panelists on all operational Automotive Master Technician panels were male, however. No real race/ethnicity balance was achieved on any of the operational Automotive Master Technician panels, as in the pilot study, with only one mathematics Panel B panelist reporting a race/ethnicity that was not White/Caucasian.

Table 39. Automotive Master Technician Operational Workshop: Panelist Distribution by Institution Type

Post-secondary Area	Content Area	Panel	Type of Institution					Total Panelists
			4-Year Public	4-Year Private	2-Year Public (Community/ Technical)	2-Year Private	Secondary	
Automotive Master Technician	Math	A	0 (0%)	0 (0%)	6 (86%)	1 (14%)	N/A	7 (100%)
		B	0 (0%)	0 (0%)	7 (88%)	1 (13%)	N/A	8 (100%)
	Reading	A	0 (0%)	0 (0%)	5 (100%)	0 (0%)	N/A	5 (100%)
		B	0 (0%)	0 (0%)	6 (100%)	0 (0%)	N/A	6 (100%)
Automotive Master Technician Totals (N = 26)			0 (0%)	0 (0%)	24 (92%)	2 (8%)	N/A	26 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

Table 40. Automotive Master Technician Operational Workshop: Panelist Distribution by Demographic Characteristics

Postsecondary Area	Content Area	Panel	Gender		Race/Ethnicity			Total Panelists
			Female	Male	White/ Caucasian	Non-White/ Caucasian	Not Specified	
Automotive Master Technician	Math	A	0 (0%)	7 (100%)	7 (100%)	0 (0%)	0 (0%)	7 (100%)
		B	1 (13%)	7 (88%)	7 (88%)	1 (13%)	0 (0%)	8 (100%)
	Reading	A	0 (0%)	5 (100%)	5 (100%)	0 (0%)	0 (0%)	5 (100%)
		B	1 (17%)	5 (83%)	5 (83%)	0 (0%)	1 (17%)	6 (100%)
Automotive Master Technician Totals (N = 26)			2 (8%)	24 (92%)	24 (92%)	1 (4%)	1 (4%)	26 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

As shown in Table 41, similar to the pilot study, all four census regions were represented on the Automotive Master Technician panels for the operational session, with the lowest percentage of panelists (19%) representing the Midwest and the highest percentage (35%) representing the South. On three of the four Automotive Master Technician replicate panels, panelists represented all four geographic regions. On reading Panel B, however, there were no panelists from the Midwest.

Table 41. Automotive Master Technician Operational Workshop: Geographic Distribution of Panelists

Postsecondary Area	Content Area	Panel	Geographic Region*				Total Panelists
			Northeast	South	Midwest	West	
Automotive Master Technician	Math	A	1 (14%)	2 (29%)	3 (43%)	1 (14%)	7 (100%)
		B	2 (25%)	3 (38%)	1 (13%)	2 (25%)	8 (100%)
	Reading	A	1 (20%)	2 (40%)	1 (20%)	1 (20%)	5 (100%)
		B	2 (33%)	2 (33%)	0 (0%)	2 (33%)	6 (100%)
Automotive Master Technician Totals (N = 26)			6 (23%)	9 (35%)	5 (19%)	6 (23%)	26 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.
*Based upon U.S. Census Bureau census regions.

Information about student populations served by the programs represented by the Automotive Master Technician operational panelists is shown in Table 42.

Table 42. Automotive Master Technician Operational Workshop: Student Populations Served by Panelists

Postsecondary Area	Content Area	Panel	Predominant Student Population Served by Program			Total Panelists
			Students Coming Directly from High School	Students Returning to School after Absence	Not Specified	
Automotive Master Technician	Math	A	4 (57%)	3 (43%)	0 (0%)	7 (100%)
		B	6 (75%)	2 (25%)	0 (0%)	8 (100%)
	Reading	A	3 (60%)	2 (40%)	0 (0%)	5 (100%)
		B	5 (83%)	0 (0%)	1 (17%)	6 (100%)
Automotive Master Technician Totals (N = 26)			18 (69%)	7 (27%)	1 (4%)	26 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

The majority of Automotive Master Technician panelists (69%) reported that their job-training programs predominantly served students coming directly from high school, which corresponds with pilot study Automotive Master Technician panelists. Across the four panels, seven panelists (27%) taught in job-training programs that predominantly served students returning to school after an absence.

Automotive Master Technician Operational Workshop Numerical Results

The following tables display standard-setting results for the Automotive Master Technician operational workshop.

When round 1 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The round 1 results are presented in Table 43.

Table 43. Automotive Master Technician Operational Workshop: Round 1 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	200	9.3	9.2	13.0	8.5
	B	166	5.4	4.6	10.4	37.0
Reading	A	321	11.9	10.3	13.8	20.2
	B	294	8.8	8.1	13.7	47.3

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 2 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 2 judgments are presented in Table 44.

Table 44. Automotive Master Technician Operational Workshop: Round 2 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	174	11.2	11.6	18.4	28.9
	B	171	2.7	2.5	5.6	31.9
Reading	A	308	12.3	9.7	12.8	32.5
	B	294	5.8	5.1	6.8	47.3

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 3 bookmarks were placed, the CAB once again calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 3 decisions are presented in Table 45.

Table 45. Automotive Master Technician Operational Workshop: Round 3 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	167	10.6	9.7	15.7	31.9
	B	171	2.5	2.4	5.6	31.9
Reading	A	308	12.3	9.7	15.6	32.5
	B	294	5.8	5.1	6.8	47.3

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.

[b] Empirical-based Standard Error.

[c] Bootstrap Standard Error.

[d] Mean Absolute Deviation.

[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

Figure 22 summarizes the MAD results across rounds for the Automotive Master Technician operational panels. While two of the four automotive panel results (mathematics Panel B and reading Panel B) were as expected (i.e., decreasing variability of panelist cut scores between rounds), the other two panels did not meet expectations. These results suggest that panelists within mathematics Panel A and reading Panel A reacted strongly and differently to the post-round feedback, and that mathematics Panel A reacted most strongly when whole-booklet feedback was presented and reading Panel A reacted most strongly when the impact data was presented. Reasons for these reactions are not readily available.

Figure 22. Automotive Master Technician Operational Workshop: Mean Absolute Deviation (MAD) of Cut Scores by Panel

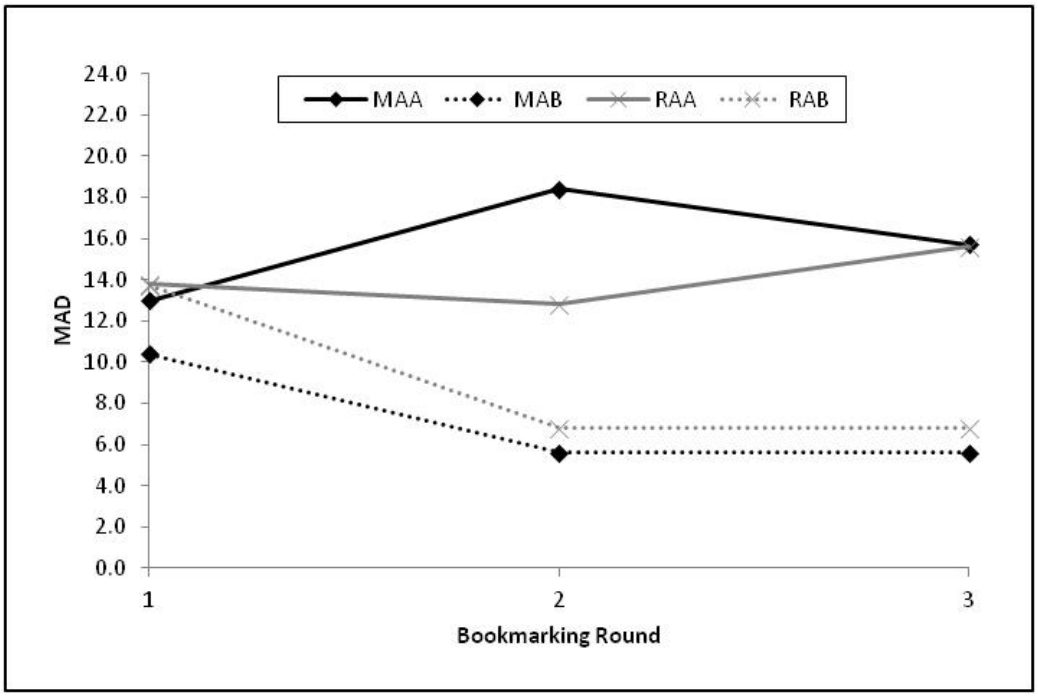


Table 46 summarizes the changes in cut scores made by individual panelists between rounds, while Table 47 presents a comparison of median and mean cut scores for each panel. The expectation was that the same number of panelists or fewer panelists changed their cut scores between rounds 2 and 3 than between rounds 1 and 2. This pattern is observed in Table 46 for three of the four panels; as indicated as well by the MAD results, reading Panel A panelists seem to have given more weight to the impact data than other groups did in making their final cut score judgments. Differences between mean and median cut scores are due to the presence of outliers (i.e., panelists who set cut scores significantly higher or lower than their peers). The largest absolute difference between the mean and median cut scores in Table 47 occurs for reading Panel A in round 2 (12.8), where the direction of the difference would result in a higher cut score being set if the mean instead of the median were used. Likewise, in the third round, the effect of outliers on mean cut scores is noticeable, mainly for reading Panel A.

Table 46. Automotive Master Technician Operational Workshop: Round-to-Round Cut Score Changes by Panel

Panel	Round	Increased N (%)	No Change N (%)	Decreased N (%)
MAA	R1–R2	0 (0.0%)	1 (14.3%)	6 (85.7%)
	R2–R3	1 (14.3%)	3 (42.9%)	3 (42.9%)
MAB	R1–R2	3 (37.5%)	3 (37.5%)	2 (25.0%)
	R2–R3	0 (0.0%)	6 (75.0%)	2 (25.0%)
RAA	R1–R2	1 (20.0%)	3 (60.0%)	1 (20.0%)
	R2–R3	1 (20.0%)	1 (20.0%)	3 (60.0%)
RAB	R1–R2	3 (50.0%)	1 (16.7%)	2 (33.3%)
	R2–R3	0 (0.0%)	6 (100.0%)	0 (0.0%)

Table 47. Automotive Master Technician Operational Workshop: Comparison of Cut Scores Based on Medians and Means

	Round 1			Round 2			Round 3		
Panel	Median	Mean	Median –Mean	Median	Mean	Median –Mean	Median	Mean	Median –Mean
MAA	200.0	197.3	**2.7**	174.0	175.9	**-1.9**	167.0	172.4	**-5.4**
MAB	166.0	168.9	**-2.9**	170.5	171.9	**-1.4**	170.5	171.6	**-1.1**
RAA	321.0	324.4	**-3.4**	308.0	320.8	**-12.8**	308.0	317.2	**-9.2**
RAB	293.5	296.0	**-2.5**	293.5	296.5	**-3.0**	293.5	296.5	**-3.0**

Automotive Master Technician Operational Workshop Exemplar Item Ratings. Following the rounds of bookmarking, panelists were asked to identify items that, if responded to correctly (i.e., MC items) or if responses earned a specified number of points (i.e., CR items), would exemplify preparedness for admission into a job-training program.

Table 48 presents a summary of the number of items presented to each panel and the number of items for which panelists expressed 100% agreement and at least 75% agreement that the item/score point would be at least OK to demonstrate preparedness. The average scale value for items selected as at least OK appears parenthetically within the table as well.

Table 48. Automotive Master Technician Operational Workshop: Exemplar Item Summary

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)
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)

Automotive Master Technician Operational Workshop Process Evaluation Results

Following this presentation of the process evaluation results, reactions to the consequences data are summarized. The process evaluation questionnaires for this workshop are presented in their entirety in Appendix Q; along with the questions, the appendix shows the frequency of responses per Likert-scale category, and the average response.

Understanding of Tasks. Across all panels, when panelists were asked to respond to the following statement: "My understanding of the tasks I was to accomplish during each round was . . ." the average response was 4.2 or higher, which corresponds to a verbal description of at least "Adequate." Average results by panel can be found in Table 49.

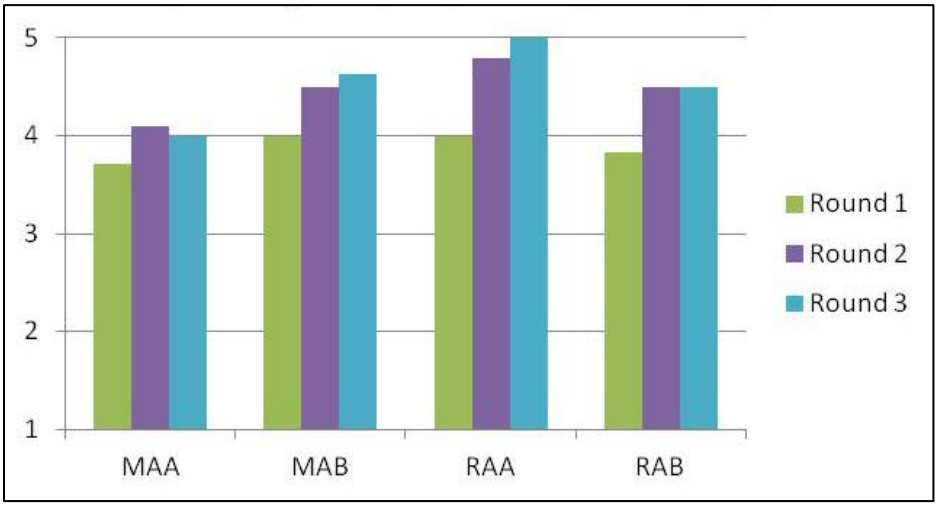
Table 49. Automotive Master Technician Operational Workshop: Summary of Selected Evaluation Items by Panel

Evaluation Item	Average Response by Panel			
	Automotive Master Technician			
	Mathematics		*Reading*	
	MAA	MAB	RAA	RAB
My understanding of the tasks I was to accomplish during each round was . . . (1 = Totally Inadequate; 5 = Totally Adequate)	4.4	4.3	4.4	4.2
The most accurate description of my level of confidence in the cut score recommendations I provided was . . . (1 = Not at All Confident; 5 = Very Confident)	3.8	4.4	4.6	4.3
I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark. (1 = Totally Disagree; 5 = Totally Agree)	3.7	4.0	4.0	3.5
The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items. (1 = Totally Disagree; 5 = Totally Agree)	2.8	3.5	3.7	3.9
I felt pressured by others in my group to make my cut score recommendation agree with theirs. (1 = Totally Disagree; 5 = Totally Agree)	1.6	1.1	1.6	1.0
During the standard setting process, I found using CAB to be . . . (1 = Not at All Helpful; 5 = Very Helpful)	4.4	4.5	4.6	4.2

Understanding of the Borderline Performance Description. After each bookmarking round, panelists were asked to respond to the following statement: *"At the time I placed my bookmark, my understanding of the BPD was . . ."* For round 1, the average response was 3.7 or higher, which corresponds to a verbal description of at least "Somewhat Adequate." Across all panels, the average reported understanding of the BPD steadily increased or remained the same between rounds, with the lowest average being 4.0 after the third round of bookmarking. Average results by panel are shown in Figure 23.

Figure 23. Automotive Master Technician Operational Workshop: At the time I placed my bookmark, my understanding of the BPD was . . .

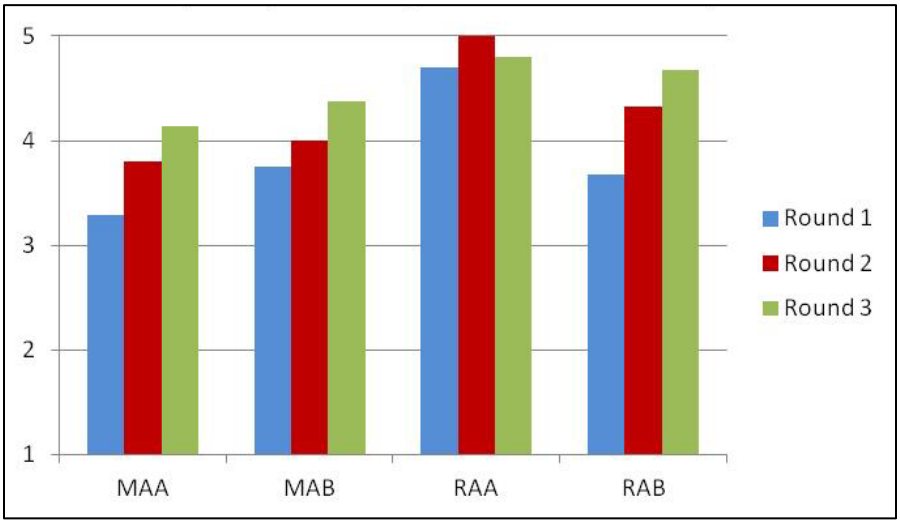
(1 = Totally Inadequate; 5 = Totally Adequate)



Comfort and Confidence. Several statements were developed to solicit how comfortable and/or confident panelists were in setting their cut scores or with components of the standard-setting process. After each bookmarking round, panelists were asked to respond to the following statement*: "The most accurate description of my level of confidence in my bookmark placement is . . ."* For round 1, the average response was 3.3 or higher, which corresponds to a verbal description of at least "Somewhat Confident." Across three of the four panels, the average reported comfort level steadily increased across bookmarking rounds, with the lowest average after the third round being 4.1. One panel, reading Panel A, maintained a high level of confidence across all rounds, with average responses recorded as 4.7, 5.0, and 4.8 for each round, respectively. Average results by panel are shown in Figure 24.

Figure 24. Automotive Master Technician Operational Workshop: The most accurate description of my level of confidence in my bookmark placement is . . .

(1 = Not at All Confident; 5 = Very Confident)

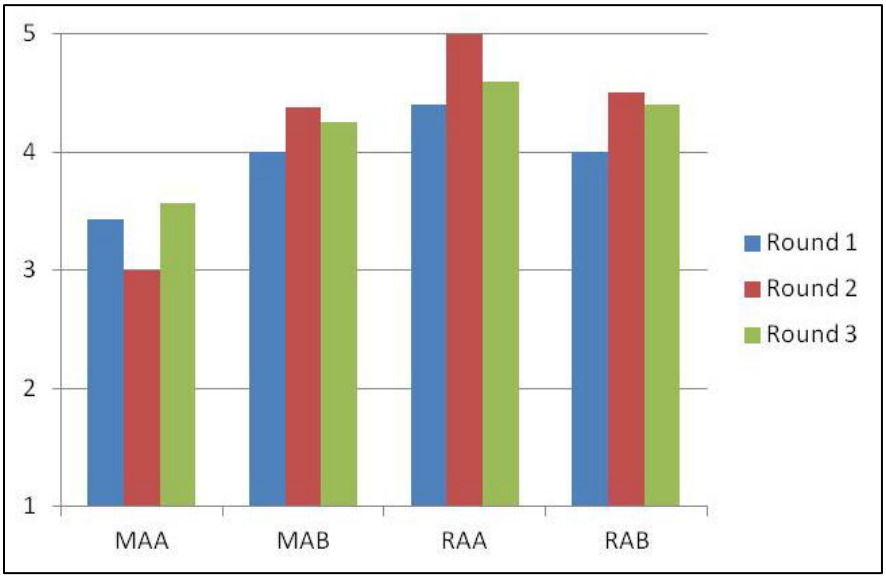


As a follow-up to this statement, panelists were asked to respond to the following statement: *"The most accurate description of my level of confidence in the cut score recommendations I provided was . . ."* Across all panels, the average response was at least 3.8, which corresponds to a verbal description of at least "Somewhat Confident." Average results by panel can be found in Table 49 (p. 131).

Additionally, after each round of bookmarking, panelists were asked to respond to the following statement: *"I believe my cut score is consistent with the BPD."* After round 1, all panels had an average response of at least 3.4, which corresponds to a verbal description of at least "Somewhat Agree." While consistent increases between rounds were not observed, averages were higher by the third round of bookmarking, with the lowest average being 3.6. Average results by panel are shown in Figure 25.

Figure 25. Automotive Master Technician Operational Workshop: I believe my cut score is consistent with the BPD

(1 = Totally Disagree; 5 = Totally Agree)



Specific to the standard-setting method are the reliance on an understanding of how to use a response probability in placing a bookmark and a general acceptance that items are ordered by relative difficulty. Thus, each of these assumptions was evaluated. Panelists were asked to respond to the following statement: *"I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark."* Across all panels, the average response was at least 3.5, which corresponds to a verbal description of at least "Somewhat Agree." Average results by panel can be found in Table 49 (p. 131).

Panelists were also asked to respond to the following statement: *"The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items."* Average responses across panels varied from 2.8 to 3.9, which corresponds to verbal descriptions between "Disagree" to "Agree." Average results by panel can be found in Table 49 (p. 131).

Independence of Judgment. Across all panels, when panelists were asked to respond to the following statement: *"I felt pressured by others in my group to make my cut score recommendation agree with theirs,"* the average response was no higher than 1.6, which corresponds to a verbal description of "Disagree" or "Totally Disagree." Average results by panel can be found in Table 49 (p. 131).

Helpfulness of Software. As the CAB software system was implemented to aid the standard-setting process, panelists were asked to respond to the following statement: *"During the standard setting process, I found using CAB to be . . ."* Across all panels, the lowest average response was 4.2, which corresponds to a verbal description of "Helpful" or higher. Average results by panel can be found in Table 49 (p. 131).

Panelists were asked to review both their final round 3 cut scores and the associated impact data. They were asked to respond to the following questions with a "Yes" or "No" response:

- *Does the [impact data] percentage reflect your expectations about the proportion of students whose NAEP score would indicate at least minimal preparedness for placement?*
- *Would you change the cut score recommended by your panel to the Governing Board if you could?*

Results across panels varied significantly, with 0% to 80% of panelists (depending on which panel they belonged to) responding "Yes" to the first question and 40% to 100% of panelists responding "No" to the second question. The full set of results can be found in Appendix Q.

Operational Session 2

The second operational JSS session was conducted on June 7–10, 2011, at the Westin Hotel in St. Louis, Missouri, and included LPN and Pharmacy Technician workshops.

For this JSS session, the JSS process was largely implemented as described in the Judgmental Standard-Setting Process subsection of this report (pp. 39–69), with modifications made as described in the Pilot Study (pp. 72–101) and Operational Session 1 (pp. 102–106) sections of this report. Based on feedback from pilot study panelists, facilitators, and observers, however, the following additional enhancements were implemented.

Following the KSA training provided by the content facilitators, panelists reviewed KSAs for the four assigned CR items in their panel groups. During this activity, process facilitators had access to anchor descriptions for all partial- and full-credit responses for each item and were instructed to use these descriptions to help guide panelists in meaningful discussions.

Two enhancements implemented for both the second and third operational JSS sessions related to the presentation of whole-booklet feedback. First, in addition to scoring guides, panelists were also given a crosswalk between the items in the whole-booklet forms and the items' locations in the OIBs and CROIBs, which they used to access the OIBs and CROIBs for item information. For reading, this crosswalk included the passage associated with each block in each whole-booklet form. Second, booklets that showed reversals between scaled scores and percentages of expected possible raw score were removed from the set of booklets from which the whole-booklet feedback was selected. This reversal is an artifact of the psychometric procedure and caused some confusion among pilot study panelists.

Finally, given observed similarities between the Automotive Master Technician, LPN, and Pharmacy Technician panelists' responses to the Content Objectives forms, the operational Automotive Master Technician BPDs were used as the starting point for the Operational Session 2 BPDs. The content facilitators consolidated the Automotive Master Technician BPDs and the Operational Session 2 panelists' responses to the online Content Objectives form to develop preliminary BPD drafts to share with the Operational Session 2 panelists. The BPD sessions on the first day of the standard-setting workshops addressed discrepancies between the KSAs identified by the Automotive Master Technician panelists and those identified by the LPN and Pharmacy Technician panelists.

As presented in subsequent sections of this report, larger percentages of panelists were recruited from private LPN and Pharmacy Technician programs than had been recruited from private Automotive Master Technician programs. Anecdotal input from LPN and Pharmacy Technician panelists suggested that the eligibility requirements from private institutions might vary across institutions and differ from those of public institutions. These differences were also discussed during BPD review sessions. The preliminary mathematics and reading BPDs developed by the content facilitators in advance of the standard-setting session and the final BPD versions agreed upon by the mathematics and reading pairs of replicate panels for each postsecondary area are provided in Appendix O.

The abbreviations used to describe panels in the second operational session are displayed in Table 50.

Table 50. Operational Session 2: Abbreviations Used to Describe Panels

Abbreviation	Content Area	Occupation	Panel
MLA	Mathematics	LPN	A
MLB			B
RLA	Reading	LPN	A
RLB			B
MPA	Mathematics	Pharmacy Technician	A
MPB			B
RPA	Reading	Pharmacy Technician	A
RPB			B

LPN Operational Workshop

The following subsections describe the LPN operational workshop panelists, numerical results, and process evaluation results.

LPN Operational Workshop Panelists

Thirty-three LPN job-training programs were represented in the pool of LPN operational panelists. From these programs, 40 LPN panelists were recruited.[15] Table 51 displays the distribution of panelists by type of institution, and Table 52 displays the distribution of panelists by demographic characteristics. The Design Document called for a total of 40 panelists for this workshop: 20 in mathematics and 20 in reading, with each group divided into two replicate panels of 10 panelists. This target was achieved.

As shown in Table 51, nearly all LPN panelists were recruited from either two-year public institutions (60%) or two-year private institutions (38%); only one panelist (3%) was recruited from a four-year public institution. The LPN panels reflected a higher percentage of private institutions than was found on the Automotive Master Technician operational panels (8%). Also in contrast to the Automotive Master Technician panels, the majority (95%) of LPN panelists were women, as shown in Table 52. While some diversity in race/ethnicity was achieved on each LPN panel, the majority (80%) of panelists reported themselves to be White/Caucasian.

[15]Five programs provided two panelists each, and one program provided three panelists. Multiple panelists from the same program were assigned to different content areas and/or panels when possible; when not possible, they were assigned to different table groups within a panel.

Table 51. LPN Operational Workshop: Panelist Distribution by Institution Type

Postsecondary Area	Content Area	Panel	Type of Institution				Total Panelists
			4-Year	2-Year Public (Community/ Technical)	2-Year Private	Secondary	
LPN	Math	A	1 (10%)	5 (50%)	4 (40%)	N/A	10 (100%)
		B	0 (0%)	6 (60%)	4 (40%)	N/A	10 (100%)
	Reading	A	0 (0%)	6 (60%)	4 (40%)	N/A	10 (100%)
		B	0 (0%)	7 (70%)	3 (30%)	N/A	10 (100%)
LPN Totals (N = 40)			1 (3%)	24 (60%)	15 (38%)	N/A	40 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

Table 52. LPN Operational Workshop: Panelist Distribution by Demographic Characteristics

Postsecondary Area	Content Area	Panel	Gender		Race/Ethnicity			Total Panelists
			Female	Male	White/ Caucasian	Non-White/ Caucasian	Not Specified	
LPN	Math	A	10 (100%)	0 (0%)	8 (80%)	2 (20%)	0 (0%)	10 (100%)
		B	10 (100%)	0 (0%)	9 (90%)	1 (10%)	0 (0%)	10 (100%)
	Reading	A	9 (90%)	1 (10%)	7 (70%)	3 (30%)	0 (0%)	10 (100%)
		B	9 (90%)	1 (10%)	8 (80%)	2 (20%)	0 (0%)	10 (100%)
LPN Totals (N = 40)			38 (95%)	2 (5%)	32 (80%)	8 (20%)	0 (0%)	40 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

As shown in Table 53, all four census regions were represented on the LPN panels, with the lowest percentage of panelists (8%) representing the West and the highest percentage (43%) representing the South. On three of the four LPN replicate panels, panelists represented all four geographic regions. On reading Panel A, however, there were no panelists from the West.

Table 53. LPN Operational Workshop: Geographic Distribution of Panelists

Postsecondary Area	Content Area	Panel	Geographic Region*				Total Panelists
			Northeast	South	Midwest	West	
LPN	Math	A	1 (10%)	5 (50%)	3 (30%)	1 (10%)	10 (100%)
		B	1 (10%)	4 (40%)	4 (40%)	1 (10%)	10 (100%)
	Reading	A	2 (20%)	5 (50%)	3 (30%)	0 (0%)	10 (100%)
		B	3 (30%)	3 (30%)	3 (30%)	1 (10%)	10 (100%)
LPN Totals (N = 40)			7 (18%)	17 (43%)	13 (33%)	3 (8%)	40 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.
*Based upon U.S. Census Bureau census regions.

Information about student populations served by the programs represented by the LPN operational panelists is shown in Table 54.

Table 54. LPN Operational Workshop: Student Populations Served by Panelists

Postsecondary Area	Content Area	Panel	Predominant Student Population Served by Program			Total Panelists
			Students Coming Directly from High School	Students Returning to School after Absence	Not Specified	
LPN	Math	A	2 (20%)	8 (80%)	0 (0%)	10 (100%)
		B	2 (20%)	8 (80%)	0 (0%)	10 (100%)
	Reading	A	1 (10%)	9 (90%)	0 (0%)	10 (100%)
		B	0 (0%)	10 (100%)	0 (0%)	10 (100%)
LPN Totals (N = 40)			5 (13%)	35 (88%)	0 (0%)	40 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

As reported by panelists, the majority of LPN panelists (88%) represented job-training programs that predominantly served students returning to school after a year or more of absence, regardless of whether the panelists came from public or private institutions. This is in contrast to the

Automotive Master Technician panels, on which only 27% reported predominantly serving returning students.

LPN Operational Workshop Numerical Results

The following tables display standard-setting results for the LPN operational workshop.

When round 1 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The round 1 results are presented in Table 55. Interestingly, panelists on mathematics Panel A had a high level of agreement on where the cut score should be placed, as evidenced by the low standard error in comparison to all other panels. Likewise, it is noteworthy that, even while mathematics Panel B experienced more variability among panelists, its cut score after the first round is only two scale score points from that of mathematics Panel A.

Table 55. LPN Operational Workshop: Round 1 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	177	1.3	1.3	4.0	25.8
	B	179	10.0	9.6	16.8	24.0
Reading	A	308	12.2	11.6	19.2	32.5
	B	282	12.2	11.7	17.5	60.4

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 2 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 2 judgments are presented in Table 56. While round 1 cut scores were close to one another for mathematics Panels A and B,

this was not the case after round 2, suggesting that Panel B reacted strongly to the whole-booklet feedback presented, while Panel A did not.

Table 56. LPN Operational Workshop: Round 2 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	177	2.0	1.6	4.4	25.8
	B	199	4.6	4.9	7.4	9.0
Reading	A	311	6.2	6.0	12.9	29.5
	B	288	3.2	3.5	7.7	54.1

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 3 bookmarks were placed, the CAB once again calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 3 decisions are presented in Table 57.

Table 57. LPN Operational Workshop: Round 3 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	177	0.5	0.5	1.7	25.8
	B	193	6.6	6.6	8.5	12.7
Reading	A	307	5.2	4.8	8.5	33.5
	B	288	2.0	2.3	4.9	54.1

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

Figure 26 summarizes the MAD results across rounds for the LPN operational panels. While the two reading panel results are as expected (decreasing variability of panelist cut scores between

rounds), the two mathematics panel results did not meet expectations. For mathematics Panel A, a slight increase in variability occurred between rounds 1 and 2, which may suggest that panelists either reacted differently to the whole-booklet feedback presented or remained strong in their original judgments. Panelists within mathematics Panel B may have reacted strongly and differently to the impact data.

Figure 26. LPN Operational Workshop: Mean Absolute Deviation (MAD) of Cut Scores by Panel

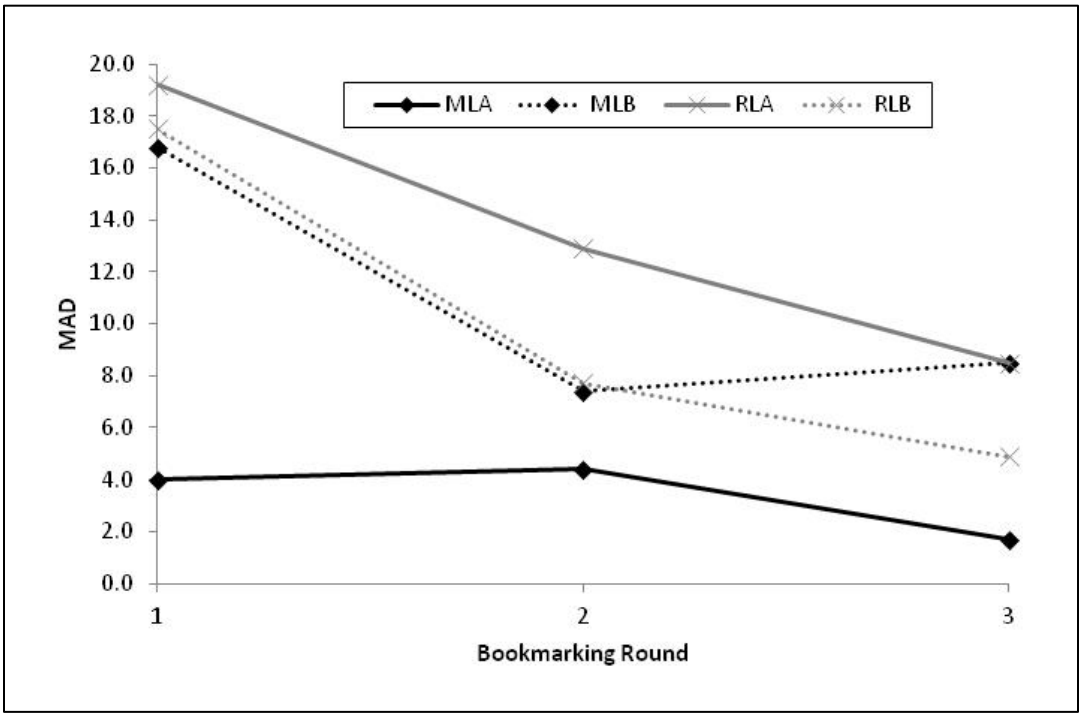


Table 58 summarizes the changes in cut scores made by individual panelists between rounds, while Table 59 presents a comparison of median and mean cut scores for each panel. The expectation was that the same number of panelists or fewer panelists changed their cut scores between rounds 2 and 3 than between rounds 1 and 2. This pattern is observed in Table 58. Differences between mean and median cut scores are due to the presence of outliers (i.e., panelists who set cut scores significantly higher or lower than their peers). The largest absolute difference between the mean and median cut scores in Table 59 occurs for reading Panel B in

round 1 (7.1), where the direction of the difference would result in a higher cut score being set if the mean instead of the median were used. The effect of outliers on mean cut scores is negligible after the third round.

Table 58. LPN Operational Workshop: Round-to-Round Cut Score Changes by Panel

Panel	Round	Increased N (%)	No Change N (%)	Decreased N (%)
MLA	R1–R2	4 (40.0%)	5 (50.0%)	1 (10.0%)
	R2–R3	0 (0.0%)	7 (70.0%)	3 (30.0%)
MLB	R1–R2	6 (60.0%)	2 (20.0%)	2 (20.0%)
	R2–R3	2 (20.0%)	2 (20.0%)	6 (60.0%)
RLA	R1–R2	5 (50.0%)	1 (10.0%)	4 (40.0%)
	R2–R3	1 (10.0%)	7 (70.0%)	2 (20.0%)
RLB	R1–R2	4 (40.0%)	2 (20.0%)	4 (40.0%)
	R2–R3	2 (20.0%)	4 (40.0%)	4 (40.0%)

Table 59. LPN Operational Workshop: Comparison of Cut Scores Based on Medians and Means

Panel	Round 1			Round 2			Round 3		
	Median	Mean	Median –Mean	Median	Mean	Median –Mean	Median	Mean	Median –Mean
MLA	177.0	177.4	**-0.4**	177.0	180.2	**-3.2**	177.0	177.5	**-0.5**
MLB	179.0	183.0	**-4.0**	198.5	195.6	**2.9**	192.5	191.7	**0.8**
RLA	308.0	313.0	**-5.0**	311.0	315.3	**-4.3**	307.0	309.9	**-2.9**
RLB	282.0	289.1	**-7.1**	288.0	288.1	**-0.1**	288.0	286.1	**1.9**

LPN Operational Workshop Exemplar Item Ratings. Following the rounds of bookmarking, panelists were asked to identify items that, if responded to correctly (i.e., MC items) or if responses earned a specified number of points (i.e., CR items), would exemplify preparedness for acceptance into a job-training program.

Table 60 presents a summary of the number of items presented to each panel and the number of items for which panelists expressed 100% agreement and at least 75% agreement that the

item/score point would be at least OK to demonstrate preparedness. The average scale value for items selected as at least OK appears parenthetically within the table as well.

Table 60. LPN Operational Workshop: Exemplar Item Summary

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)
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)

LPN Operational Workshop Process Evaluation Results

Following this presentation of the process evaluation results, reactions to the consequences data are summarized. The process evaluation questionnaires for this workshop are presented in their entirety in Appendix R; along with the questions, the appendix shows the frequency of responses per Likert-scale category, and the average response.

Understanding of Tasks. Across all panels, when panelists were asked to respond to the following statement: *"My understanding of the tasks I was to accomplish during each round was . . ."* the average response was 4.2 or higher, which corresponds to a verbal description of "Adequate" or "Totally Adequate." Average results by panel can be found in Table 61.

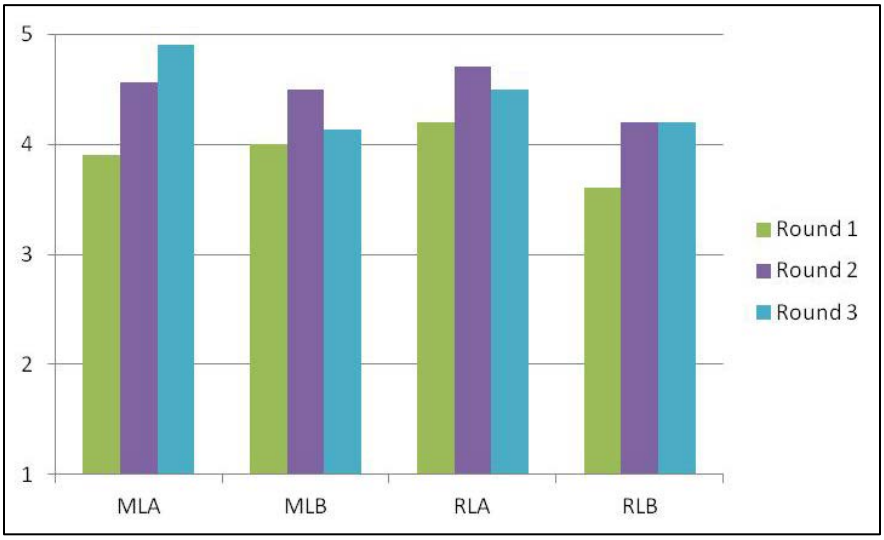
Table 61. LPN Operational Workshop: Summary of Selected Evaluation Items by Panel

Evaluation Item	Average Response by Panel			
	LPN			
	Mathematics		*Reading*	
	MCA	MCB	RCA	RCB
My understanding of the tasks I was to accomplish during each round was . . . (1 = Totally Inadequate; 5 = Totally Adequate)	4.6	4.6	4.6	4.2
The most accurate description of my level of confidence in the cut score recommendations I provided was . . . (1 = Not at All Confident; 5 = Very Confident)	4.8	4.3	4.4	4.0
I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark. (1 = Totally Disagree; 5 = Totally Agree)	3.5	3.6	3.7	3.4
The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items. (1 = Totally Disagree; 5 = Totally Agree)	3.6	3.4	3.6	3.6
I felt pressured by others in my group to make my cut score recommendation agree with theirs. (1 = Totally Disagree; 5 = Totally Agree)	1.5	1.1	1.8	2.0
During the standard setting process, I found using CAB to be . . . (1 = Not at All Helpful; 5 = Very Helpful)	4.6	4.7	4.2	3.9

Understanding of the Borderline Performance Description. After each bookmarking round, panelists were asked to respond to the following statement: *"At the time I placed my bookmark, my understanding of the BPD was . . ."* For round 1, the average response was 3.6 or higher, which corresponds to a verbal description of at least "Somewhat Adequate." Across all panels, the average reported understanding of the BPD increased between rounds 1 and 3, with the lowest average being 4.1 after the third round of bookmarking. Average results by panel are shown in Figure 27.

Figure 27. LPN Operational Workshop: At the time I placed my bookmark, my understanding of the BPD was . . .

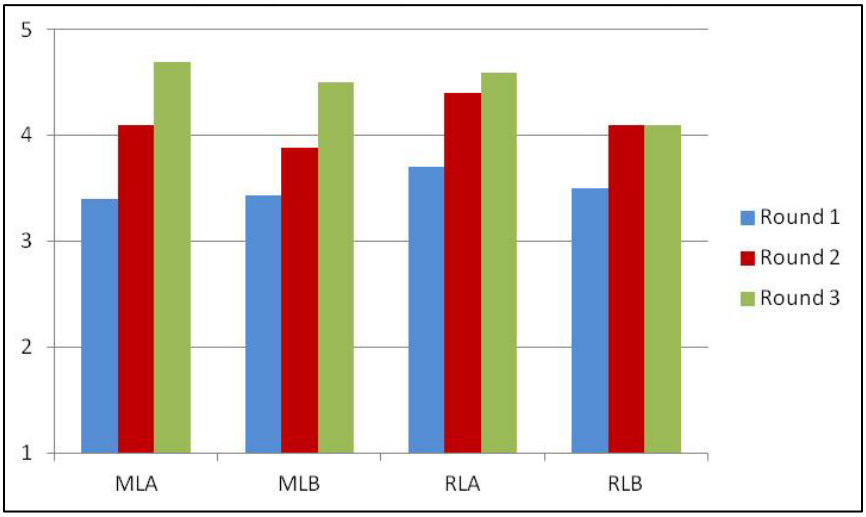
(1 = Totally Inadequate; 5 = Totally Adequate)



Comfort and Confidence. Several statements were developed to solicit how comfortable and/or confident panelists were in setting their cut scores or with components of the standard-setting process. After each bookmarking round, panelists were asked to respond to the following statement: "*The most accurate description of my level of confidence in my bookmark placement is . . .*" For round 1, the average response was 3.4 or higher, which corresponds to a verbal description of at least "Somewhat Confident." Across all panels, the average reported comfort level steadily increased or was maintained between rounds, with the lowest average after the third round of bookmarking being 4.1. Average results by panel are shown in Figure 28.

Figure 28. LPN Operational Workshop: The most accurate description of my level of confidence in my bookmark placement is . . .

(1 = Not at All Confident; 5 = Very Confident)

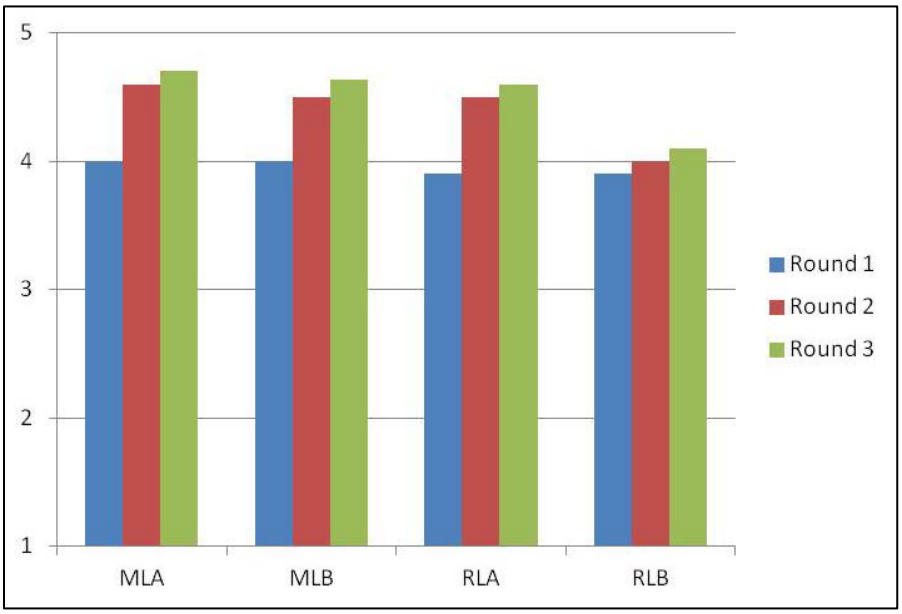


As a follow-up to this statement, panelists were asked to respond to the following statement: *"The most accurate description of my level of confidence in the cut score recommendations I provided was . . ."* Across all panels, the average response was at least 4.0, which corresponds to a verbal description of at least "Confident." Average results by panel can be found in Table 61 (p. 147).

Additionally, after each round of bookmarking, panelists were asked to respond to the following statement: *"I believe my cut score is consistent with the BPD."* After round 1, all panels had an average response of at least 3.9, which corresponds to a verbal description of at least "Somewhat Agree." Further, agreement increased for all panels by the end of the third bookmarking round with the lowest average being 4.1. Average results by panel are shown in Figure 29.

Figure 29. LPN Operational Workshop: I believe my cut score is consistent with the BPD

(1 = Totally Disagree; 5 = Totally Agree)



Specific to the standard-setting method are the reliance on an understanding of how to use a response probability in placing a bookmark and a general acceptance that items are ordered by relative difficulty. Thus, each of these assumptions was evaluated.

In order to evaluate these assumptions, panelists were asked to respond to the following statement: *"I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark."* Across all panels, the average response was at least 3.4, which corresponds to a verbal description of at least "Somewhat Agree." Average results by panel can be found in Table 61 (p. 147).

Panelists were also asked to respond to the following statement: *"The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items."* Average responses to this statement across panels only varied from 3.4 to 3.6, which corresponds to verbal

descriptions between “Somewhat Agree” and “Agree.” Average results by panel can be found in Table 61 (p. 147).

Independence of Judgment. Across all panels, when panelists were asked to respond to the following statement: *“I felt pressured by others in my group to make my cut score recommendation agree with theirs,”* the average response was no higher than 2.0, which corresponds to a verbal description of “Disagree” or “Totally Disagree.” Average results by panel can be found in Table 61 (p. 147).

Helpfulness of Software. As the CAB software system was implemented to aid the standard-setting process, panelists were asked to respond to the following statement: *“During the standard setting process, I found using CAB to be . . .”* Across all panels, the lowest average response was 3.9, which corresponds to a verbal description of “Somewhat Helpful” or higher. Average results by panel can be found in Table 61 (p. 147).

Panelists were asked to review both their final round 3 cut scores and the associated impact data. They were asked to respond to the following questions with a “Yes” or “No” response:

- *Does the [impact data] percentage reflect your expectations about the proportion of students whose NAEP score would indicate at least minimal preparedness for placement?*
- *Would you change the cut score recommended by your panel to the Governing Board if you could?*

Ninety percent to 100% of panelists (depending on which panel they belonged) responded “Yes” to the first question, and 70% to 100% of panelists responded “No” to the second question,

indicating that most panelists were satisfied with the final cut scores. The full set of results can be found in Appendix R.

Pharmacy Technician Operational Workshop

The following subsections describe the Pharmacy Technician operational workshop panelists, numerical results, and process evaluation results.

Pharmacy Technician Operational Workshop Panelists

Thirty-four job-training programs were represented in the pool of Pharmacy Technician panelists. From these programs, 36 Pharmacy Technician panelists were recruited to participate in the second operational session.[16] Table 62 displays the distribution of panelists by type of institution, and Table 63 displays the distribution of panelists by demographic characteristics.

The Design Document called for a total of 40 panelists for this workshop: 20 in mathematics and 20 in reading, with each group divided into two replicate panels of 10 panelists. Due to a series of last-minute cancellations for personal and professional reasons, two Pharmacy Technician mathematics panel slots and one Pharmacy Technician reading panel slot were unfilled, resulting in a total of 18 Pharmacy Technician mathematics panelists and 19 Pharmacy Technician reading panelists.

As shown in Table 62, the majority of Pharmacy Technician panelists (68%) were recruited from private institutions; the remaining 12 panelists (32%) were from public two-year community or two-year technical colleges. This is in contrast to the distribution of Automotive Master Technician panelists, of which 8–11% were from private institutions, and of LPN panelists, of which 38% were from private institutions. As shown in Table 63, the majority of Pharmacy

[16]Within this sample were six networks of schools that operated across multiple states. Schools within a network that operated in different states were considered to be unique. From within these six networks, three schools had campuses that provided two panelists each. Panelists coming from the same campus were assigned to different panels.

Technician panelists (62%) were women. Some degree of racial/ethnic diversity was achieved for Pharmacy Technician panels; while the majority of Pharmacy Technician panelists (59%) reported themselves to be White/Caucasian, between three (30%–33%) and four (44%) individuals on each Pharmacy Technician panel self-reported as being non–White/Caucasian.

Table 62. Pharmacy Technician Operational Workshop: Panelist Distribution by Institution Type

Postsecondary Area	Content Area	Panel	Type of Institution				Total Panelists
			4-Year	2-Year Public (Community/ Technical)	2-Year Private	Secondary	
Pharmacy Technician	Math	A	0 (0%)	5 (56%)	4 (44%)	N/A	9 (100%)
		B	0 (0%)	1 (11%)	8 (89%)	N/A	9 (100%)
	Reading	A	0 (0%)	3 (30%)	7 (70%)	N/A	10 (100%)
		B	0 (0%)	3 (33%)	6 (67%)	N/A	9 (100%)
Pharmacy Technician Totals (N = 37)			0 (0%)	12 (32%)	25 (68%)	N/A	37 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

Table 63. Pharmacy Technician Operational Workshop: Panelist Distribution by Demographic Characteristics

Postsecondary Area	Content Area	Panel	Gender		Race/Ethnicity			Total Panelists
			Female	Male	White/ Caucasian	Non-White/ Caucasian	Not Specified	
Pharmacy Technician	Math	A	5 (56%)	4 (44%)	5 (56%)	4 (44%)	0 (0%)	9 (100%)
		B	5 (56%)	4 (44%)	5 (56%)	3 (33%)	1 (11%)	9 (100%)
	Reading	A	6 (60%)	4 (40%)	7 (70%)	3 (30%)	0 (0%)	10 (100%)
		B	7 (78%)	2 (22%)	5 (56%)	4 (44%)	0 (0%)	9 (100%)
Pharmacy Technician Totals (N = 37)			23 (62%)	14 (38%)	22 (59%)	14 (38%)	1 (3%)	37 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

As shown in Table 64, all four census regions were represented on the Pharmacy Technician panels, with the lowest percentage of panelists (8%) representing the Northeast and the highest percentage (49%) representing the South. On three of the four Pharmacy Technician panels, panelists represented all four geographic regions. On mathematics Panel A, however, there were no panelists from the Northeast. It should be noted that four reading panelists came from one state (South Carolina), with each South Carolina panelist assigned to a different table group. With the exception of the HVAC reading panels, this is the only instance of four panelists within the same postsecondary area and content area representing the same state.

Table 64. Pharmacy Technician Operational Workshop: Geographic Distribution of Panelists

Postsecondary Area	Content Area	Panel	Geographic Region*				Total Panelists
			Northeast	South	Midwest	West	
Pharmacy Technician	Math	A	0 (0%)	4 (44%)	2 (22%)	3 (33%)	9 (100%)
		B	1 (11%)	5 (56%)	2 (22%)	1 (11%)	9 (100%)
	Reading	A	1 (10%)	5 (50%)	2 (20%)	2 (20%)	10 (100%)
		B	1 (11%)	4 (44%)	1 (11%)	3 (33%)	9 (100%)
Pharmacy Technician Totals (N = 37)			3 (8%)	18 (49%)	7 (19%)	9 (24%)	37 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.
*Based upon U.S. Census Bureau census regions.

Information about student populations served by the programs represented by the Pharmacy Technician operational panelists is shown in Table 65.

Table 65. Pharmacy Technician Operational Workshop: Student Populations Served by Panelists

Postsecondary Area	Content Area	Panel	Predominant Student Population Served by Program			Total Panelists
			Students Coming Directly from High School	Students Returning to School after Absence	Not Specified	
Pharmacy Technician	Math	A	4 (44%)	4 (44%)	1 (11%)	9 (100%)
		B	1 (11%)	8 (89%)	0 (0%)	9 (100%)
	Reading	A	3 (30%)	7 (70%)	0 (0%)	10 (100%)
		B	3 (33%)	6 (67%)	0 (0%)	9 (100%)
Pharmacy Technician Totals (N = 37)			11 (30%)	25 (68%)	1 (3%)	37 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

As reported by panelists, the majority of Pharmacy Technician panelists (68%) represented job-training programs that predominantly served students returning to school after a year or more of absence.

Pharmacy Technician Operational Workshop Numerical Results

The following tables display standard-setting results for the Pharmacy Technician operational workshop.

When round 1 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The round 1 results are presented in Table 66.

Table 66. Pharmacy Technician Operational Workshop: Round 1 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	174	3.8	3.3	7.6	28.9
	B	214	9.1	10.9	20.7	3.4
Reading	A	308	6.2	5.3	11.0	32.5
	B	289	6.0	5.4	12.1	52.9

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher

When round 2 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 2 judgments are presented in Table 67.

Table 67. Pharmacy Technician Operational Workshop: Round 2 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	174	3.7	3.0	6.1	28.9
	B	206	2.8	2.9	5.4	5.8
Reading	A	322	2.8	2.8	5.0	19.3
	B	299	5.5	5.4	6.8	41.9

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 3 bookmarks were placed, the CAB once again calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 3 decisions are presented in Table 68.

Table 68. Pharmacy Technician Operational Workshop: Round 3 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	174	1.8	2.1	5.2	28.9
	B	176	6.3	6.9	11.0	26.9
Reading	A	321	3.1	3.3	6.1	20.2
	B	299	5.5	5.4	6.8	41.9

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.

[b] Empirical-based Standard Error.

[c] Bootstrap Standard Error.

[d] Mean Absolute Deviation.

[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

Figure 30 summarizes the MAD results across rounds for the Pharmacy Technician operational panels. While the variability of panelist cut scores decreased for all panels between rounds 1 and 2, cut score variability increased for mathematics Panel B and (to a lesser extent) for reading Panel A. This increase in variability suggests that these panels had stronger and different reactions to the impact data presented.

Figure 30. Pharmacy Technician Operational Workshop: Mean Absolute Deviation (MAD) of Cut Scores by Panel

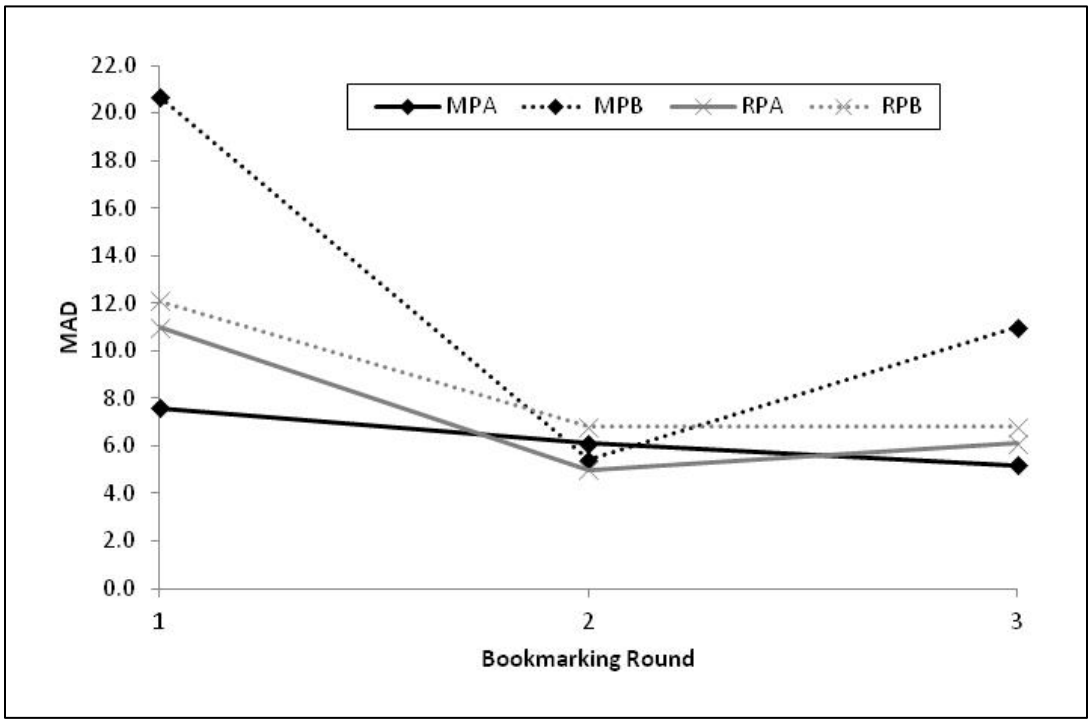


Table 69 summarizes the changes in cut scores made by individual panelists between rounds, while Table 70 presents a comparison of median and mean cut scores for each panel. The expectation was that the same number of panelists or fewer panelists changed their cut scores between rounds 2 and 3 than between rounds 1 and 2. This pattern is observed in Table 69. Differences between mean and median cut scores are due to the presence of outliers (i.e., panelists who set cut scores significantly higher or lower than their peers). The largest absolute difference between the mean and median cut scores in Table 70 occurs for reading Panel B in round 1 (17.1), where the direction of the difference would result in a lower cut score being set if the mean instead of the median were used. In the third round, the effect of outliers on mean cut scores is more negligible.

Table 69. Pharmacy Technician Operational Workshop: Round-to-Round Cut Score Changes by Panel

Panel	Round	Increased N (%)	No Change N (%)	Decreased N (%)
MPA	R1–R2	3 (33.3%)	3 (33.3%)	3 (33.3%)
	R2–R3	0 (0.0%)	4 (44.4%)	5 (55.6%)
MPB	R1–R2	3 (33.3%)	1 (11.1%)	5 (55.6%)
	R2–R3	0 (0.0%)	1 (11.1%)	8 (88.9%)
RPA	R1–R2	7 (70.0%)	0 (0.0%)	3 (30.0%)
	R2–R3	0 (0.0%)	7 (70.0%)	3 (30.0%)
RPB	R1–R2	6 (66.7%)	2 (22.2%)	1 (11.1%)
	R2–R3	0 (0.0%)	9 (100.0%)	0 (0.0%)

Table 70. Pharmacy Technician Operational Workshop: Comparison of Cut Scores Based on Medians and Means

	Round 1			Round 2			Round 3		
Panel	Median	Mean	Median –Mean	Median	Mean	Median –Mean	Median	Mean	Median –Mean
MPA	174.0	179.6	**-5.6**	174.0	180.1	**-6.1**	174.0	174.8	**-0.8**
MPB	214.0	196.9	**17.1**	206.0	203.9	**2.1**	176.0	173.0	**3.0**
RPA	307.5	312.2	**-4.7**	322.0	322.0	**0.0**	321.0	319.5	**1.5**
RPB	289.0	294.0	**-5.0**	299.0	299.1	**-0.1**	299.0	299.1	**-0.1**

Pharmacy Technician Operational Workshop Exemplar Item Ratings. Following the rounds of bookmarking, panelists were asked to identify items that, if responded to correctly (i.e., MC items) or if responses earned a specified number of points (i.e., CR items), would exemplify preparedness for acceptance into a job-training program.

Table 71 presents a summary of the number of items presented to each panel and the number of items for which panelists expressed 100% agreement and at least 75% agreement that the item/score point would be at least OK to demonstrate preparedness. The average scale value for items selected as at least OK appears parenthetically within the table as well.

Table 71. Pharmacy Technician Operational Workshop: Exemplar Item Summary

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)
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)

Pharmacy Technician Operational Workshop Process Evaluation Results

Following this presentation of the process evaluation results, reactions to the consequences data are summarized. The process evaluation questionnaires for this workshop are presented in their entirety in Appendix S; along with the questions, the appendix shows the frequency of responses per Likert-scale category, and the average response.

Understanding of Tasks. Across all panels, when panelists were asked to respond to the following statement: *"My understanding of the tasks I was to accomplish during each round was . . ."* the average response was 4.1 or higher, which corresponds to a verbal description of at least "Adequate." Average results by panel can be found in Table 72.

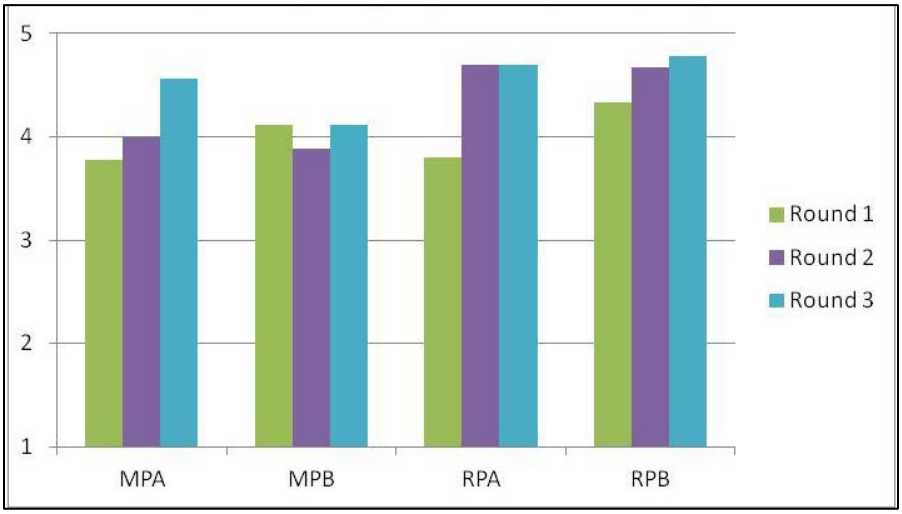
Table 72. Pharmacy Technician Operational Workshop: Summary of Selected Evaluation Items by Panel

Evaluation Item	Average Response by Panel			
	Pharmacy Technician			
	Mathematics		*Reading*	
	MCA	**MCB**	**RCA**	**RCB**
My understanding of the tasks I was to accomplish during each round was . . . (1 = Totally Inadequate; 5 = Totally Adequate)	4.5	4.1	4.4	4.6
The most accurate description of my level of confidence in the cut score recommendations I provided was . . . (1 = Not at All Confident; 5 = Very Confident)	4.6	4.0	4.2	4.7
I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark. (1 = Totally Disagree; 5 = Totally Agree)	3.9	3.6	3.8	4.1
The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items. (1 = Totally Disagree; 5 = Totally Agree)	3.4	3.8	3.8	4.0
I felt pressured by others in my group to make my cut score recommendation agree with theirs. (1 = Totally Disagree; 5 = Totally Agree)	1.3	2.3	1.2	1.9
During the standard setting process, I found using CAB to be . . . (1 = Not at All Helpful; 5 = Very Helpful)	4.8	4.6	4.4	4.7

Understanding of the Borderline Performance Description. After each bookmarking round, panelists were asked to respond to the following statement: *"At the time I placed my bookmark, my understanding of the BPD was . . ."* For round 1, the average response was 3.8 or higher, which corresponds to a verbal description of at least "Somewhat Adequate." Across all panels, the average reported understanding of the BPD steadily increased or remained the same between rounds, with the lowest average being 4.1 after the third round of bookmarking. Average results by panel are shown in Figure 31.

Figure 31. Pharmacy Technician Operational Workshop: At the time I placed my bookmark, my understanding of the BPD was . . .

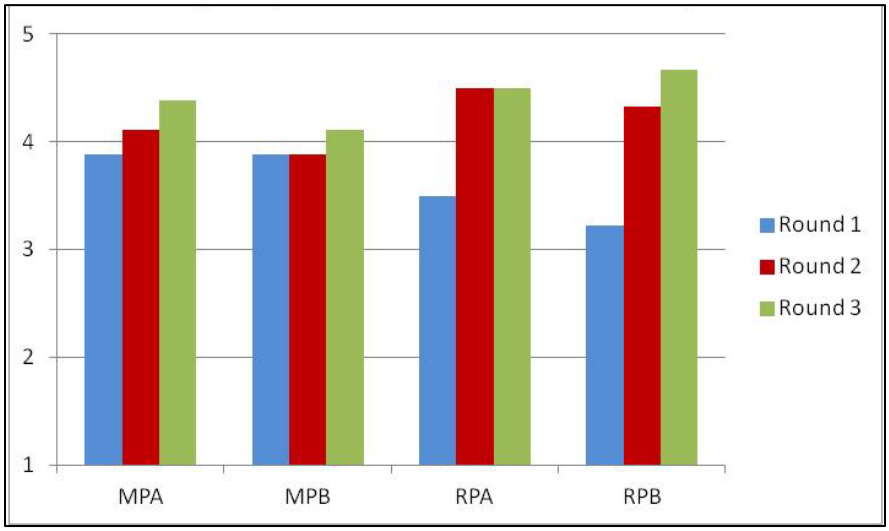
(1 = Totally Inadequate; 5 = Totally Adequate)



Comfort and Confidence. Several statements were developed to solicit how comfortable and/or confident panelists were in setting their cut scores or with components of the standard-setting process. After each bookmarking round, panelists were asked to respond to the following statement: "*The most accurate description of my level of confidence in my bookmark placement is . . .*" For round 1, the average response was 3.2 or higher, which corresponds to a verbal description of at least "Somewhat Confident." Across all panels, the average reported comfort level steadily increased or remained the same across bookmarking rounds, with the lowest average after the third round being 4.1. Average results by panel are shown in Figure 32.

Figure 32. Pharmacy Technician Operational Workshop: The most accurate description of my level of confidence in my bookmark placement is . . .

(1 = Not at All Confident; 5 = Very Confident)

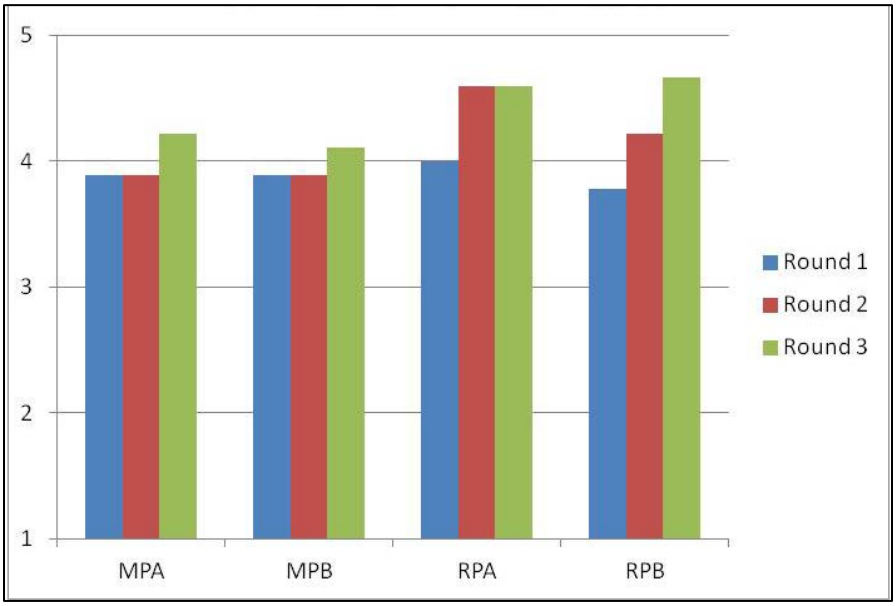


As a follow-up to this statement, panelists were asked to respond to the following statement: *"The most accurate description of my level of confidence in the cut score recommendations I provided was . . ."* Across all panels, the average response was at least 4.0, which corresponds to a verbal description of at least "Confident." Average results by panel can be found in Table 72 (p. 162).

Additionally, after each round of bookmarking, panelists were asked to respond to the following statement: *"I believe my cut score is consistent with the BPD."* After round 1, all panels had an average response of at least 3.8, which corresponds to a verbal description of at least "Somewhat Agree." Agreement with this statement slightly increased or remained the same with each successive bookmarking round, with the lowest average in the final round of bookmarking being 4.1. Average results by panel are shown in Figure 33.

Figure 33. Pharmacy Technician Operational Workshop: I believe my cut score is consistent with the BPD

(1 = Totally Disagree; 5 = Totally Agree)



Specific to the standard-setting method are the reliance on an understanding of how to use a response probability in placing a bookmark and a general acceptance that items are ordered by relative difficulty. Thus, each of these assumptions was evaluated. Panelists were asked to respond to the following statement: *"I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark."* Across all panels, the average response was at least 3.6, which corresponds to a verbal description of at least "Somewhat Agree." Average results by panel can be found in Table 72 (p. 162).

Panelists were also asked to respond to the following statement: *"The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items."* Average responses across panels varied slightly between 3.4 and 4.0, which corresponds to verbal descriptions between "Somewhat Agree" and "Agree." Average results by panel can be found in Table 72 (p. 162).

Independence of Judgment. Across all panels, panelists were asked to respond to the following statement: *"I felt pressured by others in my group to make my cut score recommendation agree with theirs."* The average response was no higher than 2.3, which corresponds to a verbal description of "Disagree" or "Totally Disagree." Average results by panel can be found in Table 72 (p. 162).

Helpfulness of Software. As the CAB software system was implemented to aid the standard-setting process, panelists were asked to respond to the following statement: *"During the standard setting process, I found using CAB to be . . ."* Across all panels, the lowest average response was 4.4, which corresponds to a verbal description of "Helpful" or higher. Average results by panel can be found in Table 72 (p. 162).

Panelists were asked to review both their final round 3 cut scores and the associated impact data. They were asked to respond to the following questions with a "Yes" or "No" response:

- *Does the [impact data] percentage reflect your expectations about the proportion of students whose NAEP score would indicate at least minimal preparedness for placement?*
- *Would you change the cut score recommended by your panel to the Governing Board if you could?*

Results across panels varied significantly, with 44% to 100% of panelists (depending on which panel they belonged to) responding "Yes" to the first question, and 44% to 100% of panelists responding "No" to the second question. The full set of results can be found in Appendix S.

Operational Session 3

The third operational JSS session was conducted on June 28–July 1, 2011, and included Computer Support Specialist and HVAC workshops.

Panelist recruitment was modified somewhat for this operational session: based on feedback from the first two operational sessions, which suggested that the job-training instructors were not familiar with the mathematics and reading skills that are taught at the high school level, secondary-level teachers of mathematics and reading were recruited to supplement the job-training instructors for the Computer Support Specialist and HVAC workshops. The inclusion of these teachers did appear to contribute to the discussions; however, content facilitators reported that in some workshops, the secondary-level instructors became too influential in establishing the content areas' BPDs.

The JSS process was largely implemented in this session as described in the Judgmental Standard-Setting Process subsection of this report (pp. 39–69), with the process enhancements applied in the first and second operational sessions (pp. 102–106 and 136–138, respectively).

An enhancement to the KSA review that was implemented only for the third operational session was the provision of item descriptions to the panelists. A description was provided for nearly every item in the OIBs. Also, unlike in the earlier workshops, panelists performed the initial KSA review with a partner, rather than individually. For the initial review using the item descriptions, the following instructions were given to the panelists:

> "A description for each item will be provided to you to help you with your task. You may use the provided description as the KSAs, you may paraphrase them, or you may decide to develop your own KSAs. You and your partner should share your thoughts and suggestions

regarding the KSAs for each item, but you do not have to agree on the KSA you record for each item." (p. 18 of the Facilitator Handbook, Appendix K)

In addition, based on feedback from panelists in prior JSS sessions, Computer Support Specialist and HVAC panelists participated in a special study to research the use of a different item map configuration in the bookmarking process. This special study is described later in this report (pp. 219–224).

The mathematics and reading BPDs that resulted from the first two operational sessions were used as the starting point for the Operational Session 3 BPDs. The content facilitators consolidated the earlier BPDs and the Operational Session 3 panelists' responses to the online Content Objectives form to develop preliminary BPD drafts to share with the Operational Session 3 panelists. The preliminary mathematics and reading BPDs developed by the content facilitators in advance of the standard-setting session and the final BPD versions agreed upon by the mathematics and reading pairs of replicate panels for each postsecondary area are provided in Appendix O.

The abbreviations used to describe panels in the third operational session are displayed in Table 73.

Table 73. Operational Session 3: Abbreviations Used to Describe Panels

Abbreviation	Content Area	Occupation	Panel
MSA	Mathematics	Computer Support Specialist	A
MSB			B
RSA	Reading	Computer Support Specialist	A
RSB			B
MHA	Mathematics	HVAC	A
MHB			B
RHA	Reading	HVAC	A
RHB			B

Computer Support Specialist Operational Workshop

The following subsections describe the Computer Support Specialist operational workshop panelists, numerical results, and process evaluation results.

Computer Support Specialist Operational Workshop Panelists

Twenty-seven Computer Support Specialist job-training programs[17] and eight secondary-level institutions were represented in the pool of Computer Support Specialist panelists. From these programs and institutions, 40 Computer Support Specialist panelists were recruited. It was observed during the recruitment process that this occupation appeared to be the most diverse of the five, with an unexpectedly large number of areas of focus within this occupation and a substantial degree of overlap between many of these areas. The process by which this diversity was addressed is described in the Panelist Recruitment Plan subsection of this report (pp. 26–38).

The Design Document called for a total of 40 panelists for this workshop: 20 in mathematics and 20 in reading, with each group divided into two replicate panels of 10 panelists. This target was achieved.

Table 74 displays the distribution of panelists by type of institution, and Table 75 displays the distribution of panelists by demographic characteristics. As intended, eight Computer Support Specialist panelists (20%) came from secondary-level institutions, with 22 panelists (55%) representing public two-year community or technical colleges, as shown in Table 74; relatively few panelists (8%) came from private institutions, while seven panelists (18%) taught in programs housed within four-year public institutions. As shown in Table 75, a slight majority of

[17]Two programs each provided two panelists, and one program provided four panelists. Panelists coming from the same program were assigned to different content areas and/or panels.

Computer Support Specialist panelists (58%) were men, with 17 (42%) women. Some diversity in race/ethnicity was achieved on each Computer Support Specialist panel, with 12 panelists (30%)—three on each panel—reporting themselves to be non–White/Caucasian.

Table 74. Computer Support Specialist Operational Workshop: Panelist Distribution by Institution Type

Postsecondary Area	Content Area	Panel	Type of Institution				Total Panelists
			4-Year Institution	2-Year Public (Community/ Technical)	2-Year Private	Secondary	
Computer Support Specialist	Math	A	3 (30%)	4 (40%)	1 (10%)	2 (20%)	10 (100%)
		B	3 (30%)	5 (50%)	0 (0%)	2 (20%)	10 (100%)
	Reading	A	1 (10%)	6 (60%)	1 (10%)	2 (20%)	10 (100%)
		B	0 (0%)	7 (70%)	1 (10%)	2 (20%)	10 (100%)
Computer Support Specialist Totals (N = 40)			7 (18%)	22 (55%)	3 (8%)	8 (20%)	40 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

Table 75. Computer Support Specialist Operational Workshop: Panelist Distribution by Demographic Characteristics

Postsecondary Area	Content Area	Panel	Gender		Race/Ethnicity			Total Panelists
			Female	Male	White/ Caucasian	Non-White/ Caucasian	Not Specified	
Computer Support Specialist	Math	A	4 (40%)	6 (60%)	7 (70%)	3 (30%)	0 (0%)	10 (100%)
		B	3 (30%)	7 (70%)	7 (70%)	3 (30%)	0 (0%)	10 (100%)
	Reading	A	5 (50%)	5 (50%)	7 (70%)	3 (30%)	0 (0%)	10 (100%)
		B	5 (50%)	5 (50%)	7 (70%)	3 (30%)	0 (0%)	10 (100%)
Computer Support Specialist Totals (N = 40)			17 (43%)	23 (58%)	28 (70%)	12 (30%)	0 (0%)	40 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

As shown in Table 76, all four census regions were represented on the Computer Support Specialist panels, with the lowest percentage of panelists (8%) representing the Northeast and the

highest percentage (48%) representing the South. On the reading panels, panelists represented all four geographic regions. However, neither of the mathematics panels had panelists from the Northeast.

Table 76. Computer Support Specialist Operational Workshop: Geographic Distribution of Panelists

Postsecondary Area	Content Area	Panel	Geographic Region*				Total Panelists
			Northeast	South	Midwest	West	
Computer Support Specialist	Math	A	0 (0%)	6 (60%)	2 (20%)	2 (20%)	10 (100%)
		B	0 (0%)	6 (60%)	2 (20%)	2 (20%)	10 (100%)
	Reading	A	1 (10%)	4 (40%)	3 (30%)	2 (20%)	10 (100%)
		B	2 (20%)	4 (40%)	2 (20%)	2 (20%)	10 (100%)
Computer Support Specialist Totals (N = 40)			3 (8%)	20 (50%)	9 (23%)	8 (20%)	40 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.
*Based upon U.S. Census Bureau census regions.

Information about student populations served by the programs represented by the Computer Support Specialist panelists is shown in Table 77.

Table 77. Computer Support Specialist Operational Workshop: Student Populations Served by Panelists

Postsecondary Area	Content Area	Panel	Predominant Student Population Served by Program			Total Job-Training Panelists*
			Students Coming Directly from High School	Students Returning to School after Absence	Not Specified	
Computer Support Specialist	Math	A	4 (50%)	4 (50%)	0 (0%)	8 (100%)
		B	7 (88%)	1 (13%)	0 (0%)	8 (100%)
	Reading	A	4 (50%)	3 (38%)	1 (13%)	8 (100%)
		B	5 (63%)	3 (38%)	0 (0%)	8 (100%)
Computer Support Specialist Totals (N = 32)			20 (63%)	11 (34%)	1 (3%)	32 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.
*Two panelists on each panel came from secondary-level schools and are not reflected in this table.

The majority of Computer Support Specialist panelists who provided this information (63%) reported coming from job-training programs that predominantly served students coming directly from high school.

Computer Support Specialist Operational Workshop Numerical Results

The following tables display standard-setting results for the Computer Support Specialist operational workshop.

When round 1 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The round 1 results are presented in Table 78.

Table 78. Computer Support Specialist Operational Workshop: Round 1 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	173	7.6	7.2	14.4	29.9
	B	190	12.6	12.7	16.9	14.8
Reading	A	293	8.6	8.6	16.5	48.4
	B	305	5.5	6.3	12.6	35.5

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 2 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 2 judgments are presented in Table 79.

Table 79. Computer Support Specialist Operational Workshop: Round 2 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	172	2.4	2.4	4.7	31.0
	B	185	4.9	4.1	9.3	18.7
Reading	A	294	5.4	5.0	9.1	47.3
	B	308	1.0	1.0	4.9	32.5

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 3 bookmarks were placed, the CAB once again calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 3 decisions are presented in Table 80.

Table 80. Computer Support Specialist Operational Workshop: Round 3 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	165	3.3	3.3	6.1	38.1
	B	185	3.8	3.6	7.3	18.7
Reading	A	292	3.2	3.2	6.5	49.6
	B	307	1.8	2.2	3.6	33.5

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

Figure 34 summarizes the MAD results across rounds for the Computer Support Specialist operational panels. While three of the four Computer Support Specialist panel results were as expected (i.e., decreasing variability of panelist cut scores between rounds), mathematics Panel A did not meet expectations between rounds 2 and 3. This suggests that panelists within

mathematics Panel A reacted strongly and differently to the impact data. Reasons for these reactions are not readily available.

Figure 34. Computer Support Specialist Operational Workshop: Mean Absolute Deviation (MAD) of Cut Scores by Panel

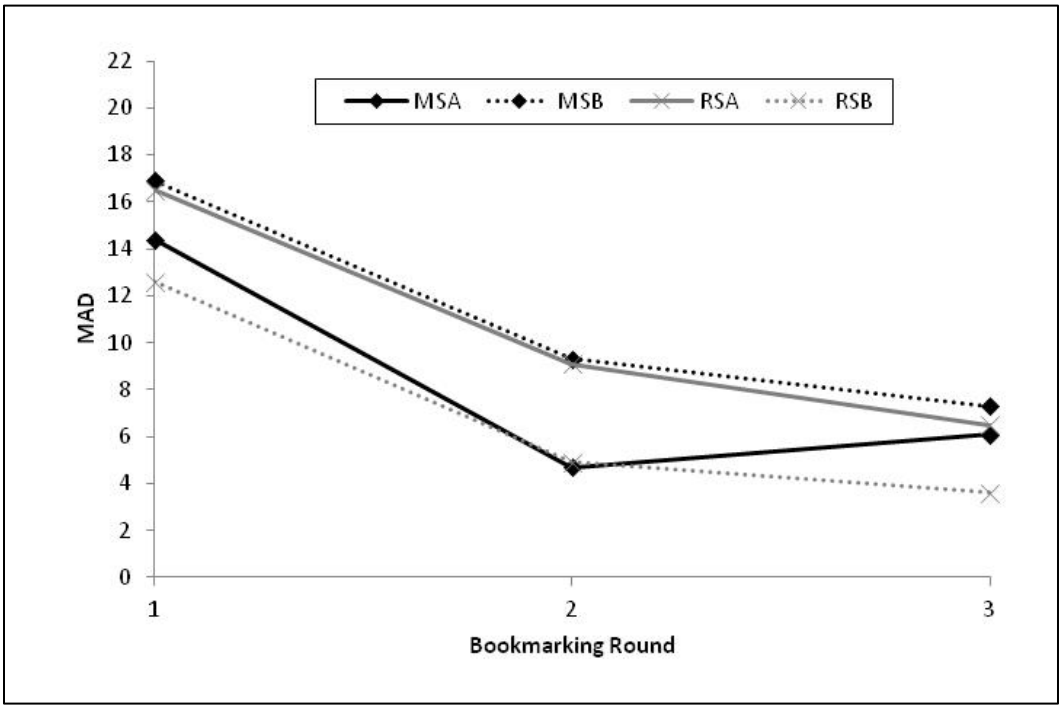


Table 81 summarizes the changes in cut scores made by individual panelists between rounds, while Table 82 presents a comparison of median and mean cut scores for each panel. The expectation is to see the same number or fewer panelists change their cut scores between rounds 2 and 3 than between rounds 1 and 2. This pattern is observed in Table 81. Differences between mean and median cut scores result from the presence of outliers (i.e., panelists who set cut scores significantly higher or lower than their peers). The largest absolute difference between the mean and median cut scores in Table 82 occurs for reading Panel B in round 1 (5.4), where the direction of the difference would result in a lower cut score being set if the mean instead of the median were used. In the third round, the effect of outliers on mean cut scores is negligible.

Table 81. Computer Support Specialist Operational Workshop: Round-to-Round Cut Score Changes by Panel

Panel	Round	Increased N (%)	No Change N (%)	Decreased N (%)
MSA	R1–R2	4 (40.0%)	0 (0.0%)	6 (60.0%)
	R2–R3	1 (10.0%)	2 (20.0%)	7 (70.0%)
MSB	R1–R2	6 (60.0%)	2 (20.0%)	2 (20.0%)
	R2–R3	1 (10.0%)	3 (30.0%)	6 (60.0%)
RSA	R1–R2	3 (30.0%)	2 (20.0%)	5 (50.0%)
	R2–R3	0 (0.0%)	8 (80.0%)	2 (20.0%)
RSB	R1–R2	7 (70.0%)	0 (0.0%)	3 (30.0%)
	R2–R3	1 (10.0%)	5 (50.0%)	4 (40.0%)

Table 82. Computer Support Specialist Operational Workshop: Comparison of Cut Scores Based on Medians and Means

Panel	Round 1			Round 2			Round 3		
	Median	Mean	Median–Mean	Median	Mean	Median–Mean	Median	Mean	Median–Mean
MSA	172.5	177.2	**-4.7**	172.0	172.3	**-0.3**	164.5	163.9	**0.6**
MSB	189.5	186.7	**2.8**	185.0	189.7	**-4.7**	184.5	187.1	**-2.6**
RSA	292.5	295.1	**-2.6**	294.0	294.5	**-0.5**	292.0	291.9	**0.1**
RSB	305.0	299.6	**5.4**	308.0	308.7	**-0.7**	307.0	304.4	**2.6**

Computer Support Specialist Operational Workshop Exemplar Item Ratings. Following the rounds of bookmarking, panelists were asked to identify items that, if responded to correctly (i.e., MC items) or if responses earned a specified number of points (i.e., CR items), would exemplify preparedness for acceptance into a job-training program.

Table 83 presents a summary of the number of items presented to each panel and the number of items for which panelists expressed 100% agreement and at least 75% agreement that the item/score point would be at least OK to demonstrate preparedness. The average scale value for items selected as at least OK appears parenthetically within the table as well.

Table 83. Computer Support Specialist Operational Workshop: Exemplar Item Summary

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)
MSA	10	30	165	1 (172)	11 (189)
MSB	10	23	185	2 (206)	7 (228)
RSA	10	16	292	3 (320)	11 (347)
RSB	10	13	307	5 (327)	9 (337)

Computer Support Specialist Operational Workshop Process Evaluation Results

Following this presentation of the process evaluation results, reactions to the consequences data are summarized. The process evaluation questionnaires for this workshop are presented in their entirety in Appendix T; along with the questions, the appendix shows the frequency of responses per Likert-scale category, and the average response.

Understanding of Tasks. Across all panels, panelists were asked to respond to the following statement: *"My understanding of the tasks I was to accomplish during each round was . . ."* The average response was 4.4 or higher, which corresponds to a verbal description of "Adequate" or "Totally Adequate." Average results by panel can be found in Table 84.

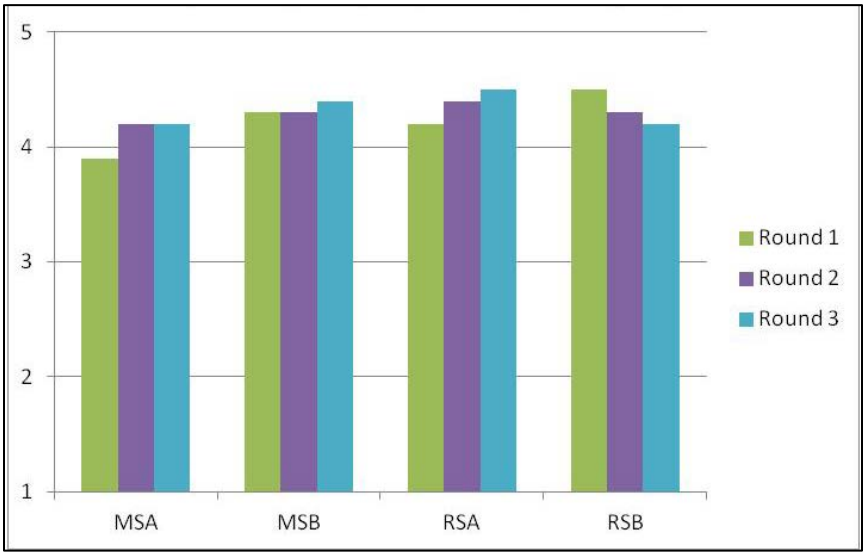
Table 84. Computer Support Specialist Operational Workshop: Summary of Selected Evaluation Items by Panel

Evaluation Item	Average Response by Panel			
	Computer Support Specialist			
	Mathematics		*Reading*	
	MCA	**MCB**	**RCA**	**RCB**
My understanding of the tasks I was to accomplish during each round was . . . (1 = Totally Inadequate; 5 = Totally Adequate)	4.4	4.7	4.5	4.4
The most accurate description of my level of confidence in the cut score recommendations I provided was . . . (1 = Not at All Confident; 5 = Very Confident)	3.8	4.6	4.4	4.6
I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark. (1 = Totally Disagree; 5 = Totally Agree)	4.0	3.9	3.7	4.1
The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items. (1 = Totally Disagree; 5 = Totally Agree)	3.3	3.9	3.5	4.0
I felt pressured by others in my group to make my cut score recommendation agree with theirs. (1 = Totally Disagree; 5 = Totally Agree)	1.7	1.3	1.3	1.1
During the standard setting process, I found using CAB to be . . . (1 = Not at All Helpful; 5 = Very Helpful)	4.3	4.4	4.7	4.9

Understanding of the Borderline Performance Description. After each bookmarking round, panelists were asked to respond to the following statement: *"At the time I placed my bookmark, my understanding of the BPD was . . ."* For round 1, the average response was 3.9 or higher, which corresponds to a verbal description of at least "Somewhat Adequate." Across three of the four panels, the average reported understanding of the BPD increased between rounds 1 and 3, but only slightly, with the lowest average being 4.2 after the third round of bookmarking. Panel RSB experienced slightly decreasing perceptions of adequacy across rounds, with averages of 4.5, 4.3, and 4.2, respectively, but not enough to correspond to a different verbal description. Average results by panel are shown in Figure 35.

Figure 35. Computer Support Specialist Operational Workshop: At the time I placed my bookmark, my understanding of the BPD was . . .

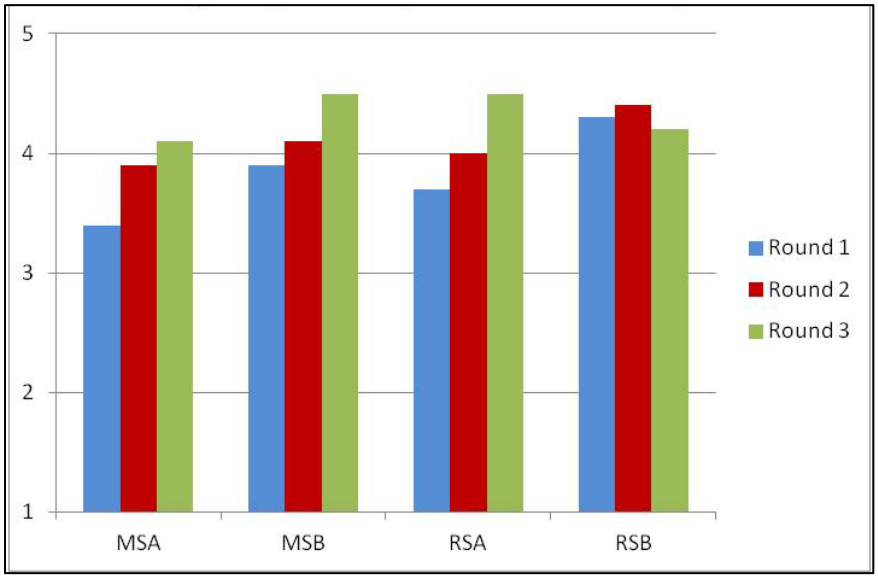
(1 = Totally Inadequate; 5 = Totally Adequate)



Comfort and Confidence. Several statements were developed to solicit how comfortable and/or confident panelists were in setting their cut scores or with components of the standard-setting process. After each bookmarking round, panelists were asked to respond to the following statement: *"The most accurate description of my level of confidence in my bookmark placement is . . ."* For round 1, the average response was 3.4 or higher, which corresponds to a verbal description of at least "Somewhat Confident." Across most panels, the average reported comfort level steadily increased between rounds, with the lowest average being 4.1 after the third round of bookmarking. Reading Panel B's level of confidence changed very little between rounds, with averages of 4.3, 4.4, and 4.2, respectively. Average results by panel are shown in Figure 36.

Figure 36. Computer Support Specialist Operational Workshop: The most accurate description of my level of confidence in my bookmark placement is . . .

(1 = Not at All Confident; 5 = Very Confident)

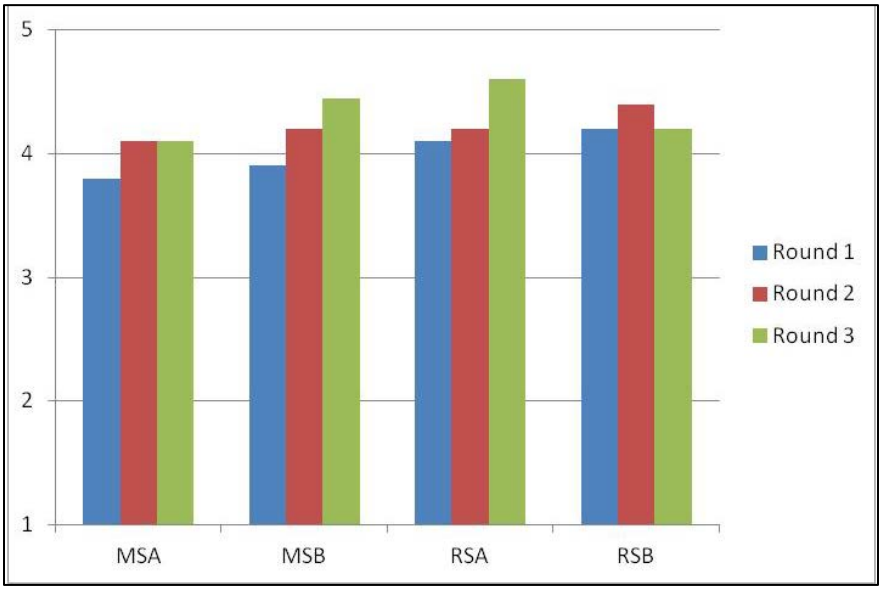


As a follow-up to this statement, panelists were asked to respond to the following statement: *"The most accurate description of my level of confidence in the cut score recommendations I provided was . . ."* Across all panels, the average response was at least 3.8, which corresponds to a verbal description of at least "Somewhat Confident." Average results by panel can be found in Table 84 (p. 177).

Additionally, after each round of bookmarking, panelists were asked to respond to the following statement: *"I believe my cut score is consistent with the BPD."* After round 1, all panels had an average response of at least 3.8, which corresponds to a verbal description of at least "Somewhat Agree." Further, agreement increased or remained the same for all panels by the end of the third bookmarking round, with the lowest average being 4.1. Average results by panel are shown in Figure 37.

Figure 37. Computer Support Specialist Operational Workshop: I believe my cut score is consistent with the BPD

(1 = Totally Disagree; 5 = Totally Agree)



Specific to the standard-setting method are the reliance on an understanding of how to use a response probability in placing a bookmark and a general acceptance that items are ordered by relative difficulty. Thus, each of these assumptions was evaluated.

In order to evaluate these assumptions, panelists were asked to respond to the following statement: *"I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark."* Across all panels, the average response was at least 3.7, which corresponds to a verbal description of at least "Somewhat Agree." Average results by panel can be found in Table 84 (p. 177).

Panelists were also asked to respond to the following statement: *"The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items."* Average responses to this statement across panels only varied between 3.3 and 4.0, which corresponds to verbal

descriptions between “Somewhat Agree” and “Agree.” Average results by panel can be found in Table 84 (p. 177).

Independence of Judgment. Across all panels, panelists were asked to respond to the following statement: *“I felt pressured by others in my group to make my cut score recommendation agree with theirs.”* The average response was no higher than 1.7, which corresponds to a verbal description of “Disagree” or “Totally Disagree.” Average results by panel can be found in Table 84 (p. 177).

Helpfulness of Software. As the CAB software system was implemented to aid the bookmarking standard-setting process, panelists were asked to respond to the following statement: *“During the standard setting process, I found using CAB to be . . .”* Across all panels, the lowest average response was 4.3, which corresponds to a verbal description of “Helpful” or higher. Average results by panel can be found in Table 84 (p. 177).

Panelists were asked to review both their final round 3 cut scores and the associated impact data. They were asked to respond to the following questions with a “Yes” or “No” response:

- *Does the [impact data] percentage reflect your expectations about the proportion of students whose NAEP score would indicate at least minimal preparedness for placement?*
- *Would you change the cut score recommended by your panel to the Governing Board if you could?*

Eighty percent to 100% of panelists (depending on which panel they belonged to) responded “Yes” to the first question, and 70% to 90% of panelists responded “No” to the second question,

indicating that most panelists were satisfied with the final cut scores. The full set of results can be found in Appendix T.

HVAC Operational Workshop

The following subsections describe the HVAC operational workshop panelists, numerical results, and process evaluation results.

HVAC Operational Workshop Panelists

Based on feedback from the first two operational sessions, secondary-level teachers of mathematics and reading were recruited to supplement the job-training instructors for the HVAC workshop. Twenty-nine HVAC job-training programs[18] and seven secondary-level schools[19] were represented in the pool of HVAC panelists. From these job-training programs and secondary-level schools, 38 HVAC panelists were recruited to participate in the third operational session. It should be noted that four reading panelists came from one state (California), with each California panelist assigned to a different table group. With the exception of the Pharmacy Technician reading panels, this is the only instance of four panelists within the same content area and postsecondary area representing the same state.

The Design Document called for a total of 40 panelists for this workshop: 20 in mathematics and 20 in reading, with each group divided into two replicate panels of 10 panelists. Due to last-minute schedule conflicts,[20] two panelists dropped from the study too late to find replacements, resulting in a total of 19 mathematics panelists and 19 reading panelists.

Table 85 displays the distribution of panelists by type of institution, and Table 86 displays the distribution of panelists by demographic characteristics. Eight HVAC panelists (21%) came from

[18]One online network of HVAC instruction provided two instructors; even though they also teach in different brick-and-mortar programs, they were counted as coming from the same program.
[19]One reading instructor and one mathematics instructor came from the same secondary school.
[20]One HVAC reading panelist dropped from the study at the last minute, and one HVAC mathematics panelist removed himself from the study after the second day due to a schedule conflict.

secondary-level institutions, with 24 panelists (63%) representing public two-year community or technical colleges, as shown in Table 85. Relatively few panelists (11%) came from private institutions, while two panelists (5%) taught in programs housed within four-year public institutions. As shown in Table 86, most of the panelists (84%) were men, with all female panelists coming from secondary-level institutions. Little diversity in race/ethnicity was achieved on each HVAC panel, with only four panelists (11%) reporting themselves to be non–White/Caucasian.

Table 85. HVAC Operational Workshop: Panelist Distribution by Institution Type

Postsecondary Area	Content Area	Panel	Type of Institution				Total Panelists
			4-Year Institution	2-Year Public (Community/ Technical)	2-Year Private	Secondary	
HVAC	Math	A	0 (0%)	7 (70%)	1 (10%)	2 (20%)	10 (100%)
		B	0 (0%)	4 (44%)	3 (33%)	2 (22%)	9 (100%)
	Reading	A	0 (0%)	8 (80%)	0 (0%)	2 (20%)	10 (100%)
		B	2 (22%)	5 (56%)	0 (0%)	2 (22%)	9 (100%)
HVAC Totals (N = 38)			2 (5%)	24 (63%)	4 (11%)	8 (21%)	38 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

Table 86. HVAC Operational Workshop: Panelist Distribution by Demographic Characteristics

Postsecondary Area	Content Area	Panel	Gender		Race/Ethnicity			Total Panelists
			Female	Male	White/ Caucasian	Non-White/ Caucasian	Not Specified	
HVAC	Math	A	1 (10%)	9 (90%)	9 (90%)	0 (0%)	1 (10%)	10 (100%)
		B	2 (22%)	7 (78%)	8 (89%)	1 (11%)	0 (0%)	9 (100%)
	Reading	A	2 (20%)	8 (80%)	6 (60%)	3 (30%)	1 (10%)	10 (100%)
		B	1 (11%)	8 (89%)	8 (89%)	0 (0%)	1 (11%)	9 (100%)
HVAC Totals (N = 38)			6 (16%)	32 (84%)	31 (82%)	4 (11%)	3 (8%)	38 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.

As shown in Table 87, all four census regions were represented on the HVAC panels, with the lowest percentage of panelists (11%) representing the Northeast and the highest percentage (37%) representing the South. On half of the panels, panelists represented all four geographic regions. On one mathematics panel and one reading panel, however, there were panelists from only three regions.

Table 87. HVAC Operational Workshop: Geographic Distribution of Panelists

Postsecondary Area	Content Area	Panel	Geographic Region*				Total Panelists
			Northeast	South	Midwest	West	
HVAC	Math	A	1 (10%)	2 (20%)	4 (40%)	3 (30%)	10 (100%)
		B	0 (0%)	4 (44%)	2 (22%)	3 (33%)	9 (100%)
	Reading	A	1 (10%)	4 (40%)	2 (20%)	3 (30%)	10 (100%)
		B	2 (22%)	4 (44%)	0 (0%)	3 (33%)	9 (100%)
HVAC Totals (N = 38)			4 (11%)	14 (37%)	8 (21%)	12 (32%)	38 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.
*Based upon U.S. Census Bureau census regions.

Information about student populations served by the programs represented by the HVAC operational panelists is shown in Table 88.

Table 88. HVAC Operational Workshop: Student Populations Served by Panelists

Postsecondary Area	Content Area	Panel	Predominant Student Population Served by Program			Total Job-Training Panelists*
			Students Coming Directly from High School	Students Returning to School after Absence	Not Specified	
HVAC	Math	A	2 (25%)	6 (75%)	0 (0%)	8 (100%)
		B	1 (14%)	4 (57%)	2 (29%)	7 (100%)
	Reading	A	7 (88%)	1 (13%)	0 (0%)	8 (100%)
		B	3 (43%)	4 (57%)	0 (0%)	7 (100%)
HVAC Totals (N = 30)			13 (43%)	15 (50%)	2 (7%)	30 (100%)

Note. Within this table, percentages are calculated by row to facilitate comparison of panelist demographics across replicate panels.
*Two panelists on each panel came from secondary-level schools and are not reflected in this table.

Half of the HVAC panelists who provided this information (50%) reported coming from job-training programs that predominantly served students returning to school after a year or more of absence.

HVAC Operational Workshop Numerical Results

The following tables display standard-setting results for the HVAC operational workshop.

When round 1 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The round 1 results are presented in Table 89.

Table 89. HVAC Operational Workshop: Round 1 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	180	11.3	11.3	21.8	23.1
	B	177	11.9	11.1	17.4	25.8
Reading	A	288	3.0	3.1	7.4	54.1
	B	283	9.0	7.9	14.2	59.2

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 2 bookmarks were placed, the CAB calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 2 judgments are presented in Table 90.

Table 90. HVAC Operational Workshop: Round 2 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	185	7.4	7.1	11.1	18.7
	B	172	2.5	2.7	5.7	31.0
Reading	A	289	1.8	1.6	5.2	52.9
	B	292	2.6	2.6	3.9	49.6

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.
[b] Empirical-based Standard Error.
[c] Bootstrap Standard Error.
[d] Mean Absolute Deviation.
[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

When round 3 bookmarks were placed, the CAB once again calculated the median cut scores for each panel and the associated impact data. The results of the panels' round 3 decisions are presented in Table 91.

Table 91. HVAC Operational Workshop: Round 3 Results

Content Area	Panel	Median NAEP Scale Cut Score[a]	Standard Error		MAD[d]	Percent of Students[e]
			EmpSE[b]	*BootSE*[c]		
Mathematics	A	177	6.6	5.9	10.9	25.8
	B	172	2.8	2.9	7.2	31.0
Reading	A	289	1.8	1.6	5.2	52.9
	B	292	2.3	2.4	2.9	49.6

[a] The mathematics scale ranges from 1 to 300 and the reading scale ranges from 1 to 500.

[b] Empirical-based Standard Error.

[c] Bootstrap Standard Error.

[d] Mean Absolute Deviation.

[e] Percent of students expected to perform at the median NAEP-scale cut score or higher.

Figure 38 summarizes the MAD results across rounds for the HVAC operational panels. While MAD results were mostly as expected (i.e., decreasing variability of panelist cut scores between rounds), there is a slight increase in variability among panelist cut scores between round 2 and 3 for mathematics Panel B, which suggests that panelists on this panel reacted differently to the impact data presented. Reasons for these reactions are not readily available.

Figure 38. HVAC Operational Workshop: Mean Absolute Deviation (MAD) of Cut Scores by Panel

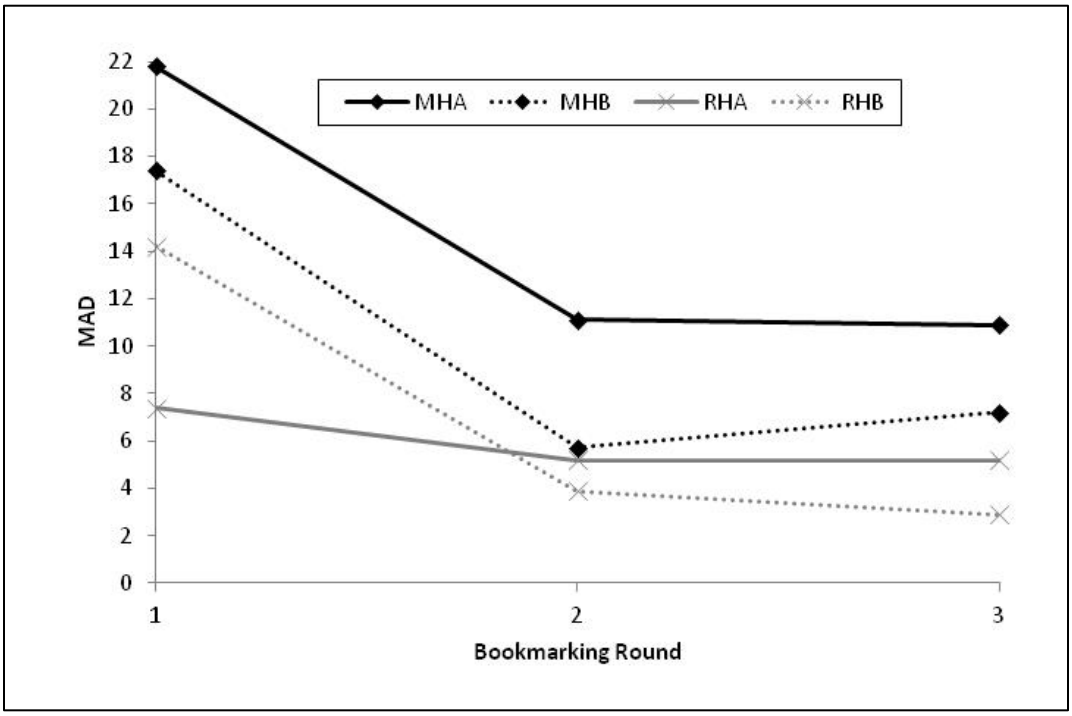


Table 92 summarizes the changes in cut scores made by individual panelists between rounds, while Table 93 presents a comparison of median and mean cut scores for each panel. The expectation was that the same number of panelists or fewer panelists changed their cut scores between rounds 2 and 3 than between rounds 1 and 2. This pattern is observed in Table 92. Differences between mean and median cut scores are due to the presence of outliers (i.e., panelists who set cut scores significantly higher or lower than their peers). The largest absolute difference between the mean and median cut scores in Table 93 occurs for mathematics panel B in round 1 (9.4), where the direction of the difference would result in a higher cut score being set if the mean instead of the median were used. Overall, in the third round, the effect of outliers on mean cut scores is minimal.

Table 92. HVAC Operational Workshop: Round-to-Round Cut Score Changes by Panel

Panel	Round	Increased N (%)	No Change N (%)	Decreased N (%)
MHA	R1–R2	6 (60.0%)	1 (10.0%)	3 (30.0%)
	R2–R3	0 (0.0%)	4 (40.0%)	6 (60.0%)
MHB	R1–R2	2 (22.2%)	0 (0.0%)	7 (77.8%)
	R2–R3	2 (22.2%)	6 (66.7%)	1 (11.1%)
RHA	R1–R2	4 (40.0%)	6 (60.0%)	0 (0.0%)
	R2–R3	0 (0.0%)	10 (100.0%)	0 (0.0%)
RHB	R1–R2	6 (66.7%)	0 (0.0%)	3 (33.3%)
	R2–R3	1 (11.1%)	6 (66.7%)	2 (22.2%)

Table 93. HVAC Operational Workshop: Comparison of Cut Scores Based on Medians and Means

Panel	Round 1			Round 2			Round 3		
	Median	Mean	Median–Mean	Median	Mean	Median–Mean	Median	Mean	Median–Mean
MHA	180.0	178.4	**1.6**	185.0	188.3	**-3.3**	177.0	182.5	**-5.5**
MHB	177.0	186.4	**-9.4**	172.0	173.4	**-1.4**	172.0	175.7	**-3.7**
RHA	288.0	287.6	**0.4**	288.5	290.4	**-1.9**	288.5	290.4	**-1.9**
RHB	283.0	290.3	**-7.3**	292.0	289.9	**2.1**	292.0	290.4	**1.6**

HVAC Operational Workshop Exemplar Item Ratings. Following the rounds of bookmarking, panelists were asked to identify items that, if responded to correctly (i.e., MC items) or if responses earned a specified number of points (i.e., CR items), would exemplify preparedness for acceptance into a job-training program.

Table 94 presents a summary of the number of items presented to each panel and the number of items for which panelists expressed 100% agreement and at least 75% agreement that the item/score point would be at least OK to demonstrate preparedness. The average scale value for items selected as at least OK appears parenthetically within the table as well.

Table 94. HVAC Operational Workshop: Exemplar Item Summary

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)
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)

HVAC Operational Workshop Process Evaluation Results

Following this presentation of the process evaluation results, reactions to the consequences data are summarized. The process evaluation questionnaires for this workshop are presented in their entirety in Appendix U; along with the questions, the appendix shows the frequency of responses per Likert-scale category, and the average response.

Understanding of Tasks. Across all panels, panelists were asked to respond to the following statement: *"My understanding of the tasks I was to accomplish during each round was . . ."* The average response was 4.3 or higher, which corresponds to a verbal description of at least "Adequate." Average results by panel can be found in Table 95.

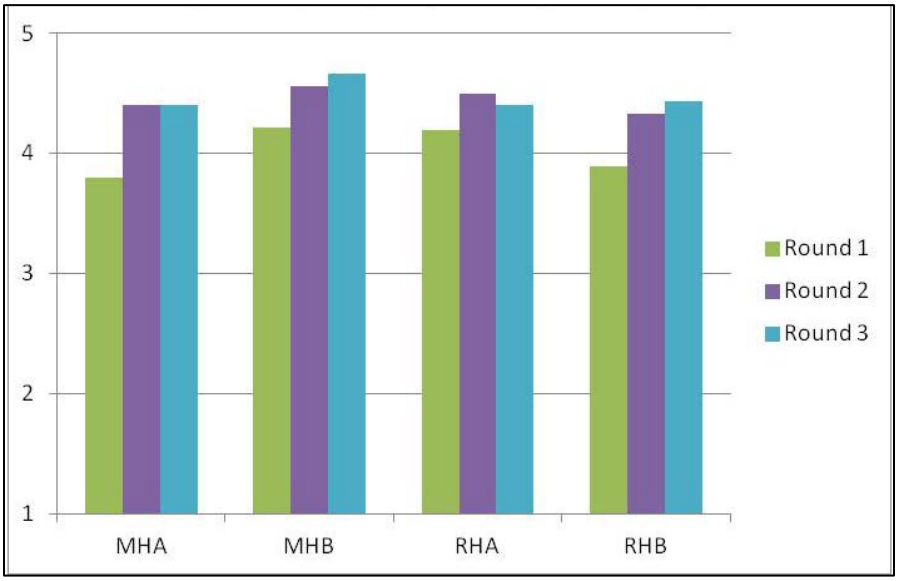
Table 95. HVAC Operational Workshop: Summary of Selected Evaluation Items by Panel

Evaluation Item	Average Response by Panel			
	HVAC			
	Mathematics		*Reading*	
	MCA	MCB	RCA	RCB
My understanding of the tasks I was to accomplish during each round was . . . (1 = Totally Inadequate; 5 = Totally Adequate)	4.3	4.4	4.5	4.3
The most accurate description of my level of confidence in the cut score recommendations I provided was . . . (1 = Not at All Confident; 5 = Very Confident)	4.3	4.4	4.5	4.4
I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark. (1 = Totally Disagree; 5 = Totally Agree)	3.4	4.1	3.8	3.9
The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items. (1 = Totally Disagree; 5 = Totally Agree)	3.1	3.6	3.5	3.2
I felt pressured by others in my group to make my cut score recommendation agree with theirs. (1 = Totally Disagree; 5 = Totally Agree)	1.4	1.8	1.5	1.3
During the standard setting process, I found using CAB to be . . . (1 = Not at All Helpful; 5 = Very Helpful)	4.1	4.6	4.5	4.4

Understanding of the Borderline Performance Description. After each bookmarking round, panelists were asked to respond to the following statement: *"At the time I placed my bookmark, my understanding of the BPD was . . ."* For round 1, the average response was 3.8 or higher, which corresponds to a verbal description of at least "Somewhat Adequate." Across most panels (MHA, MHB, RHB), the average reported understanding of the BPD steadily increased between rounds, with the lowest average being 4.4 after the third round of bookmarking. Average results by panel are shown in Figure 39.

Figure 39. HVAC Operational Workshop: At the time I placed my bookmark, my understanding of the BPD was . . .

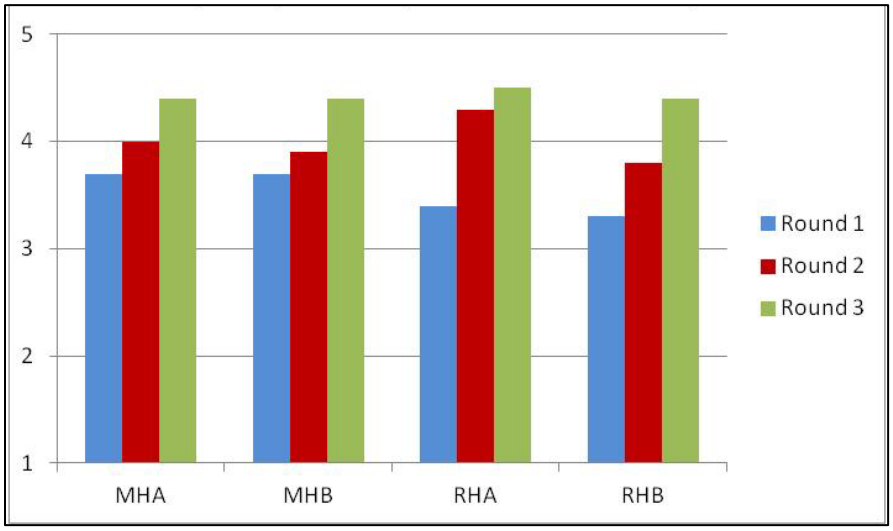
(1 = Totally Inadequate; 5 = Totally Adequate)



Comfort and Confidence. Several statements were developed to solicit how comfortable and/or confident panelists were in setting their cut scores or with components of the standard-setting process. After each bookmarking round, panelists were asked to respond to the following statement: "*The most accurate description of my level of confidence in my bookmark placement is . . .*" For round 1, the average response was 3.3 or higher, which corresponds to a verbal description of at least "Somewhat Confident." Across all panels, the average reported comfort level steadily increased across bookmarking rounds, with the lowest average after the third round being 4.4. Average results by panel are shown in Figure 40.

Figure 40. HVAC Operational Workshop: The most accurate description of my level of confidence in my bookmark placement is . . .

(1 = Not at All Confident; 5 = Very Confident)

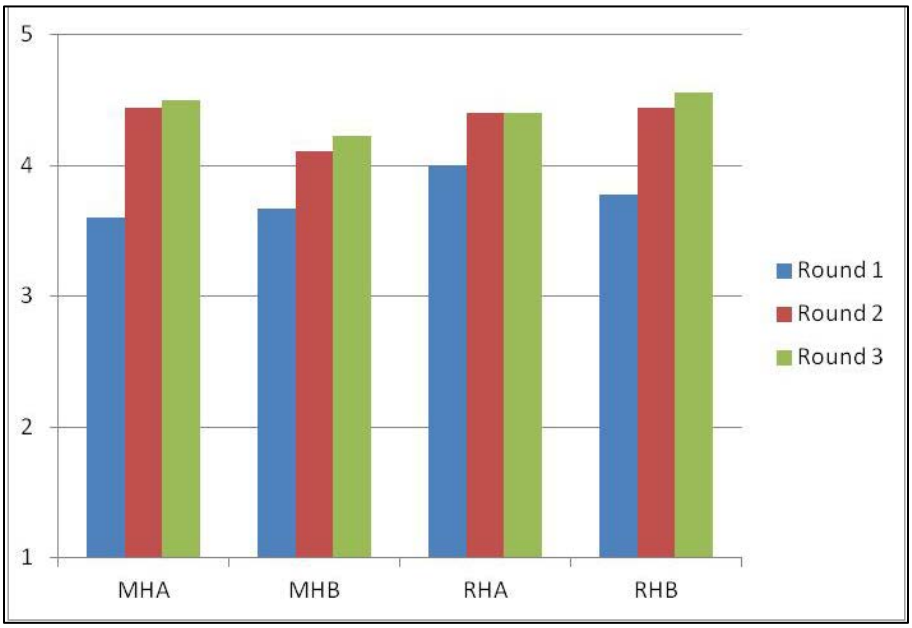


As a follow-up to this statement, panelists were asked to respond to the following statement: *"The most accurate description of my level of confidence in the cut score recommendations I provided was . . ."* Across all panels, the average response was at least 4.3, which corresponds to a verbal description of at least "Confident." Average results by panel can be found in Table 95 (p. 192).

Additionally, after each round of bookmarking, panelists were asked to respond to the following statement: *"I believe my cut score is consistent with the BPD."* After round 1, all panels had an average response of at least 3.6, which corresponds to a verbal description of at least "Somewhat Agree." Agreement with this statement slightly increased or remained the same with each successive bookmarking round, with the lowest average in the final round of bookmarking being 4.2. Average results by panel are shown in Figure 41.

Figure 41. HVAC Operational Workshop: I believe my cut score is consistent with the BPD

(1 = Totally Disagree; 5 = Totally Agree)



Specific to the standard-setting method are the reliance on an understanding of how to use a response probability in placing a bookmark and a general acceptance that items are ordered by relative difficulty. Thus, each of these assumptions was evaluated. Panelists were asked to respond to the following statement: *"I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark."* Across all panels, the average response was at least 3.4, which corresponds to a verbal description of at least "Somewhat Agree." Average results by panel can be found in Table 95 (p. 192).

Panelists were also asked to respond to the following statement: *"The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items."* Average responses across panels varied slightly between 3.1 and 3.6, which corresponds to verbal descriptions between "Somewhat Agree" and "Agree." Average results by panel can be found in Table 95 (p. 192).

Independence of Judgment. Across all panels, panelists were asked to respond to the following statement: *"I felt pressured by others in my group to make my cut score recommendation agree with theirs."* The average response was no higher than 1.8, which corresponds to a verbal description of "Disagree" or "Totally Disagree." Average results by panel can be found in Table 95 (p. 192).

Helpfulness of Software. As the CAB software system was implemented to aid the standard-setting process, panelists were asked to respond to the following statement: *"During the standard setting process, I found using CAB to be . . ."* Across all panels, the lowest average response was 4.1, which corresponds to a verbal description of "Helpful" or higher. Average results by panel can be found in Table 95 (p. 192).

Panelists were asked to review both their final round 3 cut scores and the associated impact data. They were asked to respond to the following questions with a "Yes" or "No" response:

- *Does the [impact data] percentage reflect your expectations about the proportion of students whose NAEP score would indicate at least minimal preparedness for placement?*
- *Would you change the cut score recommended by your panel to the Governing Board if you could?*

Seventy-eight percent to 100% of panelists (depending on which panel they belonged to) responded "Yes" to the first question, and 56% to 100% of panelists responded "No" to the second question, indicating that most panelists were satisfied with the final cut scores. The full set of results can be found in Appendix U.

Summary and Conclusions

This section is a summary of numerical results and process evaluations from the operational implementations of the JSS process. The cut scores presented are all in the NAEP reporting scale for grade 12 mathematics and reading. Table 96 presents the abbreviations used to refer to the 24 panels in the operational JSS meetings.

Table 96. Summary and Conclusions: Panel Identifications

Content Area	Postsecondary Area	Replicate Panel	Panel ID
Mathematics	Automotive Master Technician	A	MAA
		B	MAB
	College-Preparedness	A	MCA
		B	MCB
	Computer Support Specialist	A	MSA
		B	MSB
	HVAC	A	MHA
		B	MHB
	LPN	A	MLA
		B	MLB
	Pharmacy Technician	A	MPA
		B	MPB
Reading	Automotive Master Technician	A	RAA
		B	RAB
	College-Preparedness	A	RCA
		B	RCB
	Computer Support Specialist	A	RSA
		B	RSB
	HVAC	A	RHA
		B	RHB
	LPN	A	RLA
		B	RLB
	Pharmacy Technician	A	RPA
		B	RPB

Cut Scores and Percentages at or Above Cut Scores

Figures 42 and 43 present the cut scores resulting from the three rounds of ratings from each replicate panel during the operational meetings. For mathematics, the final panel cut scores ranged from 165 to 205. The majority of the cut scores fell within the NAEP Proficient range of 176 to 216, with the exception of both operational Automotive Master Technician panels, HVAC Panel A, Pharmacy Technician Panel A, and Computer Support Specialist Panel A, which placed final cut scores within the Basic range. For reading, the final panel cut scores ranged from 288 to 331. All of the cut scores fell within the NAEP Basic and Proficient range of 265 to 346, with seven of the 16 panels setting cut scores within the Basic range of 265 to 302.

Figure 44 presents the percentages of students who would score at or above each cut score based on the 2009 administration of the Grade 12 NAEP in mathematics and reading. Included in the charts for each content area are the percentages at or above the Basic, Proficient, and Advanced cut scores. The data are ordered based on the percentages to show the relative stringency of the cut scores relative to each other as well as to the NAEP achievement level cut scores. It is important to note that all percentages at or above the preparedness cut scores are between the percentages of students at or above the Basic and Advanced cut scores, and that, for each content area, generally half of the panel cut scores are associated with percentages that are higher than the percent of students performing at or above the Proficient level.

The relative stringency of the cut scores set by the replicate panels is also apparent in Figure 44. For example, in mathematics, Pharmacy Technician Panels A and B have replicate results of 28.9 percent and 26.9 percent, respectively, scoring at or above the cut score. In contrast, for reading, LPN Panels A and B have 33.5 percent and 54.1 percent, respectively, scoring at or

above the cut score, which does not seem to indicate replication of results. Considerable variability was observed in the 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. However, further analysis (detailed in the Technical Report) indicates that there are no detectable facilitator effects for the location of the final cut scores.

Figure 42. Cut Scores from Three Rounds of Ratings for Mathematics

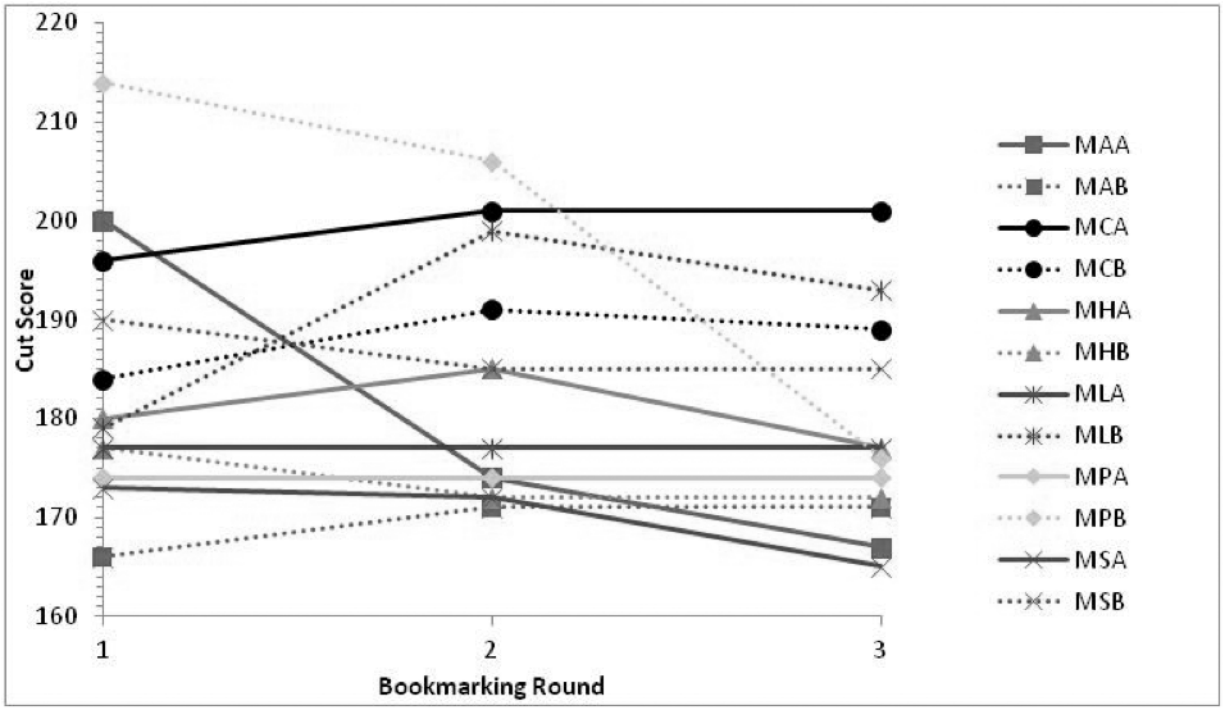


Figure 43. Cut Scores from Three Rounds of Ratings for Reading

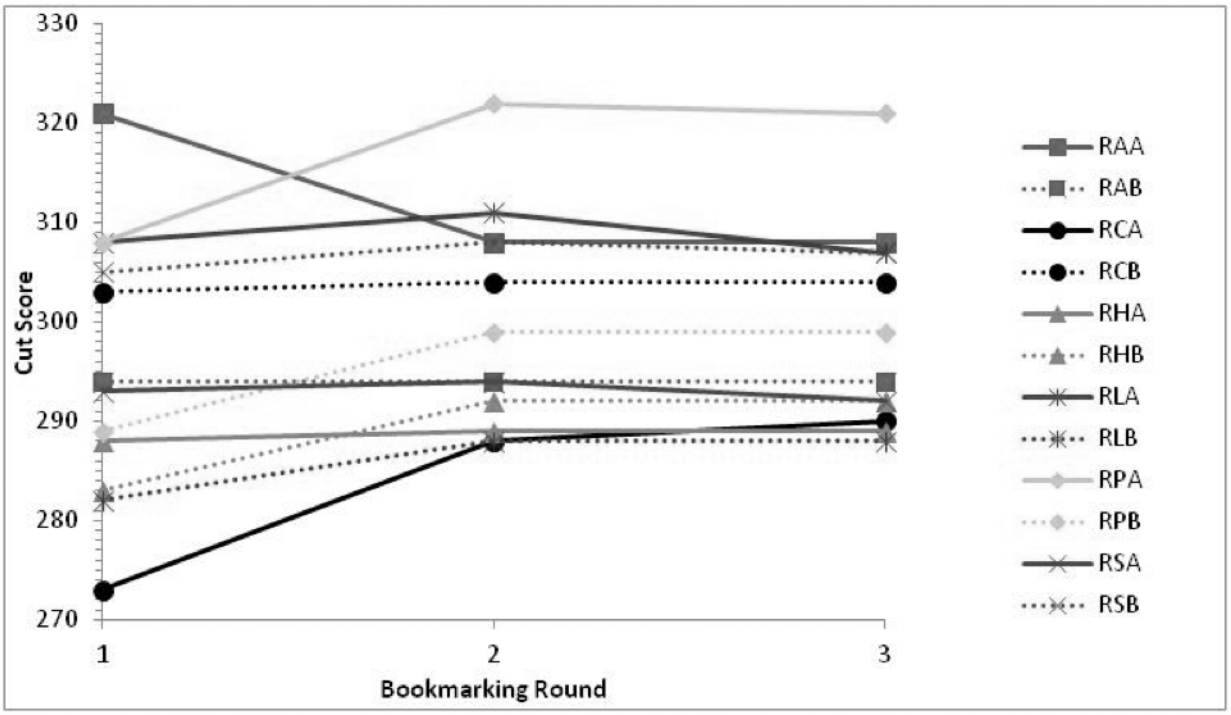
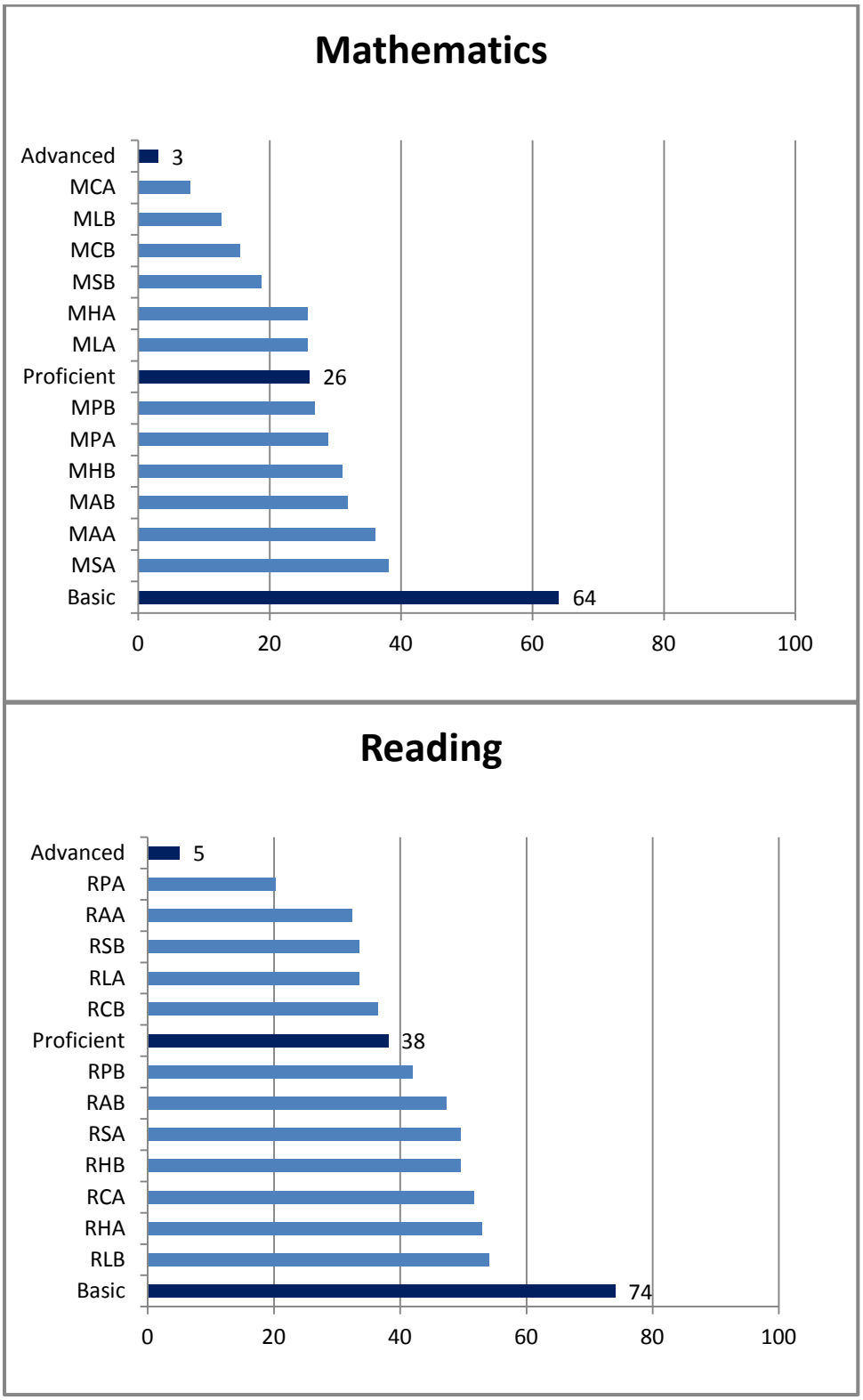


Figure 44. Percents at or Above the Cut Scores from Round 3



Distribution of Cut Scores

An analysis of Mean Absolute Deviation (MAD) was performed. Changes in MAD within panels between rounds showed that, for the first operational session, the MAD for cut scores of most mathematics panels decreased between round 1 and round 3. For the second operational session, there was more discrepancy between panels, with both mathematics Panel Bs showing an increase between rounds 2 and 3. For the third operational session, MAD decreased from round 1 to round 3, while no directional patterns were observed between round 2 and round 3, suggesting that the panels in the third session responded differently to the impact data. For reading, panels in the first operational session showed either a decrease or a moderate increase in MAD of cut scores; in the second and third operational sessions, the panels showed a fairly consistent pattern of decrease from round 1 to round 2. Looking at individual panelists' movement of cut scores between rounds, there were higher percentages of panelists in both mathematics and reading changing cut scores between rounds 1 and 2 than between rounds 2 and 3, as expected. Tables 97 and 98 below show the summary of changes between rounds for the different panels.

Table 97. Summary of Cut Score Changes Across Rounds for Mathematics

Panel	Round	Increased N (%)	No Change N (%)	Decreased N (%)
MAA	R1–R2	0 (0.0%)	1 (14.3%)	6 (85.7%)
	R2–R3	1 (14.3%)	3 (42.9%)	3 (42.9%)
MAB	R1–R2	3 (37.5%)	3 (37.5%)	2 (25.0%)
	R2–R3	0 (0.0%)	6 (75.0%)	2 (25.0%)
MCA	R1–R2	5 (50.0%)	1 (10.0%)	4 (40.0%)
	R2–R3	6 (60.0%)	2 (20.0%)	2 (20.0%)
MCB	R1–R2	5 (50.0%)	1 (10.0%)	4 (40.0%)
	R2–R3	2 (20.0%)	5 (50.0%)	3 (30.0%)
MLA	R1–R2	4 (40.0%)	5 (50.0%)	1 (10.0%)
	R2–R3	0 (0.0%)	7 (70.0%)	3 (30.0%)
MLB	R1–R2	6 (60.0%)	2 (20.0%)	2 (20.0%)
	R2–R3	2 (20.0%)	2 (20.0%)	6 (60.0%)

Panel	Round	Increased N (%)	No Change N (%)	Decreased N (%)
MPA	R1–R2	3 (33.3%)	3 (33.3%)	3 (33.3%)
	R2–R3	0 (0.0%)	4 (44.4%)	5 (55.6%)
MPB	R1–R2	3 (33.3%)	1 (11.1%)	5 (55.6%)
	R2–R3	0 (0.0%)	1 (11.1%)	8 (88.9%)
MHA	R1–R2	6 (60.0%)	1 (10.0%)	3 (30.0%)
	R2–R3	0 (0.0%)	4 (40.0%)	6 (60.0%)
MHB	R1–R2	2 (22.2%)	0 (0.0%)	7 (77.8%)
	R2–R3	2 (22.2%)	6 (66.7%)	1 (11.1%)
MSA	R1–R2	4 (40.0%)	0 (0.0%)	6 (60.0%)
	R2–R3	1 (10.0%)	2 (20.0%)	7 (70.0%)
MSB	R1–R2	6 (60.0%)	2 (20.0%)	2 (20.0%)
	R2–R3	1 (10.0%)	3 (30.0%)	6 (60.0%)

Table 98. Summary of Cut Score Changes Across Rounds for Reading

Panel	Round	Increased N (%)	No Change N (%)	Decreased N (%)
RAA	R1–R2	1 (20.0%)	3 (60.0%)	1 (20.0%)
	R2–R3	1 (20.0%)	1 (20.0%)	3 (60.0%)
RAB	R1–R2	3 (50.0%)	1 (16.7%)	2 (33.3%)
	R2–R3	0 (0.0%)	6 (100.0%)	0 (0.0%)
RCA	R1–R2	8 (80.0%)	0 (0.0%)	2 (20.0%)
	R2–R3	5 (50.0%)	4 (40.0%)	1 (10.0%)
RCB	R1–R2	4 (44.4%)	0 (0.0%)	5 (55.6%)
	R2–R3	0 (0.0%)	8 (88.9%)	1 (11.1%)
RLA	R1–R2	5 (50.0%)	1 (10.0%)	4 (40.0%)
	R2–R3	1 (10.0%)	7 (70.0%)	2 (20.0%)
RLB	R1–R2	4 (40.0%)	2 (20.0%)	4 (40.0%)
	R2–R3	2 (20.0%)	4 (40.0%)	4 (40.0%)

Panel	Round	Increased N (%)	No Change N (%)	Decreased N (%)
RPA	R1–R2	7 (70.0%)	0 (0.0%)	3 (30.0%)
	R2–R3	0 (0.0%)	7 (70.0%)	3 (30.0%)
RPB	R1–R2	6 (66.7%)	2 (22.2%)	1 (11.1%)
	R2–R3	0 (0.0%)	9 (100.0%)	0 (0.0%)
RHA	R1–R2	4 (40.0%)	6 (60.0%)	0 (0.0%)
	R2–R3	0 (0.0%)	10 (100.0%)	0 (0.0%)
RHB	R1–R2	6 (66.7%)	0 (0.0%)	3 (33.3%)
	R2–R3	1 (11.1%)	6 (66.7%)	2 (22.2%)
RSA	R1–R2	3 (30.0%)	2 (20.0%)	5 (50.0%)
	R2–R3	0 (0.0%)	8 (80.0%)	2 (20.0%)
RSB	R1–R2	7 (70.0%)	0 (0.0%)	3 (30.0%)
	R2–R3	1 (10.0%)	5 (50.0%)	4 (40.0%)

Reliability Estimates for Cut Scores

The reliability of cut scores obtained during a standard-setting session is thought of in terms of how consistent the cut scores are between 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 that lower values indicate a more reliable cut score.

The study's Design Document specifies that, within each postsecondary area, there exist two replicate panels (A and B) that each produce a median cut score. Therefore, there are only two independent observations for each postsecondary area. To calculate the standard error using two observations (Brennan, 2002), the following formula is used:

$$\hat{\sigma}_{\bar{X}} = \frac{|X_1 - X_2|}{2}$$

Tables 99 and 100 present these standard error estimates for mathematics and reading, respectively, for each postsecondary area. Also included in the tables are the 95% confidence intervals for each of the postsecondary area means. Confidence intervals are also displayed graphically in Figures 45 and 46 for mathematics and reading, respectively.

Table 99. Standard Error Estimates and Confidence Intervals for Mean Cut Scores by Postsecondary Area, Mathematics

Postsecondary Area	Panel A Cut Score	Panel B Cut Score	Mean Cut Score	Standard Error	95% Confidence Interval	
					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
HVAC	177	172	174.5	2.5	179.4	169.6
LPN	177	193	185.0	8.0	200.7	169.3
Pharmacy Technician	174	176	175.0	1.0	177.0	173.0

Table 100. Standard Error Estimates and Confidence Intervals for Mean Cut Scores by Postsecondary Area, Reading

Postsecondary Area	Panel A Cut Score	Panel B Cut Score	Mean Cut Score	Standard Error	95% Confidence Interval	
					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
HVAC	289	292	290.5	1.5	293.4	287.6
LPN	307	288	297.5	9.5	316.1	278.9
Pharmacy Technician	321	299	310.0	11.0	331.6	288.4

Figure 45. Mean Cut Scores and Confidence Intervals by Postsecondary Area, Mathematics

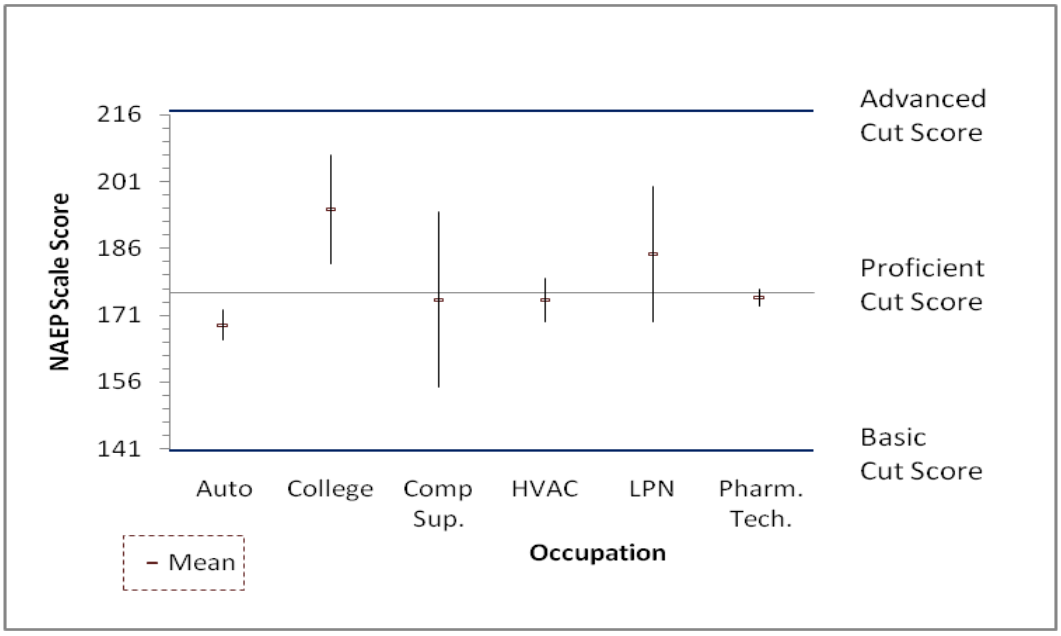
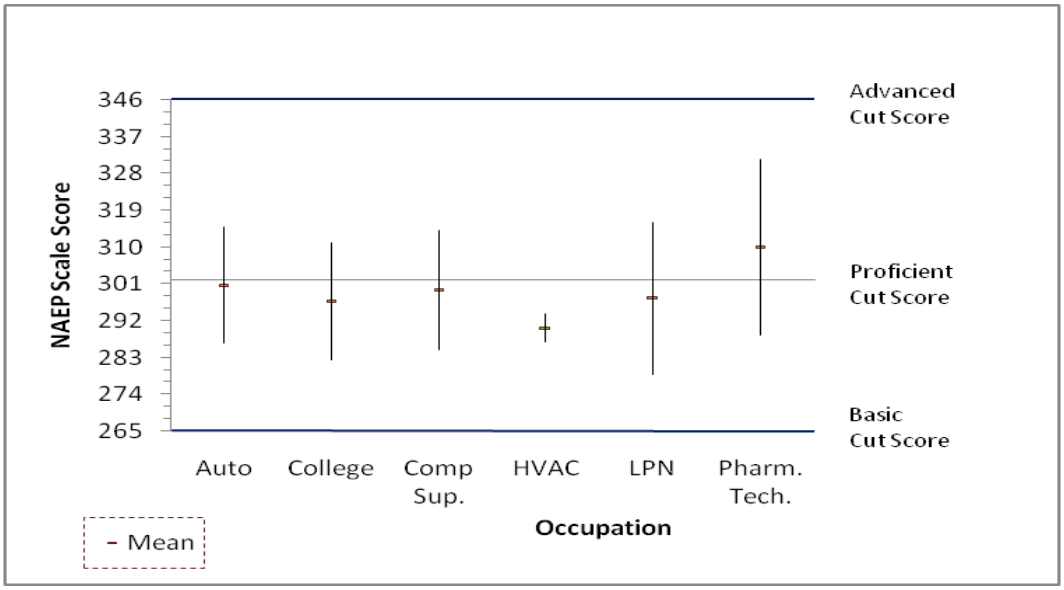


Figure 46. Mean Cut Scores and Confidence Intervals by Postsecondary Area, Reading



In Figures 45 and 46, the horizontal axis is placed at the Proficient cut score (i.e., 176 for mathematics and 302 for reading) obtained in the 2009 NAEP mathematics and reading achievement-level-setting processes 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 scores, respectively.

Based on the degree of overlap of the error bands in Figures 45 and 46, arguments could be made for collapsing findings across postsecondary areas (except for college-preparedness mathematics) for reporting cut scores. Therefore, it is reasonable to propose two cut scores for mathematics (one for college-preparedness and one for job-training). For reading, it is reasonable to propose a single cut score across all postsecondary areas.

Summary of Process Evaluations

Five process evaluations were administered during the JSS sessions for the purpose of obtaining immediate feedback from panelists regarding the clarity of directions, their comfort level with

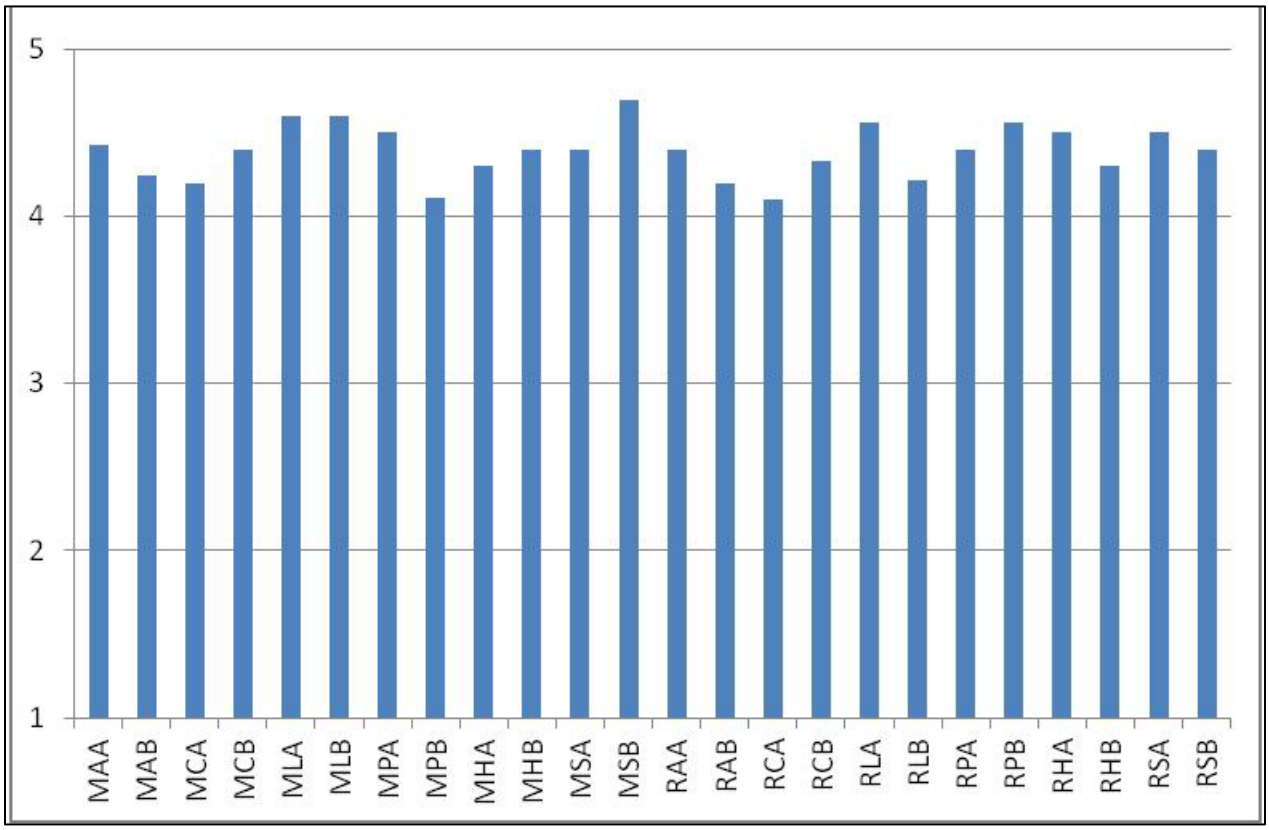
the process, and other procedural aspects that would indicate a need for intervention. Additionally, these evaluations provide some evidence for procedural validity of the process. Results for each postsecondary activity were presented in this report. This section summarizes results for selected evaluation items across the postsecondary activities for the following topics:

- Understanding of Tasks;
- Understanding of the Borderline Performance Description;
- Comfort and Confidence;
- Independence of Judgment; and
- Helpfulness of Software.

Understanding of Tasks. As shown in Figure 47, the average panelist responses across all postsecondary activities indicate that panelists had an adequate understanding of their tasks during the JSS sessions.

Figure 47. My understanding of the tasks I was to accomplish during each round was . . .

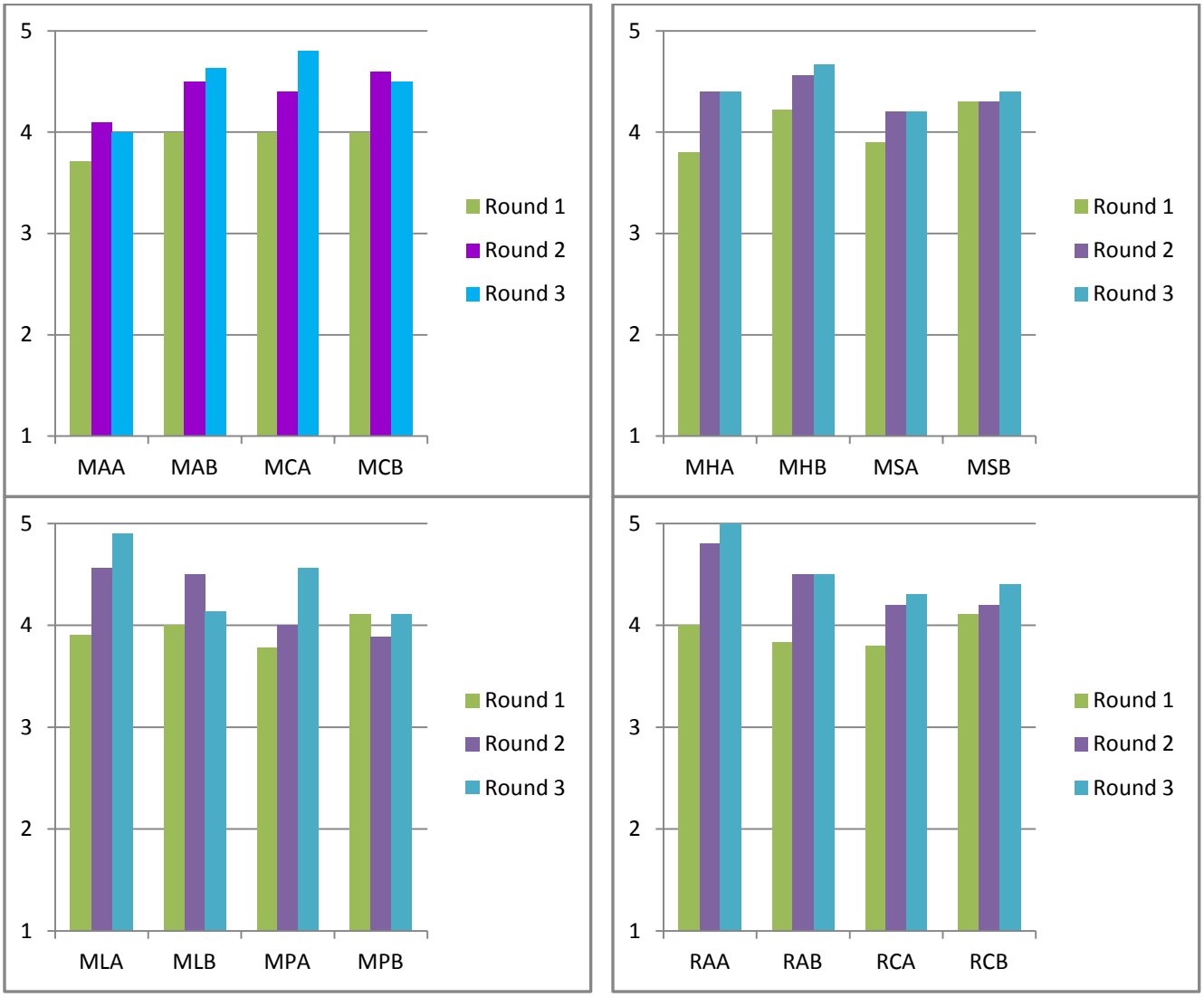
(1 = Totally Inadequate; 5 = Totally Adequate)

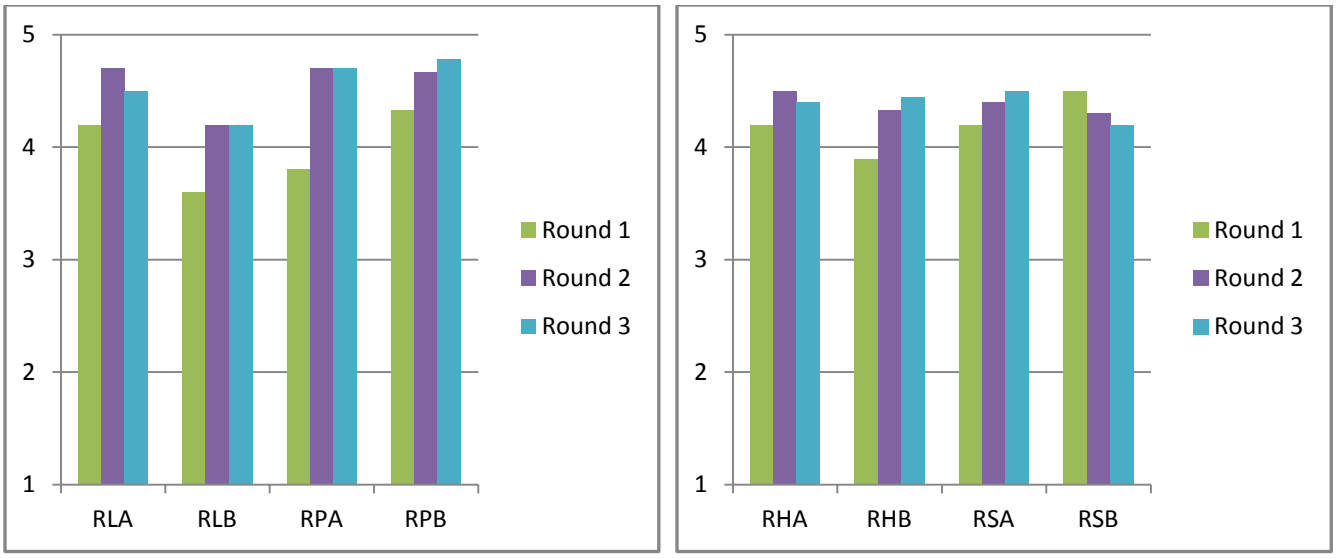


Understanding of the Borderline Performance Description. There was no time during the JSS sessions that, on average, panelists responded that they had a less than "Somewhat Adequate" understanding of the BPDs. Across the panels, panelist understanding of the BPDs was most often greater during round 3 than it was during round 1, with no panel expressing less than "Adequate" understanding during round 3. The bar charts in Figure 48 display these results.

Figure 48. At the time I placed my bookmark, my understanding of the BPD was . . .

(1 = Totally Inadequate; 5 = Totally Adequate)

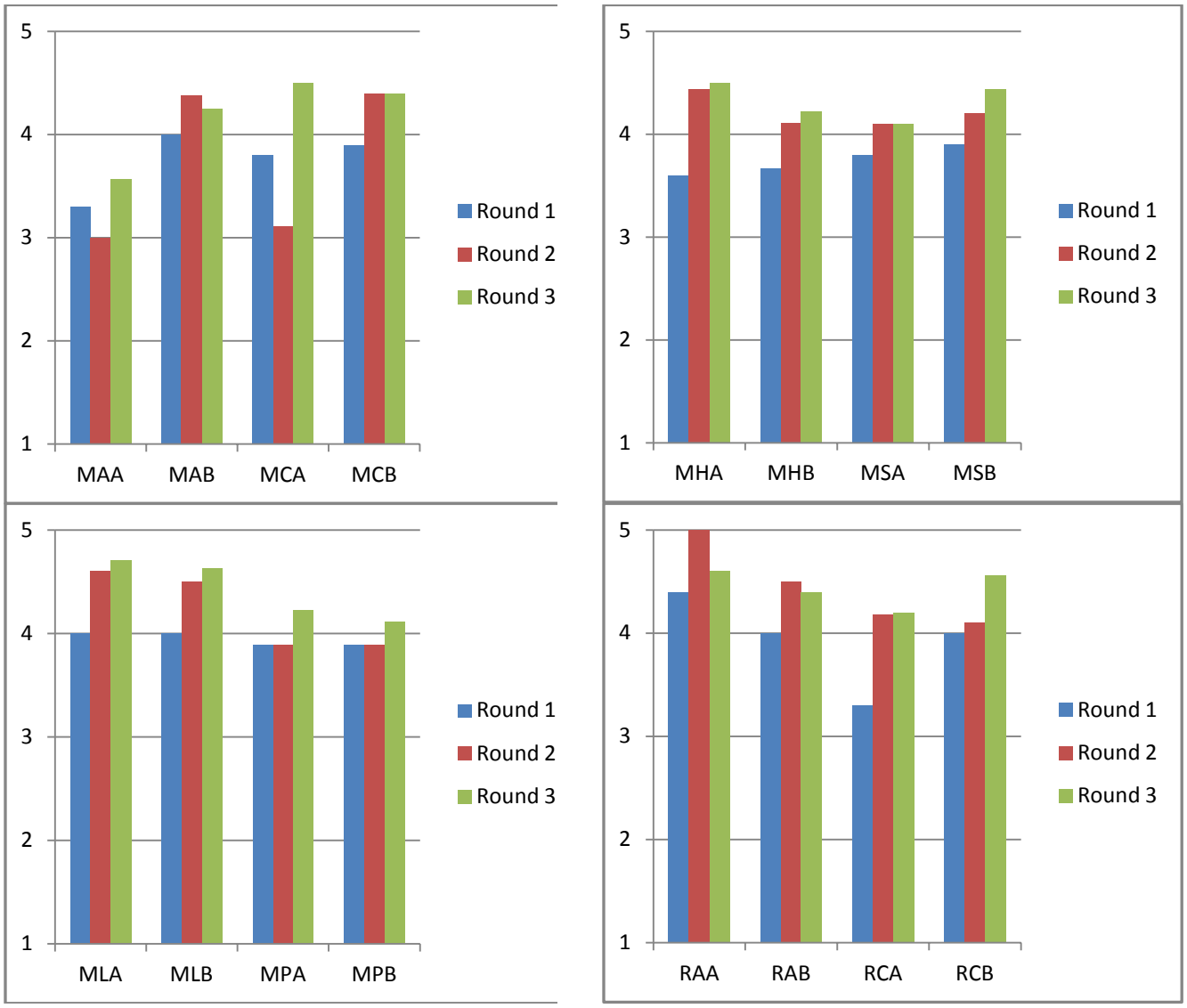


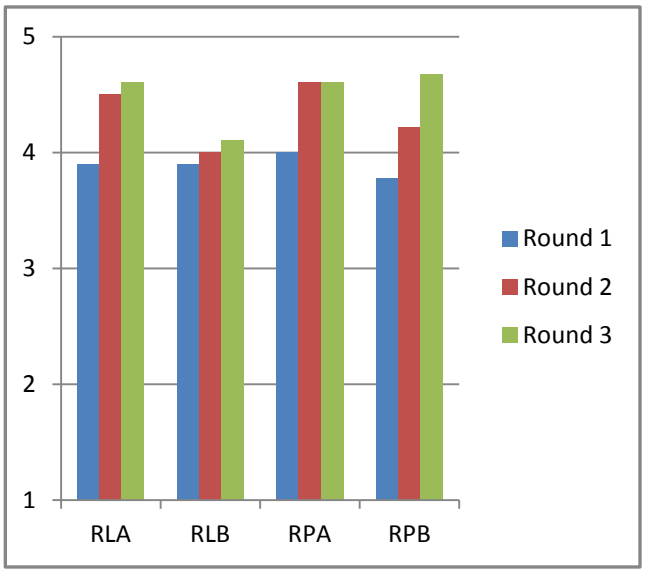


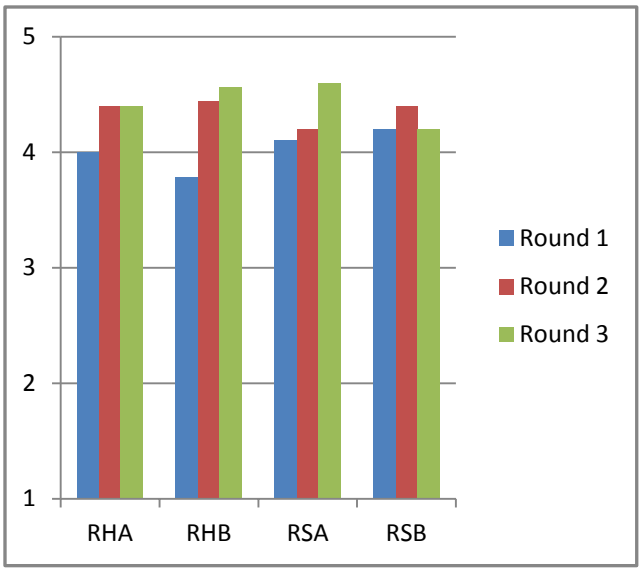
Comfort and Confidence. On average, panelists across postsecondary activities at least somewhat agreed that their cut scores were consistent with the BPDs. Further, this agreement increased between rounds 1 and 3 for most panels. Only one panel (Automotive Master Technician mathematics Panel A) had an average agreement level below a solid "Agree" (4.0) by the third round. The bar charts in Figure 49 display these results.

Figure 49. I believe my cut score is consistent with the BPD

(1 = Totally Disagree; 5 = Totally Agree)



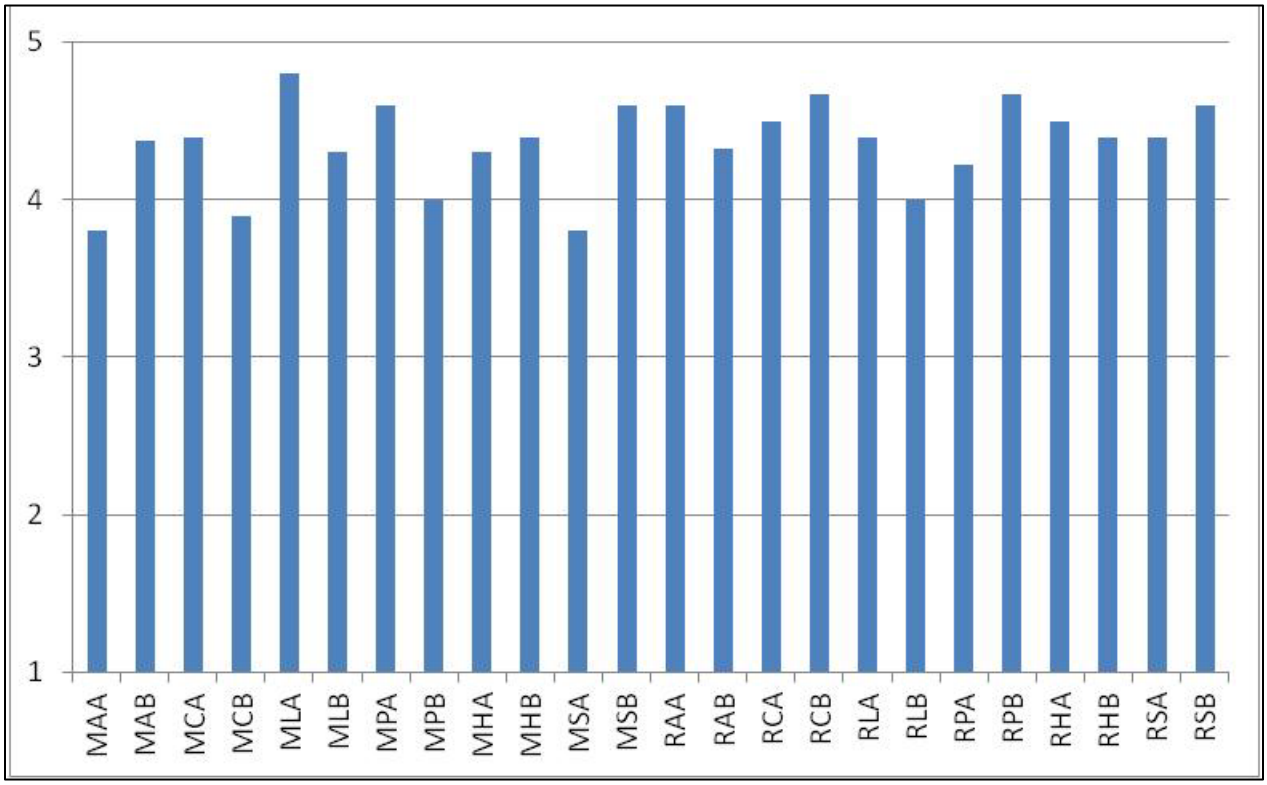

5
4
3
2
1
RLA
RLB
RPA
RPB
Round 1
Round 2
Round 3


5
4
3
2
1
RHA
RHB
RSA
RSB
Round 1
Round 2
Round 3

Additionally, Figure 50 shows that, on average, panelists were at least “Somewhat Confident” in their cut score recommendations. Panels with lower confidence include MAA, MCB, and MSA, indicating that lower confidence is not related to any particular postsecondary area. However, more variable averages were observed in the mathematics panels than among the reading panels.

Figure 50. The most accurate description of my level of confidence in the cut score recommendations I provided was . . .

(1 = Not at All Confident; 5 = Very Confident)



Relating to more specific parts of the process, panelists across the JSS sessions indicated that they at least somewhat agreed that they were comfortable with using the 0.67 probability for defining mastery in order to place a bookmark and also at least somewhat agreed with the ordering of items based on relative difficulty, except for the Automotive Master Technician

mathematics Panel A, which had an average agreement of 2.8 (with a 2.0 indicating disagreement). These results are presented in Figures 51 and 52, respectively.

Figure 51. I feel comfortable using a 2/3, or 0.67, probability for defining mastery in order to place the bookmark

(1 = Totally Disagree; 5 = Totally Agree)

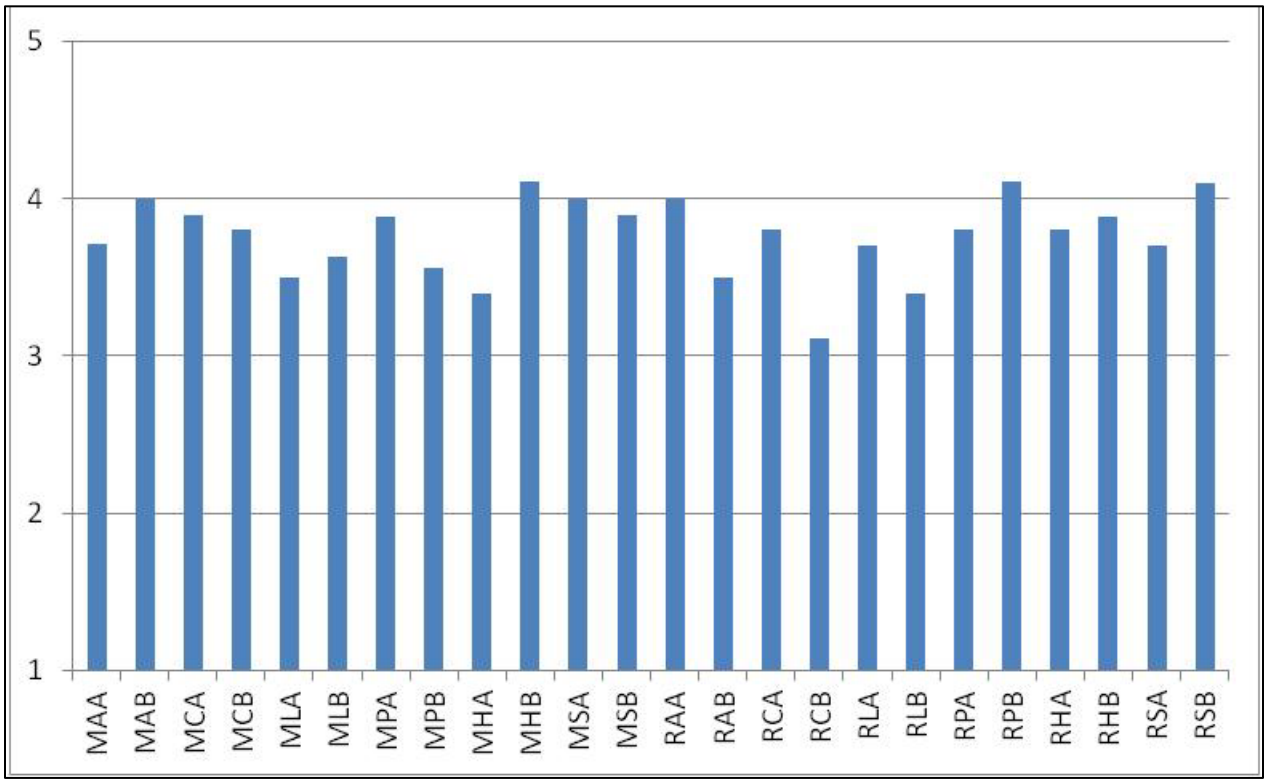
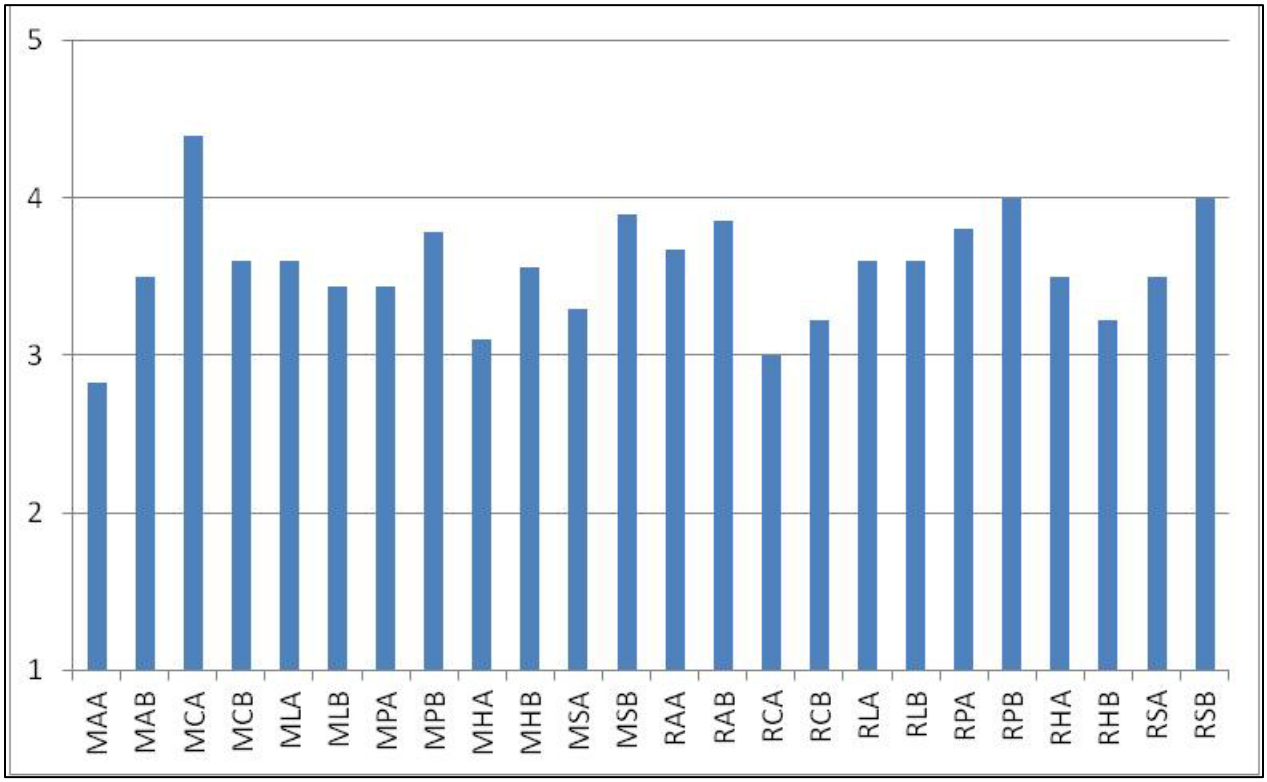


Figure 52. The ordering of the items in the OIB corresponded with my perceptions of the relative difficulty of the items

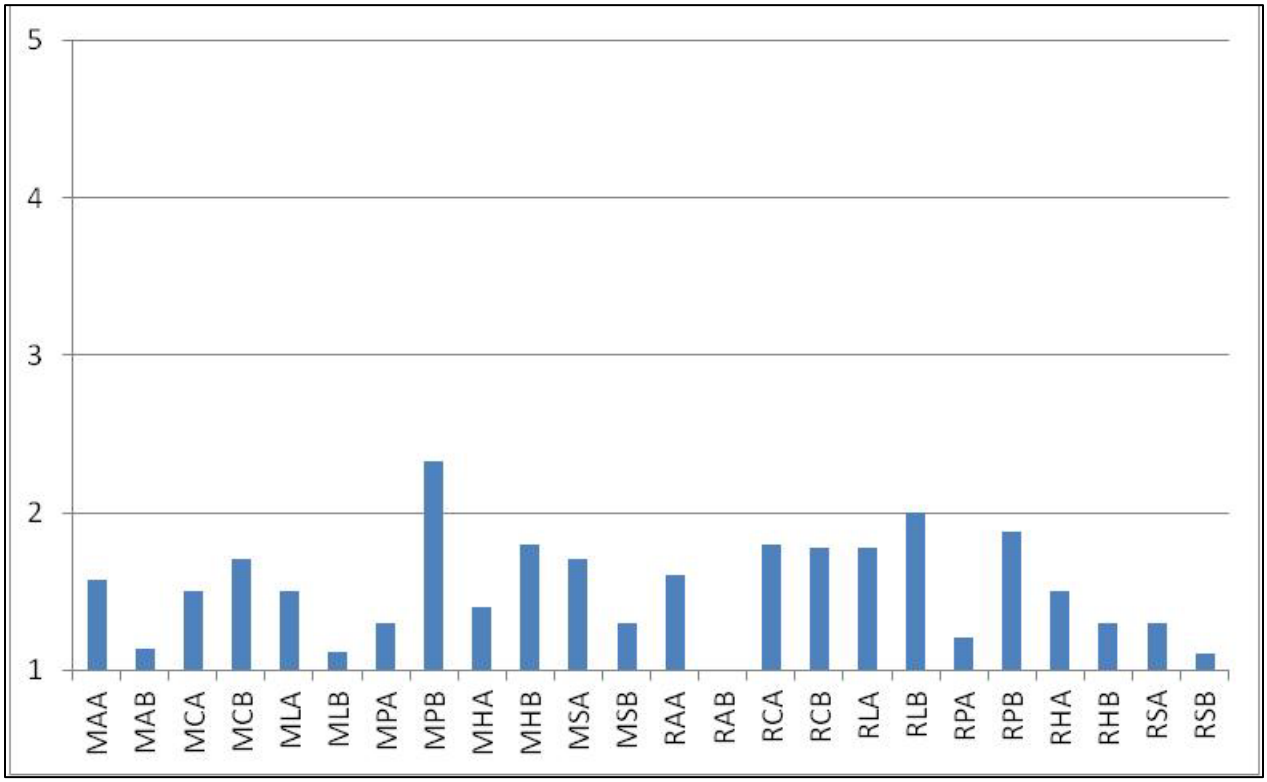
(1 = Totally Disagree; 5 = Totally Agree)



Independence of Judgment. Only one panel (Pharmacy Technician mathematics Panel B) indicated, on average, that some panelists may have felt pressure by others to make cut score recommendations that agreed with those of others in the group. All other panels indicated disagreement with this statement, as observed in Figure 53.

Figure 53. I felt pressured by others in my group to make my cut score recommendation agree with theirs

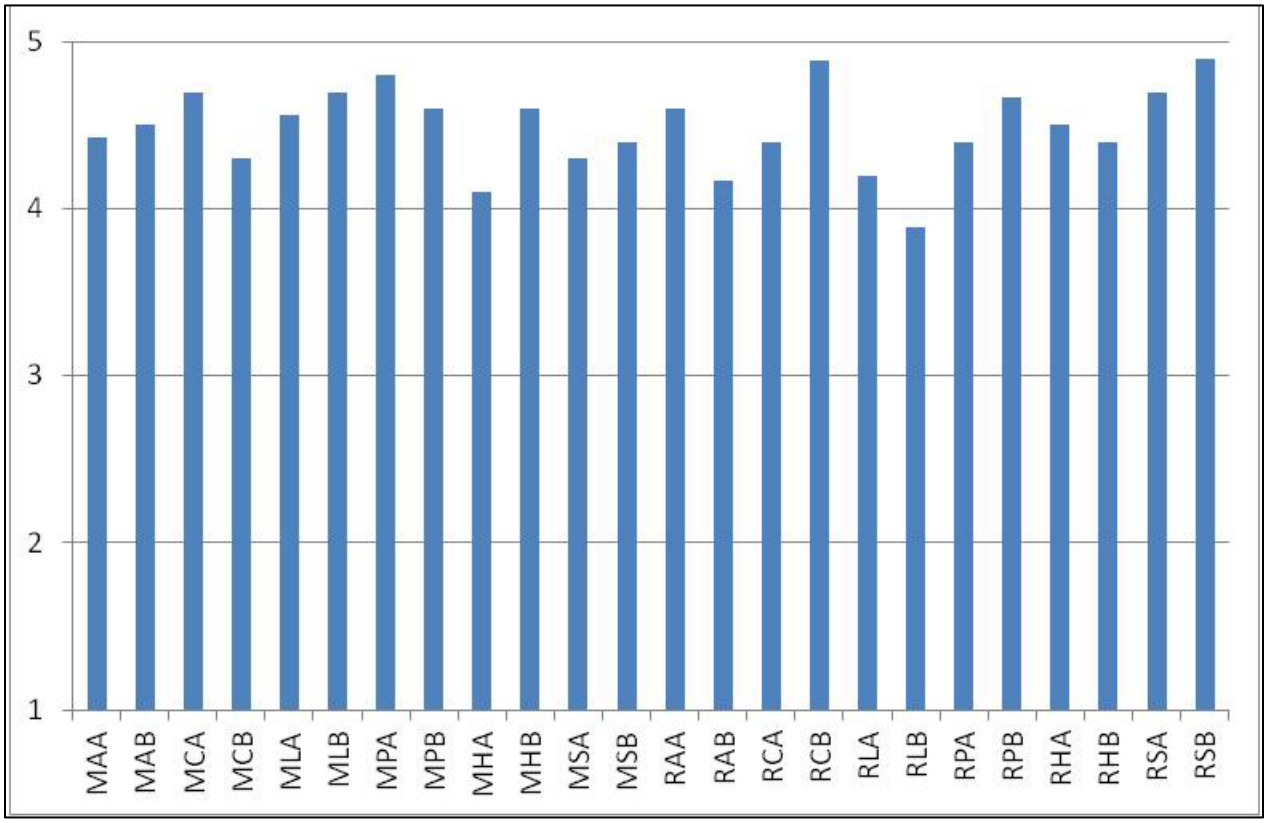
(1 = Totally Disagree; 5 = Totally Agree)



Helpfulness of Software. As indicated in Figure 54, across panels, the CAB was found to be helpful during the standard-setting process. Only one panel, LPN reading Panel B, indicated that the CAB was less than helpful ("Somewhat Helpful").

Figure 54. During the standard setting process, I found using CAB to be . . .

(1 = Not at All Helpful; 5 = Very Helpful)



Overall, no clear differences in responses to evaluations across postsecondary activities and content areas were observed. In general, there was very little variation in panel means across evaluation items, except in the following instances:

1. Automotive Master Technician mathematics Panel A expressed lower comfort and confidence with procedural aspects related to the BPDs and with the ordering of items based on relative difficulty;
2. There was more variation among mathematics panel means in relation to confidence in cut score recommendations than among the reading panel means; and

3. The mean response for Pharmacy Technician mathematics Panel B that related to feeling pressure to set cut scores in agreement with other panelists indicated that some panelists within this panel did feel such pressure.

Special Study

The purpose of the special study was to explore the utility of an alternative item map format within the context of these studies, based on observations that panelists had difficulty with the content of the assessments. Panelists 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. Item maps that grouped items by content were used in this special study, since, in previous bookmark-based standard setting studies for NAEP, the items had been grouped by content area on the item maps. 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 separate bookmarks for each content domain (e.g., for mathematics, Number Properties and Operations, Measurement and Geometry), 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 not only differentiated by color but were also separated into columns within the item maps. A sample item map for mathematics that was used for the special study is shown in Figure 55. Written instructions to the panelists and evaluation questions related to the special study are provided in Appendix V. Before 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 the occupation.

Figure 55. Special Study: Sample Modified Item Map

Redacted

Results from the first part of the special study are presented in Tables 101 and 102 (no results are presented for HVAC mathematics Panel A, because that panel did not participate in the special study). All results are reported on the NAEP scale. For each panel, the minimum, maximum, and median of panelists' cut scores within a content domain are presented. The corresponding medians are averaged based on the weights assigned to each content area according to the assessment framework. With the exception of the cut score for HVAC mathematics Panel B, the resulting cut scores are close to the cut scores set by each panel in round 3.

Table 101. Special Study: Content Domain Cut Scores for Mathematics

Content Area	Statistic	MHA	MHB	MSA	MSB
Number Properties and Operations (10%)	Minimum	N/A	126	160	175
	Maximum	N/A	279	224	213
	Median	N/A	210	171	185
Measurement and Geometry (20%)	Minimum	N/A	173	166	115
	Maximum	N/A	301	202	203
	Median	N/A	203	168	184
Data Analysis and Probability (25%)	Minimum	N/A	152	156	175
	Maximum	N/A	223	200	205
	Median	N/A	175	166	189
Algebra (35%)	Minimum	N/A	152	140	169
	Maximum	N/A	227	198	201
	Median	N/A	177	166	182
Weighted Average of Medians Round 3 Cut Score		N/A 177	188 172	167 165	185 185

Table 102. Special Study: Content Domain Cut Scores for Reading

Content Area	Statistic	RHA	RHB	RSA	RSB
Literary (30%)	Minimum	262	240	287	278
	Maximum	292	322	308	308
	Median	288	289	291	***305***
Informational (70%)	Minimum	278	288	288	289
	Maximum	299	306	358	324
	Median	293	292	298	307
Weighted Average of Medians Round 3 Cut Score		291 289	291 292	296 292	306 307

Tables 103 and 104 present the numbers of items panelists deemed irrelevant for their training programs. No results are presented for HVAC reading Panel A and mathematics Panel A, because those panels' respective facilitators provided directions that differed substantially from the directions provided by all other facilitators. Two criteria were used when counting the number of irrelevant items identified by each panel. The first criterion was that an item was considered irrelevant if at least one panelist indicated that it was irrelevant. The second criterion

was that an item was considered irrelevant if at least half of the panelists indicated that it was irrelevant. The results are presented by content domain. The numbers in parentheses under each content domain represent the total numbers of items presented for the two groups.

Table 103. Special Study: Number of Items Deemed Irrelevant, Mathematics

Criterion	Content Area	MHA		MHB		MSA		MSB	
		N	%	N	%	N	%	N	%
Deemed Irrelevant by at Least One Panelist	Number Properties and Operations (A = 9; B = 12)	N/A		11	92	5	56	9	75
	Measurement and Geometry (A = 47; B = 40)	N/A		38	95	47	100	40	100
	Data Analysis and Probability (A = 21; B = 30)	N/A		30	100	18	86	20	67
	Algebra (A = 36; B = 29)	N/A		27	93	34	94	25	86
Deemed Irrelevant by at Least Half of the Panelists	Number Properties and Operations (A = 9; B = 12)	N/A		6	50	1	11	6	50
	Measurement and Geometry (A = 47; B = 40)	N/A		8	20	37	79	31	78
	Data Analysis and Probability (A = 21; B = 30)	N/A		12	40	2	10	10	33
	Algebra (A = 36; B = 29)	N/A		15	52	5	14	18	62

Table 104. Special Study: Number of Items Deemed Irrelevant, Reading

Criterion	Content Area	RHA		RHB		RSA		RSB	
		N	%	N	%	N	%	N	%
Deemed Irrelevant by at Least One Panelist	Literary (A = 28; B = 29)	N/A		25	86	9	32	4	14
	Informational (A = 68; B = 71)	N/A		33	46	35	51	7	10
Deemed Irrelevant by at Least Half of the Panelists	Literary (A = 28; B = 29)	N/A		5	17	0	0	0	0
	Informational (A = 68; B = 71)	N/A		0	0	3	4	0	0

After identifying irrelevant items, panelists were asked to answer six questions regarding setting cut scores for academic preparedness by content domain and dealing with irrelevant items. Panelist instructions and the process evaluation questionnaire are provided in Appendix V.

A summary of panelists' responses to each question is presented in Table 105. These results represent only 67 panelists, as panelists from HVAC mathematics Panel A did not have an opportunity to respond to the questionnaire. It is notable that 78% (52, including "Totally Agree," "Agree," and "Somewhat Agree" responses) of the responding panelists responded that, if cut scores were computed by averaging content-area cut scores, the computed values would be comparable to the round 3 cut scores. Additionally, 85% (57) responded that the cut scores set on individual content domains enabled them to more accurately represent minimal preparedness requirements, and 72% (48) indicated that they at least somewhat agreed that dealing with irrelevant items was easier when setting the cut scores on separate domains.

Table 105. Special Study: Process Evaluation Results

	Total Number of Panelists Responding					
	Totally Agree N (%)	Agree N (%)	Somewhat Agree N (%)	Disagree N (%)	Totally Disagree N (%)	Not Answered N (%)
I understood how to place the bookmark for each content area to represent minimal preparedness for entry into a job training program for HVAC technicians/Computer Support Specialists.	24 (36%)	34 (51%)	5 (7%)	0 (0%)	3 (4%)	1 (2%)
If an average cut score were computed from the combination of the bookmarks I just placed for the items in the different content areas, I think the cut score would be about the same as my cut score in round 3.	9 (13%)	26 (39%)	17 (25%)	12 (18%)	1 (2%)	2 (3%)
Setting a cut score in each content area enabled me to represent the level of performance for a minimally prepared student more accurately than when the cut score was based on all items.	10 (15%)	28 (42%)	19 (28%)	6 (9%)	1 (2%)	3 (4%)
Placing a bookmark in each content area is more difficult than placing a single bookmark using all items together.	1 (2%)	12 (18%)	13 (19%)	32 (48%)	4 (6%)	5 (7%)

	Total Number of Panelists Responding					
	Totally Agree N (%)	**Agree N (%)**	**Somewhat Agree N (%)**	**Disagree N (%)**	**Totally Disagree N (%)**	**Not Answered N (%)**
I was able to deal with "irrelevant" items more easily when setting bookmarks in separate content areas than when setting the bookmark using all items.	8 (12%)	23 (34%)	17 (25%)	11 (16%)	2 (3%)	6 (9%)
I found it easier to use the bookmarking process as we did it in the first three rounds than in this process for bookmarking in each content area.	3 (4%)	17 (25%)	16 (24%)	24 (36%)	2 (3%)	5 (7%)

Recommendations

This set of standard-setting studies is an important component of the Governing Board's larger postsecondary preparedness initiative, and the focus on career-preparedness activities provides timely and useful information that will inform discussions surrounding the degree of overlap between preparedness for college and preparedness for the workplace. The methodology prescribed by the Governing Board, as described in the Design Document and implemented by Measured Progress, was thorough and comprehensive. Despite the rigor of the study design and its implementation, however, certain challenges arose, in particular within career-preparedness panels. In response to those challenges, the following lessons learned and recommendations are submitted for future standard-setting studies of this type.

Recruiting Procedures. This subsection begins with a description of challenges related to the recruiting procedures implemented for college-preparedness and career-preparedness panelists and suggests approaches to improve this process in future studies.

College-Preparedness Panelists. Recruitment of postsecondary panelists for this project's college-preparedness panels was largely successful, with a substantial pool of qualified candidates from which to select optimal panelists. It is likely that offering to provide an honorarium to panelists assisted in recruitment; however, several potential candidates representing prestigious four-year postsecondary institutions declined to participate, stating that the honorarium was less than they would typically earn for consulting work. Recruitment of secondary-level panelists proved somewhat more difficult, perhaps in part due to the timing of the pilot study and the first operational session (late April and late May, respectively—i.e.,

toward the end of the academic year), although sufficient numbers of qualified secondary-level panelists were recruited for the college-preparedness panels.

Career-Preparedness Panelists. Recruitment of job-training instructors to serve on certain career-preparedness panels was more difficult, for various possible reasons. Across the occupations, most job-training program heads and instructors were not familiar with NAEP and the type of activity entailed in standard setting studies; successfully explaining the importance, purpose, and approach of this type of study proved more difficult than when recruiting instructors from more traditional academic programs. The timing of the first and second operational workshops coincided with the end of the academic year for many, a difficult time to be away from classes; also, within at least some job-training programs, authorization to take a full week (four days for each workshop, plus a day for travel) away from classes appeared to be difficult to obtain. In addition, based on correspondence with a several nominees, it is not uncommon for job-training instructors in some occupations (e.g., Automotive Master Technician, Computer Support Specialist, HVAC) to also work as practicing technicians, thus making it even more difficult to commit to the amount of time required for each workshop.

WestEd submits the following lessons that it learned through this study's recruitment process for consideration by the Governing Board and/or its contractors when recruiting for future studies.

- *Consider variability between occupations in responsiveness to recruitment efforts.* The Governing Board thoroughly and systematically reviewed a pool of potential occupations when selecting the five exemplar occupations to be included in this study, considering, among other factors, the availability of eligible programs and panelists. Even though the number of formal job-training programs varied by occupation, all exemplar occupations seemed likely to produce the requisite numbers of panelists. Through its recruitment

efforts, however, WestEd discovered that occupations varied dramatically in how they train their workforces, how their job-training programs are accredited or certified, and how communication flows to and among job-training programs; it also found that response rates differed considerably among job-training programs. For future studies involving occupations, advance planning and research is needed to estimate the amount of time required for successful recruitment to be completed.

- *Streamline nomination materials.* During initial recruitment efforts, WestEd used an introductory letter used for recruiting from typical academically focused audiences to request nominations from the heads of job-training programs. While these materials were effective in recruiting college-preparedness panelists, they were less appealing to job-training instructors. Therefore, during the recruitment process, WestEd transitioned to more graphical, streamlined materials. Response to the streamlined materials was greater.
- *Ensure that panelists' eligibility requirements are appropriate to the task.* It became apparent through the standard-setting process that some panelists in some of the occupational workshops lacked the content knowledge and skills to effectively interact with the NAEP content, particularly in mathematics. While panelists were required to be familiar with the content-specific knowledge, skills, and abilities required in their programs, they were not required to teach courses specifically addressing either content area. Across occupations, it is not reasonable to expect to find educators who teach reading-specific courses; however, in some occupations it is common for job-training programs to include mathematics-specific courses (such as Math for Pharmacy Technicians). When recruiting from occupations that offer such courses in their job-

training programs, it may be advisable to target recruitment to job-training instructors who teach those courses.

Advance Webinars in the Development of BPDs. The use of online webinars for training panelists and engaging them in the development of preliminary BPDs in advance of the standard-setting sessions was an innovation for NAEP standard setting. An evaluation of the effectiveness of this approach could inform its use in future standard-setting studies.

Facilitator Training. Given the scheduling of the JSS sessions, a total of eight process facilitators were required to participate in each session; these facilitators were selected from three organizations: WestEd, Measured Progress, and EPIC. While all facilitators had the requisite qualifications for conducting standard-setting on the NAEP assessments in mathematics and reading, they reflected the somewhat different styles of their organizations, and addressing these styles while the standard-setting sessions were in process posed some challenges. If future studies involve working with a large group of facilitators, it would be advisable to provide extensive training in advance of standard-setting and/or recruit all facilitators from the same organization.

Standard-Setting Procedures. The bookmark method stipulated by the Design Document worked well for the college-preparedness pilot study and operational workshop. Panelists for the college-preparedness workshops came from traditional college and secondary-level academic programs and were, as a whole, relatively familiar with NAEP and with the type of activities required of them. Recruitment of these panelists and implementation of the standard-setting process proceeded largely as planned. However, on the whole, job-training instructors recruited for the career-preparedness workshops were less familiar with the objectives and structure of

NAEP and with standard setting in general. As a group, they tended to struggle more than the college-preparedness panelists with the language of the NAEP frameworks and with the pools of NAEP items assigned to them. In addition, panelists from some occupations were not well versed in the academic prerequisites for their occupations. The diversity among occupations, panel groups, and panelists posed unique challenges to the implementation of the bookmark method as planned within the job-training workshops.

In response to the challenges represented by these "lessons learned," modifications to the JSS process were made over the course of the pilot study and the three operational sessions, which yielded a refined implementation by the third operational session.

- The instructions and guidance provided to panelists when identifying knowledge, skills, and abilities (KSAs) for the items were refined through the pilot study and the first two operational sessions. It is recommended that item descriptions be provided to panelists—especially panelists not recruited from traditional academic programs—for use in the process of developing KSA annotations for future studies.

- The Design Document called for the sharing of content facilitators across pairs of replicate panels, and this design seemed appropriate for the college-preparedness panels. However, it is recommended that, for future standard-setting involving occupations, each job-training panel group be assigned its own content facilitator. Assigning a content facilitator full-time to a panel will allow more time and opportunities for panelists to seek guidance and consultation regarding content-related issues, such as KSAs.

- The decision to include secondary-level teachers in the final JSS session (for the Computer Support Specialist and HVAC workshops) was made to increase the content

knowledge and skills represented among the panelists. However, in some workshops, the secondary-level instructors became too influential in establishing the content areas' BPDs. The inclusion of such instructors in the process should be carefully considered and their roles explicitly communicated in future studies.

- Across all workshops, the computerization of aspects of the standard-setting process proved successful for this project, and the continued use of computerized procedures in future studies is recommended.

Handling of Irrelevant Items. A number of panelists across the pilot study and the first two operational JSS studies reported NAEP items to be irrelevant to their job-training programs; therefore, the Governing Board requested and designed a special study to be implemented on the last day of the third operational JSS session to explore this issue. A systematic strategy for instructing panelists on how to rate seemingly irrelevant items, drawing upon information gleaned from this special study, is recommended for future standard-setting studies of this nature.

References

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

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

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

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

Brennan, R. L. (2002). *Estimated standard error of a mean when there are only two observations* (CASMA Technical Note 1). Iowa City, IA: University of Iowa, College of Education, Center for Advanced Studies in Measurement and Assessment.

Council of Chief State School Officers. (2001). *State student assessment programs annual survey*. Data Volume II. Washington, D.C.: Author.

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

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures using behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

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

Measured Progress & WestEd. (2011). *National Assessment of Educational Progress grade 12 preparedness research project judgmental standard setting (JSS) studies: Technical report.* Dover, NH: Measured Progress.

National Assessment Governing Board. (2008). *Mathematics framework for the 2009 National Assessment of Educational Progress*. Washington, D.C.: U.S. Department of Education.

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

National Assessment Governing Board. (2010b). *Reading framework for the 2011 National Assessment of Educational Progress*. Washington, D.C.: U.S. Department of Education.

National Center for Education Statistics. (2008–09). *Common Core of Data, Public Elementary/Secondary School Universe Survey (v.1b)* [Data file]. Retrieved from http://nces.ed.gov/ipeds/datacenter/

National Center for Education Statistics. (2009). *Integrated Postsecondary Education Data System* [Data file]. Retrieved from http://nces.ed.gov/ipeds/datacenter/

National Center for Education Statistics. (2010). *The nation's report card: Grade 12 reading and mathematics 2009 national and pilot state results* (NCES No. 2011-455). Washington, D.C.: U.S. Department of Education, Institute of Education Sciences.